Various RNA-binding proteins and their conditional networks explain miRNA biogenesis and help to reveal the potential SARS-CoV-2 host miRNAome system

Upendra Kumar Pradhan[1,3], Prince Anand[2,3], Nitesh Kumar Sharma[1,3] , Prakash Kumar[1,3], Ashwani Kumar[1,3], Rajesh Pandey[4], Yogendra Padwad[2,3] and Ravi Shankar[1,3*]

[1] Studio of Computational Biology & Bioinformatics,

CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT),

Palampur (HP), 176061, India.

[2]Pharmacology and Toxicology Lab,

CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT),

Palampur (HP), 176061, India.

[3]Academy of Scientific and Innovative Research (AcSIR),

CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT)

Palampur (HP), 176061, India

[4]Genomics and Molecular Medicine Unit,

CSIR-Institute of Genomics & Integrative Biology, Delhi, 110007, India

*Corresponding Author**:** ravish@ihbt.res.in

Abstract

Background: In-spite of ubiquitous expression of DROSA/DICER, miRNA formation and maturation are highly spatiotemporal implying involvement of other factors in their biogenesis. Several key studies have elucidated functions of few other RNA-binding proteins (RBPs) in miRNAs biogenesis, making it necessary to look miRNA biogenesis models with fresh approach.

Results: A comprehensive study of >25TB of high-throughput data revealed that various combinations of RBPs and their networks determine the miRNA pool, regardless of DROSHA/DICER. The discovered RBP and miRNA associations displayed strong functional alliances. An RBP, AAR2, was found highly associated with miRNAs biogenesis, which was experimentally validated. The RBPs combinations and networks were tested successfully across a large number of experimentally validated data and cell lines for the observed associations. The RBP networks were finally modeled into a XGBoosting-regression based tool to identify miRNA profiles without any need of doing miRNA-seq, which scored a reliable average accuracy of 91% on test sets. It was further tested across >400 independent experimental samples and scored consistently high accuracy. This tool was applied to reveal the miRNAome of Covid19 patients about which almost negligible information exists. A significant number of Covid19 specific miRNA targets were involved in IFN-gamma, Insulin/IGF/P3K/AKT, and Ub-proteasome systems, found in cross-talk with each other and down-regulated heavily, holding promise as strong candidates for therapeutic solution. A large number of them belonged to zinc-finger family.

Conclusion: There are several RBPs and their networks responsible for miRNA biogenesis, regardless of DROSHA/DICER. Modeling them successfully can reveal miRNAomes with deep reaching impact.

Keywords: RBP, miRNA, CLIP, Bayesian, Network, XGBoost, COVID, SARS-CoV-2

Background

The regulatory sRNAs in eukaryotes, commonly called as miRNAs, are supposed to regulate at least one third of the genes post-transcriptionally. This observation was made in 2003 by Lewis et al [1] when hardly few hundreds of miRNAs were known. One can imagine the extent of regulation by these sRNAs at present when there are more than 2,000 miRNAs reported alone for human in miRBase [2] and total miRNAs >38,000. In year 2015, our group had reported 11,234 regulatory sRNAs in human system alone if one looks beyond the canonical characterization for miRNAs [3]. In general, barring some exceptions, RNAs are essentially accompanied by some proteins, and miRNAs are not exception. These proteins are essential for RNA synthesis, maturation, storage, transport, and regulation of stability and function (Additional file 2: Figure S1). The RNA molecules are constantly associated and chaperoned by highly dynamic protein complexes which contribute to the fate of the bound RNA. The scope of this regulation is crucial for the complexity of the organism. This association between the RNA binding proteins (RBP) and RNAs define another level of regulation whose understanding has been largely limited to the process of translation and nascent for the rest. However, projects like FANTOM [4] and ENCODE [5] have compellingly pushed us to look beyond the protein coding biology only. In the last 10 years there has been an explosion of sequencing data due to revolutions in sequencing technologies, opening gates to much bigger universe of unexplored entitites of biology and same time posing new challenges and questions. This has resulted into a drastic change in our understanding about the genomic system where a lot of stake is there through non-coding RNAs and different regulatory mechanisms. There are various types of non-coding RNAs, including well-known RNAs with specific functions like rRNAs or tRNAs, nucleolar snoRNAs, snRNAs which guide chemical modifications of other RNAs and help in splicing, and finally the two big classes of regulatory ncRNAs: small non-coding RNAs (i.e. miRNAs) and lncRNA (>200nt) with a growing list of different functions, including molecular sponging/buffering, the regulation of chromatin accessibility and transcription etc to name a few. The defining features of small RNAs (sRNAs) are

their short length and their association with members of the Argonaute family of proteins (AGO1-4) which guide to the regulatory targets, typically resulting into bringing down of expression of the target genes. Beyond these defining features, different small RNA classes guide diverse and complex schemes of gene regulation at post-transcriptional level. There is a growing number of studies now which suggest that despite of having a common mechanism of function, these sRNAs originate from different sources through different mechanisms and RBP associations [6,7]. Additional file 2: Figure S2 illustrates few of them.

Our understanding towards RBPs with roles beyond translation is very primary and continuously increasing with recent advances in high throughput technologies. Recent efforts to identify new RBPs by a screening technique like RNA interactome capture revealed that there might be about 1,500–1,900 RBPs in human cells and many of them lack the typical RNA binding domains (RBD) [8]. Most of the RBPs have wide range of diverse functions arising from combinatorial effects of these domains. RBPs can either exist in the nucleus as well as in the cytoplasm as they are often shuttle in between and generally facilitate nuclear transport also. To date, more than hundred RNA modifications have been reported with most common changes caused as RNA capping, polyadenylation, RNA editing, methylation, alternate splicing, and degradation. In fact the most important processes like translation and completion of transcription themselves are a big example of the dynamic of RBP:RNA interactions [9]. The process of RNA maturation through splicing is mediated by interactions of the core spliceosome and an array of accessory RBPs like ACIN1, ELAVL1, WTAP, FMR1, FUS, HNRNPH1, HNRNPA1 etc. [10]. Studies on localization process in various eukaryotic systems have discovered involvement of numerous RBPs. The ‘zipcodes’ in the untranslated regions (UTR) form secondary structures that serve as a docking site for the RBPs and promote the transport process [11]. AU rich elements (AREs) have been found critical for RNA stability and work as a cis regulatory element hosting binding sites for many RBPs [12]. There is

also evidence for other RBPs (ALKBH5, ELAVL1, FMR1, FXR1, FXR2, HNRNPD, HNRNPM, IGF2BP3) in regulation of RNA stability.

The biogenesis of miRNAs is a complex process primarily dependent upon some well recognized RBPs like DROSHA, DGCR8, and DICER. The canonical miRNA biogenesis starts with primary miRNAs (pri-miRNAs) which is processed by a complex of RNase III proteins, DROSHA and DGCR8 to release the precursor miRNA (pre-miRNA) [13]. The pre-miRNA is transported to the cytoplasm by another RBP, exportin-5 protein (XPOT5). In the cytoplasm, the RNase III DICER binds the pre-miRNA and cleave it to produce the mature miRNA duplex. With the help of the dsRBPs TRBP or PACT, one of the two strands is loaded into the RNA induced silencing complex (RISC) where it directly interacts with a member of the Argonaute protein family (AGO1-4) to cause suppression of the interacting target RNA [14]. An increasing amount of evidence also suggests that Argonautes also play a role in the processing of a pri-miRNA in DICER independent manner [15]. Besides this, a good number of miRNAs have been found originating directly as product of splicing, called as mirtrons [16].

For a longtime this has been argued that the DGCR8/DROSHA and DICER are indispensable components in miRNA biogenesis. Though these RBPs express themselves in almost all tissues, the miRNAs are mostly highly spatio-temporal and specific in their expression. Also, as already discussed above, there are number of regulatory sRNAs whose biogenesis involves many other RBPs besides these regular RBPs. In fact, in 2008, a long pending puzzle of regulation of *let*-7 miRNAs was resolved which would express in highly tissue specific manner where its regulation was thought to be dependent on some transcription factor. It was found that RBP LIN28 interaction with *let*-7 miRNA blocked its maturation [17]. This is perhaps the first study implicating other than usual miRNA processing RBPs in miRNA regulation. By the end of year 2012, a very interesting work with pri-miRNA transcription identified that FUS/FET proteins co-trascriptionally bind the

pri-miRNA sequence, which in turn facilitates the binding of DROSHA/DGCR8 complex [18]. Deleting FET protein blocked the miRNA formation and DROSHA/DGCR8 working. RBPs could also regulate miRNA biogenesis by affecting the expression or stability of the canonical proteins in miRNA pathway. For example, the RBP AUF1 regulates the general miRNA biosynthesis by inhibiting DICER expression by binding to DICER mRNA and decreasing its half-life [19]. Another good example of miRNA biogenesis beyond canonical pathway is AGO2 led miRNA formation [20].

In 2011, while characterizing miRNAs for their regulatory properties, our group had came across findings which were implicating RBPs in miRNA formation [21]. In this pioneering work, we had reported that miRNAs precursors host a number of RBP binding spots and the corresponding RBPs displayed strong expression relationships with associated miRNAs while fitting to the Regulon model. A hypothesis was proposed for miRNA maturation, where RBPs other than canonical miRNA processing RBPs were proposed to regulate miRNA formation (Additional file 2: Figure S3). At that time, due to the scarcity of high-throughput data for cross-linking, the study was limited. However, this entire work made the foundation of the present work with two logical motivations: 1) RBPs displayed strong associations with miRNA along with binding sites across the precursors which could add further dimensions in the regulatory control of cell system, and 2) There are more than 2,000 miRNAs in human cells which canonical DROSHA/DICER system may process, everywhere. Yet, all miRNAs are not expressed in every condition. There must be some other RBPs other than DROSHA/DICER system to control the process of miRNA formation and provide such spatio-temporal expression pattern of miRNAs besides the transcriptional control.

Due to technological advancement caused by NGS methods, transcriptome wide RBP interactome can now be discovered. These methods include sRNA-seq, RNA-seq and cross-linking based sequencing techniques like CLIP-seq. A list of different experimental methods (classified as RNA-

centric and Protein-centric) for identification of protein-RNA interaction is provided in Additional file 1:Table S1 along with their merits and demerits. The volume of CLIP-seq data has an increasing trend post 2010 (Additional file 2: Figure S4), clearly suggesting the acknowledgment of RBP led regulation in cell system. If the sRNA-seq, RNA-seq, and CLIP-seq data are considered in a relevant and condition specific manner, an enormous level of information regarding miRNA:RBP regulatory system could be revealed. In fact, in this direction, very recently a study was conducted which used two cell lines CLIP-seq data for 126 RBP from ENCODE and looked for their overlap with miRNA regions. They validated 10 RBPs for the observed miRNA binding coordinate overlaps and their stake in miRNA processing [22]. Their study was limited to two cell lines, though they made observation that these RBP interactions are suggestive of being cell line and condition specific. This underlines that here is a need of universal theory of miRNA regulation by RBPs in combinatorial manner which could hold across wide range of cell lines and conditions, clearly explaining miRNA biogenesis for any given condition. High spatio-temporal behavior of miRNA biogenesis can be explained by consideration of various interacting RBPs in conditional and networked fashion. Consideration of wide range of conditions and cell lines and multiple angle analysis will always give better picture. With above mentioned history and motivation for this study, an effort has been made here to leverage from the high-throughput sequencing data burst to derive universal models for miRNAs regulation by RNA binding proteins and their conditional networks which could be applied with high confidence across wide range of conditions. The miRNA biogenesis models were finally implemented into a machine learning approach to accurately predict miRNAs profile of any given condition in the absence of miRNA profiling experiments like miRNA-seq or in case of non-availability of miRNA informaiton. This software has been tested across a large number of experimental data where it displayed high accuracy, consistency, and robustness. Further to this, we went few more steps ahead and applied this software across Covid19 patients to reveal the miRNAome of Covid19 host about which there is almost no information available at present despite of the fact that such information is one of the most important

requirements for fight against Covid19. The Covid19 host miRNAome in turn has revealed highly valuable molecular information about this disease which may prove itself very critical in countering this pandemic which has put entire mankind under big threat.

Results and Discussion

miRNAs are mostly transcribed as pri-miRNAs where some are transcribed autonomously while some are host gene dependent [23,24]. In this study 1kb flanking region from both sides of published pre-miRNA sequences were considered as putative pri-miRNA as there is barely any resource available for pri-miRNA identification. To find out the possible binding sites of 155 RBPs across these miRNAs and their precursors, publicly available CLIP-seq reads were mapped across them and downstream studies were carried out at different levels.

Pri-miRNAs and Pre-miRNAs shared multiple RBPs binding sites

Total 1,230 CLIP-seq samples for 155 RBPs collected from ENCODE and GEO were mapped across the considered sequences of each known human miRNA using the protocol described in the methods section. After executing the mapping step, 138 RBPs qualified the criteria for interactions with pri-miRNA and 126 RBPs with pre-miRNA. The binding sites were clustered based on RBPs, pre-miRNAs and pri-miRNAs. The detailed statistics of binding sites of each RBP for different pri-miRNAs and pre-miRNAs is provided in the Additional file 1:Table S2. Detailed number of RBPs binding on each pri-miRNA and pre-miRNA is provided in Additional file 1:Table S3. It was noticed that SBDS, APOBEC3F, APOBEC3G, NSUN2, WTAP, METTL14, RBM47, SRSF3, RBM6 and TRA2B were ten such RBPs which had binding sites across pri-miRNA exclusive regions only. Out of 1,881 pre-miRNA 1,769 pre-miRNAs had binding sites for at least one RBPs, whereas 1,879 pri-miRNAs had binding sites for at least one RBP. Important RBPs in miRNA biology like DROSHA had binding sites in 787 pri-miRNA and 311 pre-miRNA, whereas DICER

had binding sites in 919 pri-miRNA and just 59 pre-miRNA, suggesting that their binding information may not be captured enough by CLIP-seq due to their fast endonucelolytic action over the substrate RNA.

It was noticed that each pri-miRNA exclusive region had binding sites for at least eight different RBPs. Similarly, most of the pre-miRNAs had binding sites for several other RBPs than those traditionally associated with miRNA biogenesis like DROSHA/DICER. Some RBPs showed positional preferences. DROSHA had higher binding sites in the pri-miRNA regions whereas AGO2 had bias for pre-miRNA regions. Cases like ACIN1 did not show any positional preference and binds uniformly in both primary exclusive and pre-miRNA regions. From the binding sites data, it was found that greater number of RBPs bound in pri-miRNAs than pre-miRNAs. But when normalized over length, only 424 pri-miRNAs out of 1,881 had higher number of RBP binding sites than their corresponding pre-miRNA. The normalized and original value of binding sites of number of RBPs on each pri-miRNA and pre-miRNA, along for each miRNA is provided in Additional file 1: Table S4. ELAVL1, AGO2, IGF2BP2, CPSF6, HNRNPC, FUS, FMR1, PTBP1, RNPS1, and CSTF2T have binding sites on more than 1,800 pri-miRNAs suggesting their widespread roles in miRNA timeline. Similarly, ELAVL1, AGO2, IGF2BP2, IGF2BP3, FUS, FMR1 and CPSF6 had binding sites on more than 500 pre-miRNA. Binding information for 30 RBPs on different pre-miRNAs were collected from the proteomics-based study reported by Treiber *et al.*, 2017 [25]. These 30 RBPs were common between both studies and were considered as high confidence cases by Triber *et al*. A total of 423 different combinations of RBP:pre-miRNA were collected from the cases reported by Triber *et al.* 386 of these combinations were found reported by our study also. The details of these combinations are given in Additional file 1: Table S5.

Binding sites of RBPs are strongly associated with their expression level

The dynamics of regulation by RBPs can be evaluated by relating the CLIP-seq data with expression data. For this purpose, one would require RNA-seq data to derive the expression levels of RBP genes and primary and precursor miRNA sequences. sRNA-seq reads evaluate the impact of RBPs on the expression level of the mature miRNAs. Combined consideration of these all in network form would help us to reconstruct the regulatory control system laid by RBPs for miRNA processing. The CLIP-seq data was available and collected for 253 experiments which covered 82 conditions. For RNA-seq and sRNA-seq data 47 experimental conditions were studied, and a total of 32 experimental conditions were common among RNA-seq, sRNA-seq and CLIP-seq data for 64 RBPs. For remaining RBPs and CLIP-seq conditions, no RNA-seq/sRNA-seq data was available. Therefore, there was a need to fathom if expression data of RBPs reflects its binding level, so that in the absence of CLIP-seq data for any given experimental condition, the expression data could reflect the binding potential of the RBP. In this study we had expression data for different RBPs, on the basis of which it was possible to draw some inference whether an RBP was binding or not in those 47 experimental conditions. To understand the association of binding site of RBP with its expression, CLIP-seq data and RNA-seq data for 73 RBPs were available commonly for any given experimental conditions. This set was analyzed for possible correlation between expression of RBP genes and their binding rate, collectively. The processed CLIP-seq data reads were mapped across the human genome (hg38) to locate the possible binding sites of the RBPs. To find out the expression of RBP in the particular condition, RNA-seq expression analysis was carried out following the protocol described in the material and method section. Total number of binding sites of RBP in each sample for the given experimental condition was counted. A correlation analysis was performed between the amount binding sites observed and expression of the RBP for each experimental condition (Additional file 3: Figure S5).

For strong association between number of binding sites and its expression, a correlation coefficient cutoff of >= 0.8 was considered. 60 RBPs out of 73 RBPs (82%) displayed significantly strong

correlation between the RBP expression level and its CLIP-seq binding level, which indicates that expression has usually a positive association with number of binding sites of RBP. Also, a Poisson regression analysis was performed considering number of binding sites as dependent variable and expression as independent variable for these 60 RBPs. From the analysis it was noticed that there exists a highly significant (p-values < 0.01) positive slope between amount of binding sites and expression in all experimental conditions for these 60 RBPs. The rank correlation value between number of binding sites and expression of RBP for all 73 RBPs in different experimental conditions is provided in Additional file 1: Table S6. For 13 RBPs it was noticed that the correlation coefficient was greater than 0.8 for some conditions, while in other conditions it was lesser than 0.8. To find out the reason behind their observed behavior co-expression network analysis was performed to find out possible auxiliary factors for such RBPs which may be responsible for such observed differences in relationship between binding and expression level. A very important observation about these RBPs was that they follow regulon model of regulation where genes participating in some common functions were found commonly regulated by some common factors [26]. In this study the regulatory chain for each RBP was considered where initially possible partners of each RBP were collected from STRING database [27]. Before performing co-expression network analysis, at first all possible partners of any given RBP were included in the network along with their expression level for the given experimental condition. Nine out of 13 remaining RBPs (ACIN1, HNRNPA1, HNRNPF, HNRNPU, MBNL2, TRA2B, IGF2BP2, FUS and ELAVL3) displayed correlation coefficient between expression and binding level greater than 0.8 in those conditions where expression was high and correlation coefficient less than 0.8 when expression was low. To check the significance difference of RBP expression in different experimental conditions, a t-test was carried out and was found to be significant at p-value <= 0.05. Therefore, the study suggests that these RBPs have strong correlation between their expression and their binding level reflected in CLIP-seq data when their expression is high. For other four RBPs (AGO2, HNRNPA2B1, METTL3 and NSUN2) it was noticed that there was no significant difference of

expression between experimental conditions, still there exists some correlation (i.e. correlation more than 0.8 in one experimental condition and less than 0.8 for other condition). In such cases there might be possibilities for involvement of other factors having synergistic or antagonistic association with the RBP which may act as the limiting factors for binding with RNA.

To find out the possible interactions between different RBPs, a co-expression network analysis was carried out for these four RBPs following the protocol described in methods section. For AGO2 from the network it was found that TNRC6A/GW182 (which is also an RBP) played a crucial role in the binding of AGO2 [28]. AGO2 interaction with RNA was also found dependent upon another factor LSM6. During binding of AGO2, LSM6 also co-expressed and appears positively regulating AGO2. The favorable condition for AGO2 binding appears to when TNRC6A displays a positive association with LSM6. LSM6 also displays a positive association with AGO2. The detailed network is presented in Additional file 3: Figure S6. TNRC6A, also known as GW182, interacts directly with the AGO2 and facilitates active miRNA repressor complexes (miRISC). TNRC6A is involved in RNA deadenylation.

HNRNPA1B2 associates with pre-mRNAs in the nucleus and appears to influence pre-mRNA processing. It is often associated with RNA metabolism and transport. HNRNPA1B2 is one of the most abundant core proteins of hnRNP complex and plays a key role in the regulation of alternative splicing. From co-expression network analysis, it was found that ELAVL1 plays a positive role in binding of HNRNPA1B2. However, at lower expression ELAVL1 does not show any positive association with HNRNPA2B1. There is an evidence of HNRNPA1B2 and ELAVL1 cooperation to inhibit protein translation [29]. The detailed network for HNRNPA2B1 regulation is presented in Additional file 3: Figure S7.

NSUN2 acts as a methyltransferase that catalyzes the methylation of cytosine to 5-methylcytosine (m5C) tRNA precursors [30]. It is involved in cell growth and chromosomal segregation. NPM1 is involved in different cellular processes, including protein chaperoning, centrosome duplication and histone integration transport. From the co-expression network analysis it was found that NPM1 has strong positive association with NSUN2 (Additional file 3: Figure S8). Sakita-Suto *et al* had reported strong association of NPM1 with NSUN2 during cell division [31].

METTL3 is involved in a N6-methyltransferase complex formation where it methylates adenosine residues at N6 position in RNAs and regulates various processes such as differentiation of embryonic haematopoietic stem cells, circadian clock, and primary miRNA processing [32-35]. N6-methyladenosine (m6A) plays a role in RNA stability, processing, translation efficiency and editing. METTL3 mediates methylation of pri-miRNAs, marking them for recognition and processing by DGCR8. To discover the possible co-factor for METTL3 a co-expression network analysis was performed. From the network it was noticed that GTF2H1 shows a positive association with METTL3, helping it in binding to RNA. Whenever expression of GTF2H1 increases, it appears to positively regulate METTL3 in binding to the RNA. The associated co-expression network is illustrated in Additional file 3: Figure S9.

Evaluation of the Bayesian approach for miRNA:RBP association network reconstruction

To decipher the involvement of RBPs and their PPI partners in miRNA biogenesis, a Bayesian Network Analysis (BNA) was performed by combining CLIP-seq derived information on binding sites of RBPs across different miRNAs, expression data of miRNAs, RBPs, and associated interactors. BNA was performed for all the 47 experimental conditions separately. To benchmark this BNA approach, we used the flow cytometry dataset[36] which is considered as a gold standard network test set. The original dataset consists of n = 7,466 observations of p = 11 continuous variables corresponding to different genes in human immune system. A network consisting of all

well-established causal interactions between these molecules was constructed based on biological experiments and literature. This network is used as a benchmark to assess the accuracy of Bayesian Network learning algorithms on real data. Performance was evaluated and compared with well-established established R-package "bnlearn". It was found that our approach was able to significantly recover more true edges, similar to bnlearn at a 95% confidence interval. For precision, our approach (91%) outperformed bnlearn (87%) which was statistically significant at a level of 0.05 by the two-sided paired sample t-test considering all edges in the network.

Also, bnlearn was run across all 47 experimental conditions taken in the present study to establish miRNA:RBP association. In each experimental condition there were more than 1,000 nodes. "bnlearn" failed to persist the good performance in large feature space (> 1000 nodes) whereas our approach performed steadily as more nodes were involved in the network with a precision of more than 85%. Our approach was capable of handling networks in both small and large feature regimes effectively. High precision achieved by our approach indicates that the networks recovered by it are reliable.

Network analysis reveals miRNA:RBP functional association following the Regulon model

After establishing the association between binding site and expression of RBP, a Bayesian network analysis (BNA) was performed to interpret the association of RBPs at different stages of miRNA biogenesis. RNA-seq and sRNA-seq data were collected for same experimental conditions to obtain expression profile of RBP, pre-miRNA, and mature miRNA. From RNA-seq data, pre-miRNAs and RBPs expressions were calculated. Mature miRNA expression was calculated using sRNA-seq sequencing data. Also, the expression data for the genes that were present in the regulatory chains of different RBPs were calculated from the RNA-seq data. The inputs for the Bayesian network analysis were expression data of pre-miRNA, mature miRNA, RBPs having binding sites in those miRNAs, and their PPI data along with corresponding expression data of the interacting genes. Any

particular RBP in the pre-miRNA:RBP interaction modeling was included if it had binding sites in the pri-miRNA and pre-miRNA. Similarly, any particular RBP was included in mature miRNA:RBP interaction modeling if it had binding sites in the pre-miRNA sequence for the corresponding mature miRNA. This analysis was followed for each experimental condition separately and the miRNA:RBP associations were obtained for miRNAs (i.e from pri-miRNA to pre-miRNA and pre-miRNA to mature miRNA processing direction) expressed in that particular condition. A total of 8,047 pre-miRNA:RBP (RBPs supposedly involved in processing of pre-miRNA from pri-miRNA) and 10,100 mature miRNA:RBP (RBPs supposedly involved in processing of mature miRNA from pre-miRNA) unique combinations were obtained considering all experimental conditions together. These pre-miRNA:RBP and mature miRNA:RBP combinations obtained from BNA were also checked based on expression correlation in each experimental condition. In each experimental condition more than 89% miRNA:RBP combinations (both pre-miRNA:RBP and mature miRNA:RBP) obtained from BNA were found showing similar type of association (positive/negative) with strong expression correlation (>=|0.6|) for any given experimental condition.

The miRNA:RBP combinations obtained from this network analysis were independently validated across eight different normal tissues viz. bladder, testis, brain, breast, lungs, pancreas, placenta and saliva. The main interest of this validation work was to find out if the observed behavior of miRNA:RBP associations in this study was similar across other tissues which were not included earlier in the current study. Here only the mature miRNA:RBP combinations were evaluated due to unavailability well established expression data of pri-miRNA v/s pre-miRNA expression data. Mature miRNAs expression data were collected from miRmine database [37] and RNA-seq expression data were collected from GTEx [38], ARCHS4 [39] and Array Express [40]. Correlation analysis was performed between mature miRNA expression and RNA-seq expression. Those combinations obtained for mature miRNA:RBP from the BNA were evaluated across all the eight

tissues considering an absolute correlation coefficient value (>=|0.6|). The detailed work-flow for this analysis and the distribution of unique miRNA:RBP combinations overlapping in different tissues is given in Additional file 3: Figure S10. It was found that 96.9% the BNA identified miRNA:RBP combinations were present in at least one of these eight tissues. Also, it was noticed that 49% of miRNA:RBP combinations were present in more than four tissues. There were 270 such miRNA:RBP combinations which appeared in all the eight tissues. To check the significant existence of miRNA:RBP combinations obtained in this study, a Fisher's exact test was performed using a total of 1,64,662 random miRNA:RBP combinations. The random miRNA:RBP combinations were searched against the considered eight tissues. Only 144 random miRNA:RBP combinations (0.09%) were found existing across the eight tissues. The details of distribution of miRNA:RBP combinations across these eight tissues for both the datasets is illustrated in Additional file 3: Figure S10. The Fisher's exact test was found to be highly significant (p-value <0.01) implying that the observed miRNA:RBP combinations across the eight different independent normal tissue were not random at all. Also, each combinations obtained in this study were tested separately using a binomial test against the miRNA:RBP combinations in random dataset. Out of 10,100 miRNA:RBP combinations 9,786 combinations (96.9%) were found significantly existing (p-value <0.01). The combinations which appeared in four or more tissues is provided in Additional file 1: Table S7 along with their p-value. To visualize the association of miRNA with RBP considering all the eight tissues, a corr-gram plot was prepared considering 50 randomly chosen mature miRNA (Additional file 3: Figure S11).

A hierarchical cluster analysis was performed for clustering of different miRNAs based on their expression data considering all the eight tissues together. A total of nine clusters were formed for these 50 randomly selected miRNAs. It was found that the miRNAs which clustered in one group shared more than 80% common RBPs which bound to their pre-miRNAs. This fitted the regulon model hypothesis very well [26]. A functional enrichment analysis for pathways, molecular

function, and biological process was performed considering those miRNAs belonging to a common cluster, for each cluster separately. Enrichment analysis was done for the miRNA targets belonging to any given cluster of miRNAs. It was noticed that miRNAs belonging to same cluster had at least 70% common pathways, biological process, and molecular function for each cluster. The miRNA members belonging to a cluster were found associated with high number of common RBPs, altogether strongly suggesting that they work towards some common functions and purpose. Further to this, to find if RBPs associated with a cluster of mature miRNA displayed similar functional enrichment, cluster specific functional analysis was performed. It was noticed that for all the nine clusters more than 68% common pathways, 76% common biological processes and 80% common molecular function were present. The number of common pathways, biological processes and molecular functions for these randomly selected 50 miRNAs and associated RBPs is illustrated in Figure 1. This study was extended for all miRNAs while considering complete data for all the eight tissues. miRNAs which expressed themselves in at least 50% samples were considered for clustering, resulting into a total of 632 miRNAs fulfilling this criteria. These miRNAs were clustered based on the combined expression data for all the eight tissues using hierarchical clustering algorithm. A total of 124 clusters formed in which minimum 80% RBPs were common for the associated corresponding miRNAs present in the cluster. Functional enrichment analysis was performed for both miRNAs target, and for the miRNA associated RBPs. For each cluster all enrichment categories were compared between miRNAs targets and RBPs. Similar pattern was observed for this full data set analysis also. It was found that out of 124 clusters, 96 clusters (75.9%) had more than 56% common pathways, 71% common biological process and 73% common molecular function between miRNAs targets and RBPs. From this overall analysis it emerged very confidently that the observed binding of RBPs across the miRNAs had enormous functional impact on the roles of miRNA, where formation of miRNA itself could be influenced by the binding RBPs.

From the available literature search 104 different combination of miRNA:RBP evidences were collected which included both RBP association with miRNA at pri-miRNA to pre-miRNA processing level and at the level of mature miRNA processing from pre-miRNA. A total of 85 out of 104 different such literature supported combinations were successfully detected in the present study, further validating the observations (43 combinations pri-miRNA to pre-miRNA and 42 combinations from pre-miRNA to mature miRNA). Those miRNA:RBP combinations having literature support for miRNA biogenesis are provided in Table 1.

Contribution of different RBPs at different level of miRNA biogenesis

To study the potential involvement of different RBPs in miRNA biogenesis, the different combinations with RBPs for pri-miRNA to pre-miRNA and pre-miRNA to mature miRNA reconstructed by BNA were analyzed. The shares of potential contribution of each RBP at pri-miRNA to pre-miRNA and pre-miRNA to mature miRNA processing levels were evaluated based on the unique combinations observed. At primary to pre-miRNA processing level, a total of 8,047 pre-miRNA:RBP combinations existed where the most abundant RBPs were ACIN1, DROSHA, and DGCR8. ACIN1 has RRM domain and is mainly involved in splice site recognition and alternate splicing via spliceosome. Out of total 8,047 pre-miRNA:RBP combinations ACIN1 shared 4.54% combinations while covering 365 pre-miRNAs forming from pri-miRNA. Similarly, DGCR8 shared 4.5% and it covered 362 pre-miRNAs formation from pri-miRNA. DROSHA shared 4.05% and it covered 326 pre-miRNAs from pri-miRNAs. The details of pre-miRNAs processed by each RBP from pri-miRNA is listed in Additional file 1: Table S8. For mature miRNA processing from pre-miRNA a total of 10,100 mature miRNA:RBP combinations were obtained from the current study. RBPs apparently responsible for processing of mature miRNA from pre-miRNA had abundance for IGF2BP3, AGO2, FMR1, and FUS. Out of total 10,100 unique combinations, IGF2BP3 contributed 5.26% combinations covering 531 mature miRNA, AGO2 contributed 3.9%

covering 393 mature miRNA, FMR1 contributed 3.38% covering 341 miRNA, and FUS contributed 3.21% while covering 325 miRNA. The details are provided in Additional file 1: Table S9. IGF2BP3 is mainly involved in RNA stability and its localization to cytoplasm. IGF2BP3 is available in both cytoplasm and nucleus. It has two RRM domain and four KH-domain. All KH domains contribute binding to target RNA. FMR1 is mainly involved in RNA stability, RNA transport and splicing. As already discussed, transportation and splicing are two important activity in mature miRNA processing. FMR1 has two Agenet-like domain and two KH-domain. The KH domains are necessary for mediating miRNA annealing to specific RNA targets [49]. FUS belongs to FET family having a RRM domain, mainly involved in RNA stability and RNA splicing [50].

AAR2 controls a large number of DICER independent miRNA formation

Besides the direct involvement of RBP in miRNA biogenesis, the current study made a very interesting observation. AAR2, an RBP, was found highly involved in the processing of pre-miRNA from pri-miRNA covering 467 miRNAs (5.8%). It was found involved in mature miRNA formation pre-miRNA while covering 781 miRNAs (9.7%). Also, AAR2 expressed more than that of well-studied miRNA biogenesis associated RBPs like DICER and DROSHA in most of the tissues. Not much is known about AAR2. To gain more insight about this RBP, structural analysis of AAR2 was done following protocol of homology modeling to predict its structure from template with >20% sequence identity as structure of AAR2 is not available in PDB. After that model was built using Swiss-Modeller with best model having 99.9% query coverage. The best template derived was PDB id-4ilj.1.A chain. To draw inference from structure to function, domain knowledge was mandatory which was collected from InterPro. This protein contains pre-mRNA splicing factor-8 both in A and B-chains of its trimeric structure and belongs to ribonuclease-H superfamily. Its C-chain belongs to cistron splicing factor. Its structural detail and domain information is provided in Additional file 4: Figure S12.

With above observations where AAR2 appeared very critical in miRNA biogenesis, it was decided to experimentally validate its found association with miRNA biogenesis. Randomly two AAR2 associated miRNAs were considered, viz. hsa-miR-25-3p and hsa-miR-206, for validation of their observed independence from DICER and dependence on AAR2 for their formation. To assess this process of maturation of miRNAs, cells with diminished DICER and AAR2 expression levels were created using shRNA mediated knockdown of DICER and AAR2. shRNA plasmid constructs were transfected in the CAL27 cells and grown in appropriate selection medium. Fluorescence microscopic imaging analysis suggested the prominent expression of GFP reporter protein in transfected cells (CAL27 sh) as compared to non-transfected control (CAL27 NT) (Figure 2 (A,D)). In addition to this, western blot analysis was performed for both the groups to check the protein expression levels of GFP and DICER (Figure 2 (B)). Results suggested significantly diminished expression of DICER in the shRNA1(sh1), 2 and 3 as compared to non-transfected control and scrambled control (SC) groups . The findings clearly evidenced and validated the generation of DICER knockdown ($^{-/-}$) in CAL27 cells. Along with it, results obtained from fluorescence microscopy implied the transfection of AAR2 in cells in sh3 group as compared to non-transfected control and scrambled control (SC) groups and generation of AAR2 knockdown ($^{-/-}$).

The quantification of expression levels of mature miRNAs, qPCR assays were performed using cDNA converted out from miRNAs. Results suggested almost basal level expression and definitely no down regulation in expression of the two mature miRNAs post DICER knockdown (Figure 2 (C)). It clearly implies that the knockdown of DICER did not affect the maturation of AAR2 associated miRNAs in CAL27 cells (Figure 2 (e)). Furthermore, the same set of miRNAs were found down regulated in the AAR2$^{-/-}$ cells in qPCR assays. It suggested consonance in our findings that maturation of the selected miRNAs is independent of DIECR and dependent on AAR2.

miRNA:RBP associations can be classified as synergistic or antagonistic

The miRNA:RBP combinations obtained were categorized based on correlation (Significant-positive/Significant-negative /Non-significant) between miRNA and RBP, and RBP with RBP. There were 27 types of different miRNA:RBP combinations obtained for both pri-miRNA to pre-miRNA and pre-miRNA to mature miRNA processing levels. All 27 different combinations for both the stages of miRNA biogenesis are provided in Additional file 1 : Table S10 and Table S11. Due to multiple RBP binding sites on same miRNA, it was noticed that both competitive and cooperative associations exist between RBPs and miRNA in spatio-temporal manner for each stage of miRNA biogenesis. For example, during pre-miRNA processing from pri-miRNA of hsa-mir-4477b, the RBPs involved are CPSF2 and MOV10. In the pri-miRNA sequence it was found that CPSF2 and MOV10 were having overlapping binding sites. With the help of expression data, it was noticed that both RBPs have positive association with the processed pre-miRNA, but there exists a negative association between CPSF2 and MOV10 i.e. they have antagonistic association among themselves. hsa-mir-4477b expressed in nine experimental conditions out of the considered 47 experimental conditions. CPSF2 was found to process hsa-mir-4477b in four experimental conditions from its pri-miRNA sequence, whereas MOV10 did so for the remaining five experimental conditions, clearly showing mutually exclusive behavior. It was found that both these RBPs had a competition for binding for a common spot and also both RBPs had similar type of biological roles, mainly involved in RNA splicing and RNA stability. Similarly, during processing of hsa-mir-6723 from its primary the RBPs involved were U2AF2 and DDX42. When these combinations were studied across the considered 47 experimental conditions, it was noticed that the pre-miRNA was processed in 15 different tissues where both RBPs had no association with each other but had a strong positive association with the pre-miRNA. This may be possible that these two RBPs work at different times in pre-miRNA processing from its pri-miRNA. Also this was noticed that pre-miRNA was processed where both RBPs were expressed. The pre-miRNA was not processed in those

experimental conditions where only one RBP was expressed, indicating that both RBPs have synergistic association in pre-miRNA processing where sequential timing appears playing some role. Case like hsa-mir-3918 processing from its primary, HNRNPA1 and ACIN1 displayed positive association with each other. Also, it was noticed that both RBPs were essential for pre-miRNA processing, where as in the absence of anyone of them the mature miRNA expression was lowered.

During mature miRNA processing from pre-miRNA also RBPs displayed synergestic and antagonistic associations among themselves. For example, during processing of hsa-miR-619-5p from hsa-miR-619, three RBPs were found significantly involved: CPSF3, MOV10 and IGF2BP3. In the pre-miRNA sequence, it was found that CPSF3 and MOV10 had overlapping binding sites on hsa-mir-619. With the help of expression data, it was noticed that both RBPs have positive association with the mature miRNA, but there exists a negative association between CPSF3 and MOV10 i.e. they have antagonistic association among themselves. Out of the 47 experimental conditions, hsa-miR-619 expressed in 13 different conditions. Out of these 13 conditions, seven conditions had CPSF3 potentially involved in mature miRNA processing, whereas MOV10 had a negative correlation with CPSF3 in these seven different experimental conditions in mature miRNA processing. In the rest of six experimental conditions MOV10 was highly expressed than CPSF3 and was potentially involved in processing of mature miRNA from pre-miRNA. It was negatively correlated with CPSF3 in these conditions. In all the thirteen conditions IGF2BP3 expressed and it had a role in mature miRNA processing. It showed positive association with both the proteins in their respective control states, but in mutually exclusive manner as CPSF3 and MOV10 competed for overlapping binding region. Similarly, during processing of hsa-miR-18a-3p from its precursor, the potential RBPs involved in mature miRNA processing were RBM22 and IGF2BP3. It was noticed that the mature miRNA was processed in 11 different experimental conditions where both RBPs had no association among each other but had a strong positive association with the mature miRNA. This may be possible that these two RBPs work at different time in mature miRNA

processing from its pre-miRNA. Also, it was noticed that mature miRNA was processed where both RBPs were expressed. The mature miRNA was not processed in those conditions where only one RBP was expressed, indicating that both RBPs have synergistic association in mature miRNA processing. In case like processing of hsa-miR-3175 from hsa-mir-3175, FXR1 and RBFOX2 displayed positive association with each other. Additionally, it was also noticed that both RBPs were essential for mature miRNA processing, where as in the absence of anyone of them the mature miRNA expression was lowered. Figure 3 illustrated the cooperative and competitive associations between RBPs for the above discussed cases.

RBPs can act as bio-markers similar to their associated miRNAs

An attempt was made to distinguish disease and normal conditions on the basis of RBP expressions and their associated miRNAs separately to see if both can replace each other as markers as formation of miRNAs clearly looked dependent on these associated RBPs. The miRNA:RBP combinations obtained from BNA in different experimental conditions were collected. The expression data of miRNA and their associated RBPs for their associated experimental conditions were collected together and K-means clustering was performed to differentiate between the experimental conditions (i.e disease v/s normal state). For this discrimination between samples of different experimental conditions, six different combinations of RBPs and miRNAs were studied separately. In the first step the experimental conditions were classified considering the expression data of total miRNA expressed in both experimental condition. In the second step the two conditions were distinguished considering only expression data of the RBPs which were found associated with the considered miRNAs. In the third step those miRNAs common in both conditions were discarded and the remaining miRNAs were considered to distinguish the states through the clustering. In the fourth step those RBPs associated with the common miRNAs present in both condition were removed and clustering was done with the remaining associated RBPs. In the fifth step the experimental conditions were classified combining the expression of total miRNA and

their associated RBPs. In last step the miRNAs after discarding common miRNAs in both the condition and their associated RBPs were considered in clustering the instances of the experimental conditions. These tests were done to measure how effectively the clustering between different experimental conditions was possible when considering the miRNAs and their associated RBPs as biomarkers.

These six combinations were tried to distinguish between lungs cancer state and normal lung tissue condition. List of miRNAs and their associated RBPs were collected for both conditions. Considering the expression data (Z-score) of RBPs and miRNAs the disease and normal samples were classified using a K-means clustering for all the six combinations of miRNAs and RBP. The miRNAs based clustering sharply clustered the cancer and normal samples. The RBPs associated with the miRNAs displayed the similar clustering. After excluding the common miRNAs in both experimental condition, when the miRNAs were combined with their associated RBPs and considered together, the clustering improved further. The plots for K-means clustering of all six combinations of miRNA:RBP is given in Additional file 4 : Figure S13. Similarly, observation was made when clustering was performed based on the expression data of miRNA and their associated RBP for normal and thyroid cancer tissues. It was pretty evident that the associated miRNAs and RBP cluster similarly and can replace each other. This also underlines that functional association is well reflected in this similar clustering patterns between miRNAs and associated RBPs.

Machine learning model of RBP conditional networks successfully predicts miRNA profile

In the above sections, multiple validations and testings made it very clear that miRNA biogenesis is dependent upon combinations of various RBPs and their conditional networks reasoning the spatio-temporal expression of miRNAs beyond the usual implications of routine enzymes like DROSHA/DICER. If these observations and hypothesis hold so much true then they can also be successfully implemented computationally to simulate and predict the miRNA profile for any given

condition just using the expression data for the network components, without needing any expression data for miRNAs. Such tool becomes very important to work for conditions where miRNA profile is unknown and cost cutting on running miRNA profiling experiments is desired. More so when large number of samples and conditions are to be studied. To implement such tool a machine learning approach, XGBoost regression was used. It is an extreme gradient boosting ensemble technique where prediction is done by an ensemble of simple estimators giving higher weights on difficult learning instances to minimize the loss function using gradient [51]. The RNA-seq and sRNA-seq expression data were collected from TCGA database for seven different tissues such as bladder, brain, breast, cervix, esophagus, kidney and head-neck (containing both normal and cancer conditions) for model building and testing. Different types of machine learning (both shallow and deep learning) regression techniques were tested and XGBoost regression approach outperformed others with consistence and an average 91% accuracy (Accuracy range: 87(lowest) – 94%(highest) for the modeled miRNAs (1,204). Each dataset was randomly shuffled and split into to a training and testing dataset in 70:30 proportion, respectively. The prediction rule was fine tuned by identifying the optimal combination of hyper-parameters that further minimized the objective function. With the optimal values of the hyper-parameters and number of trees, the regression was retrained and applied to the withheld testing data to predict a new series of miRNA profiles and evaluate the accuracy of the model. The model accuracy was tested based on the RMSE (Root mean square error) and RMAPE (Relative Mean absolute percentage error). Where as to check the model consistency 10-fold cross-validation was implemented with consistency measured on the basis of RMSE. The 10-fold RMSE was always found consistence (not much fluctuation in RMSE) when compared to the model RMSE which indicated the consistency of these models. The observed and predicted expression levels obtained from the XGBoost regression model for six miRNAs as examples is given in Additional file 4: Figure S14. Further level of testing was done for the developed models across 431 different samples covering different conditions which included different tissue types (both normal and cancer), such as liver, prostrate, pancreas, ovary, lungs

LAUD, lungs LUSC, adrenal gland ACC and adrenal gland PCPG, which were not covered in any of the studies described above. Out of 431 samples 308 (72%) scored more than 90% accuracy, while 118 samples (27%) had more than 85% accuracy. The high degree of testing score consistency high accuracy and stability in performance confirmed a very reliable to the developed tool to profile miRNAs in any given condition without any need to do miRNA sequencing or profiling experiments. The developed tool has been also implemented at a companion web server available at https://scbb.ihbt.res.in/miRbiom and will also be made available at https://scbb.ihbt.res.in/SCBB_dept/Software.php. More details are given about it in the following section.

A companion server for RBP:miRNA interaction study and miRNA profiling

A companion server with name of miRbiom has been provided here which can be seen into mainly two parts: 1) miRbiom miRNA profiler software as discussed above, and 2) Information and analytics portal. The above mentioned tool to discover the miRNA profile without any need of sRNA-sequencing data has been implemented here as a software server. User needs to profile the RNA-seq data for any given condition. This data is run through the miRNA biogenesis models implemented through RBP:miRNA conditional networks based implementation of XGBoost regression, which generates a relative regression score for various miRNAs capturing the potential expression profile of miRNAs for the given condition. It generates a plot of expression profiles of various miRNAs in interactive fashion. Selections can be made here to study the miRNA targets for their functional enrichment as well as pathways analysis. miRNA target information from various databases like miRTarbase etc has been provided. Provisions have also been made to map the miRNA targets in collective fashion and view them in KEGG pathways maps. The implementation details of this part is given in Additional file 4: Figure S15.

The objective of this study was to identify miRNA:RBP associations and their regulatory impacts while considering different types of high throughput NGS data. While doing so, CLIP-seq data, RNA-seq and sRNA-seq data were collected from different platforms. These data were studied for relation with each other and were connected accordingly. Proper structuring and handling is necessary to derive meaningful and relevant information. In this regard, a number of visually rich and useful representations as well as larger supplementary data have been made available at the associated portal where a user could explore into the further details. The website was build using advanced web development packages viz. HTML, Java-script, CSS and PHP. In this webpage, chart and table were build using libraries like D3, Plotly, JSON, Amcharts, Highcharts, jquery, sunburst-chart, circlepack chart, icical chart, bootstrap, ajax, jszip, pdfmake and vfs fonts. A user can search important key elements of this study through navigation bar. This server describes the number of RBPs binding across different miRNAs (both pri-miRNA and pre-miRNA) by an interactive bar-chart where by clicking on each miRNA at the X-axis a new window opens displaying information on the binding positions of different RBPs for that miRNA. For CLIP-seq data a sunbrust-chart was used for visualization, where lower most circle represents root circle. By clicking on the root circle one can get whole CLIP-seq data description. An interactive heat-map chart was created to display the RBP association with mature miRNA processing from pre-miRNA, which is an alternative and interactive representation of the correlogram used in the manuscript. An interactive nested pie chart was created using amChart library which illustrates the contribution of different RBPs in each stage of miRNA biogenesis. Besides these all, there are other interactive visualization and representation of the data arising from the study. From the supplementary page, user can select curated tables with multiple filter support and interactive provisions. All supporting materials of this study are hosted here.

Application: COVID19 patient miRNAome profile and system discovery using miRbiom

As transpired from the above sections now it is possible to discover the miRNA profiles for any given condition despite of no sRNA-seq experiment. Such tool becomes very important in situations like infectious diseases where no molecular information is easily available. The living example is current global health emergency caused by SARS-CoV-2 pandemic. Till the date more than 6 million Covid19 cases have been reported and this number is just spiraling up, putting huge threat to human civilizations. The emergence of the novel human corona-virus SARS-CoV-2 in Wuhan, China has caused a pandemic of respiratory disease (Covid19). So far no effective drug and vaccine have been found to deter it. The big scientific concern is that to this date very scarce and uncertain molecular information is available about the Covid19 patient's molecular system as not much high-throughput studies have been carried out so far. There is almost absolutely no information on host miRNAs response during Covid19 infection. Availability of such molecular information is perhaps most important to understand Covid19 infection system and device therapies targeting the suitable targets. It is just an irony actually not much work has been done in a direction which would help the most in the combat against Covid19. However, we found two RNA-seq studies carried out on Covid19 patients' lung sample which could be immensely helpful to reveal the miRNAome of Covid19 patients and molecular system specific to it. Using these patient samples we run the above mentioned tool to get the potential miRNAome profile of the Covid19 patients lung samples for the first time ever. The study also revealed the system impact of the Covid19 specific miRNAome which will be an indispensable resource in devising therapeutic strategies against Covid19.

RNA-seq data of four covid19 patients (accession number: CRA002390) and six control samples were collected from BIG Data Center (https://bigd.big.ac.cn/). In this study transcriptome sequencing was done for the long RNAs isolated from the bronchoalveolar lavage fluid (BALF) and peripheral blood mononuclear cells (PBMC) specimens of Covid19 patients [52]. Only lungs tissue specific data were considered in this study. RNA-seq data from another study over Covid19 patients

lungs biopsy was collected from GEO (GSE147507) [53]. Gene expression analysis was performed for both studies using standard protocols described in the methods section. The gene expression profiles for Covid19 patients were used as input to the miRbiom tool for miRNA profile discovery. In the BALF data from Wuhan, a total 498 miRNAs was discovered among which 378 miRNAs were up-regulated and 120 were down regulated. In the lungs biopsy data a total of 486 miRNAs were found differentially expressed, in which 405 were up-regulated and 81 down regulated. The up and down regulated miRNAs were decided based on their $\log_2$ fold-change values. We found that 360 miRNAs were common across both the independent studies, displaying a good agreement and high confidence on the found miRNAs involved in Covid19. Out of these 360 common miRNAs 326 miRNAs were up regulated and 34 were down regulated in both the studies (Figure 4(A)). The list of miRNAs (up and down regulated) obtained in both the studies and the common set of miRNAs obtained in both the studies are provided in Additional file 1: Table S12.

In the next step, miRbiom identified the high confidence targets of such miRNAs using experimentally validated miRNA targets data reported in databases like miRTarBase database [54] for both up and down regulated miRNAs for both the studies separately. The identified targets were further supported with anti-correlation between the miRNA and target gene based on their expression. Those cases which scored the expression anti-correlation value of 0.7 or more were considered for further study. The top twenty miRNAs based on their expression and target are given in Figure 4(B). The detailed list of number of targets obtained for each miRNA in both study is provided in Additional file 1: Table S13. The combined list of unique anti-correlated target genes were analyzed for pathways, biological process, molecular functions and sub-cellular locations using Enrichr [55]. The enrichment analysis was performed for individual miRNA targets as well as combining all miRNA targets together. Those pathways and Gene ontology termed significant at P-value <= 0.05 were considered in this study. The significant biological processes and molecular functions were ranked based on their occurrence across different miRNAs along with number of

genes reported in Additional file 1: Table S14. The pathways enrichment was done for three different databases *viz*. KEGG, Wiki, and PantherDB [56]. Mostly of the pathways were found overlapping with each other. The top 20 pathways for each database is given in Figure 4 (D, E, F). The detailed gene and miRNA list for each pathways obtained is given in Additional file 1: Table S15. The most important pathways which were found being affected in Covid19 condition are mainly for targets for the up-regulated miRNAs. The most significantly affected pathways my over-expressed miRNAs were related Apoptosis, FoxO signaling, Insulin signaling, EGF/VEGFA-VEGFR2/Angiogenesis, Sphingolipid/PDGF signaling, Interleukins, and CCKR signaling. Many of these pathways are reported to be involved in cancer were found enriched. Figure 4 (D, E, F) provides a good idea of the functional associations. The list of miRNA possessed number of pathways is provided in Additional file 1: Table S16. The miRbiom tool also provides the mapping of genes of interest in KEGG pathways. All the genes were also classified for the protein class whose distribution is given in Figure 4(G). In terms of protein class, the most dominant class of target genes down regulated by miRNAs in Covid19 are protein modifying enzyme (13.70%), more specifically protein involved in ubiquitinylation process, gene-specific transcriptional regulator and transcription factors, RNA binding and processing proteins, extracellular matrix proteins, transmembranal proteins like GPCRs, and proteins involved in defense and immunity, specifically IG-gamma receptor. A very striking observation was that a huge number of Zinc finger family proteins related to the process of ubiquitin-proteasomal pathways were found down-regulated by miRNAs over-expressed in Covid19. For more detail distribution of protein class are provided in Additional file 1: Table S17.

These are very detailed findings about SARS-CoV-2 infected patient which could be possible using the findings made here on how RBP associated networks control miRNA biogenesis. The findings made here will be a huge resource for further detailed study to design strategies and therapeutic interventions against SARS-CoV-2 infection which is currently not in the present scope of this

work. Yet, the system which attracted our attention the most for Covid19 pathophysiology is the cross-talk relationship between Insulin/IGF/AKT/mTOR signaling pathways, IFN-Gamma related pathways, and associated Ubiquitin-Proteasomal pathways which were connected and were found immensely being down-regulated by miRNAs in Covid19 infection, resulting into compromise of immune system. This axis of cell control system appears very important to device therapeutic studies to counter SARS-CoV-2 infection. A snapshot of this axis is given in Figure 5. The Insulin related aminopeptidase (IRAP) and IFN-gamma were found central to this axis. IRAP is found critical for glucose transport by GLUT4 and is involved in several important functions. Also it is a receptor of Angiotensin-IV which blocks IRAP. Chains coming to IRAP were even obstructed with best example of Insulin receptor, RTK, PI3K, AKT like crucial genes being down regulated, halting PI3K-AKT-mTOR signaling system influencing glucose uptake system controlled through IRAP-GLUT4, as well as obstructing availability of IRAP for antigen processing from viral proteins, which in turn compromises MHC-I based immune system as IRAP-MHC class-I work together towards antigen processing and presentation to T-cells. This all is also being regulated with IFN-gamma which was found down-regulated along with its receptor, severely compromising immune response, including IRAP-MHC-I combined function for immunity. Both these systems, Insulin/IGF/IRAP based steps and IFN-gamma led MHC class-I immune system heavily depends upon Ubiquitin-proteasome pathway genes, a large number of which were found down-regulated with many of them belonging to TRIM gene family which is essential for proper working of immune system. IFN-gamma system is critical to control a large number of genes related to immunity at transcriptional level, which includes several genes of Ubiquitin-proteasome pathways itself, PI3K-AKT-mTOR pathway genes, and includes IRAP and MHC-I themselves. IFN-gamma is also at the root of NF-kappa directed hypoxia related response and controls HIF/PHD/Ubiquitinylation cycle as well as attenuates HIF through repressing HIF-beta. The Insulin/Akt/mTOR also cross-talks here with HIF-1/PHD cycle. PHD, an oxygen and iron quencher, which degrades HIF-1 gene through ubiquitin-proteasome pathway was also found down regulated.

The study strongly suggests the explore this suggested axis of control for Covid19 therapeutic solution. Detailed data related to this part of the study is given in Additional file 1: Table S18.

Conclusion

miRNAs have emerged as major regulatory controller of cell systems. For a long time it has been believed that there are certain RNA binding proteins on which formation and action of miRNAs depend. While on the other side of this all, very recently importance of other kind of post-transcriptional regulators, the RNA binding proteins, has started to emerge. Till date more than 1,500 RBPs have been identified in human and it appears that this number would go much higher as recent findings are reporting that RBPs are breaking old perceptions about proteins. Enzymes long believed to be involved in metabolic processes are emerging as RBPs. Unlike long perceived belief that RNA binding domains are characteristics of RBP, recent findings are implicating intrinsically disordered regions in majority. Myths are breaking and so is going to happen with our understandings about the process of miRNA biogenesis which was mostly seen as an event managed by DROSHA and DICER proteins. miRNA precursors have been found to possess binding sites for various RBPs which could be the reason for spatio-temporal expression of miRNAs despite of ominous expression of DROSHA/DICER genes. With arrival of various next generation sequencing techniques, this is now quite possible to reveal this facet of regulation where these two major post-transcriptional regulator come together for biogenesis and action of one of them, miRNAs. Experiments towards identifying more RBPs and their interactions is therefore very much the need of the time. The present study has systematically delved into a huge amount of such NGS data to reveal the relationships which could regulate miRNA formation. Multiple layers of evidences have been provided for these associations in the present study. An RBP, AAR2 is being reported first time for its role in miRNA biogenesis. It was found highly associated with miRNA biogenesis and was subsequently validated experimentally that it is critical for miRNA biogenesis. Using the findings from the study a machine learning based tool was developed which could

identify miRNAs profile for any given condition with very high accuracy and consistency. Such software will be highly useful in dealing with conditions where miRNA information is not available or one desires to skip sRNA-seq sequencing for cost cutting. An apt application of this tool was made to reveal the first ever miRNAome and its regulatory impact information for SARS-CoV-2 (Covid19) infected host, releasing plethora of valuable information. The entire process of miRNA biogenesis appears a highly concerted and contextual process where RBPs combinations may be the deciding factor. A system level view has helped us in understanding this. The findings made here will impact at large scale on our fundamental understandings about cell system and their controls, which will have far reaching outcomes.

Methods

Source of NGS Data and Data processing

Data for sRNA-seq and RNA-seq based high throughput studies were collected from Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA) for 21 experiments (data volume ~15.6TB) which included 47 different experimental conditions. Complete list of various experimental conditions (for both RNA-seq and sRNA-seq) and source of data along with detailed experimental description are available in Table 2. The miRNA sequences of human (1,881 pre-miRNAs, 2,588 mature miRNAs) were collected from mirBase database version 21 [2]. 1,230 CLIP-seq experimental samples were collected for 155 RBPs from ENCODE and GEO database (~10.8TB). These data were derived from different types of CLIP-seq experimental techniques such as: PAR-CLIP, HITS-CLIP, eCLIP, iCLIP, CLEAR-CLIP and irCLIP. The detailed description of CLIP-seq datasets is given in Additional file 1:Table S19.

RNA-seq data was filtered using filteR [57] and sRNA-seq data was processed using trimometric [58]. For mapping RNA-seq reads across the human genome build 38 assembly (hg38) Seqmap [59] was used which is based on Bowtie platform. The alignment results were saved in SAM format for expression quantification. rSeq [59] was used for quantification of gene expression from SAM files. For expression analysis of pre-miRNAs, same RNA-seq pipeline was used where RNA-seq reads were mapped over known pre-miRNA sequences collected from mirBase. The alignment results were stored in a SAM file. rSeq was used for quantification of pre-miRNA expression from RNA-seq data. mirDeep2 [60] was used for expression analysis of mature miRNAs from sRNA-seq data. For mapping sRNA-seq reads on known mature miRNAs, mapper.pl script was used. Quantifier.pl script was executed for quantification of mature miRNAs expression from the saved SAM files. The detailed pipeline followed for expression analysis in pre-miRNA, RBPs and mature miRNA is presented in Additional file 4 : Figure S16.

Out of 155 RBPs, CLIP-seq data of 46 RBPs were collected from ENCODE database. CLIP-seq data for 109 RBPs were collected from GEO database. All these raw reads were processed following a standard protocol (Additional file 4 :Figure S16). In this study only those RBPs were considered which had at least two samples reported. Out of 155 RBPs, 12 RBPs data were from studies done on only one sample. These RBPs were discarded from the current study. For binding sites of RBP on different miRNAs a criteria was set that at least five reads should map to the considered region of miRNA and at least two different samples must support it. After considering these stringent criteria, we were left with only 138 RBPs, which were considered in the current study.

Till now there is limited information available regarding pri-miRNA and very few tools are available for pri-miRNA identification. Therefore, pre-miRNAs collected from mirBase database

were extended upto 1kb flanking region on both sides. Such regions were considered as putative pri-miRNA in our study. For processing of CLIP-seq data FastX (http://hannonlab.cshl.edu/fastx_toolkit/) was used. The adapters reported in literature for respective CLIP-seq data were used for filtering. To remove reads with lower quality, only those reads were kept which had at least 75% of bases with a quality score of 25 or more (-Q33 -q 25 -p 75). Unique reads were selected. The unique CLIP-seq reads were mapped across pri-miRNA and pre-miRNA regions using Bowtie considering a maximum of two mismatches to get possible binding sites (-f -S -n 2 -a). The binding sites were further normalized using the criteria discussed above to get high confidence miRNA:RBP interaction sites. These binding sites were clustered for each RBP and associated primary and pre-miRNAs to get a clear statistics of binding sites distribution for each RBPs across different miRNAs and number of different RBPs binding to any particular miRNA.

Expression and network analysis

The CLIP-seq data collected for 138 RBPs did not cover all experimental conditions for RNA-seq and sRNA-seq data. The CLIP-seq data collected for 138 RBPs has 82 experimental conditions. The data collected for RNA-seq and sRNA-seq covers 47 experimental conditions. Total 32 experimental conditions were common among all the three types of high-throughput data which covered 64 RBPs. Therefore, there was a need to establish a general relationship between binding sites and expression of these RBPs, which would enable us to use the binding sites of RBPs on miRNA for each experimental condition even in the absence of CLIP-seq data. In order to observe if any correlation exists between the binding site density and RBP expression, CLIP-seq data and RNA-seq data in same experimental conditions were collected for the 73 RBPs from GEO and ENCODE databases. These CLIP-seq data were processed following the protocol described previously in Additional file 4 :Figure S16 and mapped on human genome (hg38) to find possible binding sites for that particular RBP in different experimental conditions. Similarly, RNA-seq

expression analysis was performed to quantify the expression of RBP in the given experimental condition. The association between number of binding sites and expression of different RBPs in different experimental conditions were studied based on rank correlation analysis. Due to smaller number of samples (3-7) for each considered experimental condition it was decided that a high correlation coefficient value of 0.8 or above would be a better choice.

It was found some RBPs showed distinctive behavior in correlation *i.e.* some conditions had correlation greater than 0.8, where as in other conditions it was less than 0.8. The difference in correlation of same RBP in different experimental conditions might be due to interference of other auxiliary factors (other proteins/RBPs) which impact the expression and binding of that particular RBP in any given experimental condition. To address this distinctive behavior, possible partners were collected for each RBP from STRING database [27]. To trace out possible interaction (synergistic/antagonistic) a co-expression network analysis was carried out using an in-house developed R-script. Based on the expression data of RBPs and their partners, positive and negative association were evaluated from co-expression network. To perform co-expression analysis, initially a correlation matrix was constructed based on expression data of RBP and its protein-protein interaction (PPI) partners. Adjacency matrix of co-expression network was created based on the correlation matrix of RBP and its PPI partners. In the next step those edges having correlation coefficient value lesser than 0.8 were removed. From the remaining edges of adjacency matrix, network was constructed using Prim's minimum spanning tree (MST) algorithm. MST was used to accommodate multidimensional gene expression data. The main idea behind this representation is that each cluster of the expression data represent to a sub-tree of the MST, which converts a high-dimensional clustering problem to a tree partitioning problem. The co-expressed modules were identified based on the edge betweenness property. A module is regarded as a densely connected subnetwork within a larger network. In biological networks, the genes belonging to a common module are more likely to share same common properties or play related roles towards some

molecular process. By dividing the networks into modules, a large system can be reduced and more precisely the roles of its components can be deciphered in a relevant manner. The detailed protocol followed for coexpression network analysis is illustrated in Additional file 4: Figure S17. The protocol used for co-expression network analysis was developed using "vsn", "igraph" and "genefilter" R packages.

Reconstruction of miRNA:RBP association using Bayesian network analysis

Bayesian network (BN) is a type of probabilistic graphical model that describes conditional dependencies of a set of variables through a directed acyclic graph (DAG). The structure of a DAG represented by two sets: (I) the set of nodes which represent random variables, and (II) the set of directed edges. The directed edges in a BN structure model the dependencies between the nodes. Bayesian network models alter with respect to assumptions about the local probability distribution. As our data (expression data of RBP, pre-miRNA and mature miRNA) continuous in nature, it has assumed a multivariate Gaussian distribution for all nodes through out the study. For a BN, probability is more epistemological which define its belief on the occurrence of an event. This belief is known as prior probability which derived from its previous experience. BN utilizes Bayes theorem to accumulate the prior probabilities and likelihood from the observed data to obtain the posterior probability of the event. Posterior probability is the updated belief on the probability of an event happening given the prior information and the observed data.

We hypothesize that the mature miRNAs result from the concerted action of RBPs and associated factors at each stage of miRNA biogenesis. To study the involvement of potential RBPs and their PPI partners in miRNA biogenesis, a Bayesian Network Analysis (BNA) was performed based on CLIP-seq derived information on binding sites of RBPs across different miRNAs, expression data of miRNAs, RBPs, and possible PPI partners of RBPs. BNA was performed for all 47 experimental conditions separately and its input consisting of expression data of pre-miRNAs, mature miRNAs,

RBPs and their respective PPI data. All miRNAs are not expressed in each experimental condition. Therefore, in each experimental condition only those pre-miRNA and mature miRNA were considered which expressed at least in three independent samples. The RBPs having binding sites across those expressed miRNAs(including pri/precursors) for the given experimental condition were considered for BNA for the given experimental condition along with associated PPI partners.

Referring to the main idea of miRNA biogenesis where RBPs are the causal factors of miRNA processing, a more general approach was adopted via structural equation modeling (SEM). The basic idea in SEM is to estimate the potential RBPs involved in each step of miRNA biogenesis. The basic model used in the current study is a p-dimensional random vectors $X= (X_1,X_2,. . . , X_p)$ with joint distribution $P(X_1,X_2,. . . , X_p)$. BNs are directed graphical model and their edges encode conditional independence constraint implied by the joint distribution of **X,** which is:

$$\square x_j \square P\left(X_{1,} X_{2,\ldots}, X_p\right) = \prod_{j=1}^{p} \square pa\left(x_j, \theta_j\right) \square \qquad \textbf{(1)}$$

Here $pa(X_j = \{X_i : X_i \square E\}$ is the parent set of X_j and θ_j encode parameters that defines the conditional probability distribution (CPD) for X_j . In this approach, initially CPDs were estimated between RBP and miRNA, RBP and its PPI partners based on the expression data and prior information. The number of RBPs having binding sites in each miRNA was used as the prior information in this study. From total DAG, significant DAGs were filtered considering a suitable convergence criteria. The significant DAGs are directly modeled via a generalized linear model. It was assumed through out the study that the data are generated from a multivariate Gaussian distribution, where covariance matrix is positive definite. Such a model can be written as a set of gaussian structural equation. If $X = [X_1 | X_2 | \ldots . | X_P]$ is an n×p matrix of *i.i.d* observations, then the set of structural equation can be rewrite in matrix notation as:

$$X = \mathbf{B}^T X + E \qquad \textbf{(2)}$$

This is called a SEM for total observations **X**. **B** is the weighted adjacency matrix of a directed graph, which can be written as $B=[|\beta_1|,\ldots,|\beta_p|] \square R^{p\times p}$ and E□ N (0, w_j^2). In this approach, we have to make sure the adjacency matrix **B** to be acyclic. The nodes in **B** are in one to one correspondence with the random variables $(X_1,X_2,X_3,\ldots,X_P)$ in the model. Algorithms for building Bayesian network are generally divided into three type: score-based methods, constraint-based methods and hybrid methods. In the present context the network structure is mostly estimated by score-based techniques due to its high-dimensional nature. In the present approach a score-based method was used.

In this study, the number of RBPs having binding site on a particular pri-miRNA or a pre-miRNA were used as prior information. Here, it was noticed that greater number of RBPs have binding sites in each miRNA compared to their number of samples in each experimental condition. Binding sites of RBPs on both pri-miRNA and pre-miRNA were considered for pre-miRNA formation from pri-miRNA. For mature miRNA processing from pre-miRNA only those RBPs were considered having binding sites on pre-miRNA. Therefore, it is not possible to estimate the association of all RBPs with the miRNA, which is a high dimensional problem (i.e. n << p). In this study we have used a sparse regularized penalty which controls over-fitting by penalizing the maximum likelihood with respect to the number of model parameters [61,62,63].

To learn a BN from data, regularized maximum likelihood estimation is applied to generate scores while using experimental data. Suppose is $X \square R^{n\times p}$ a matrix of observations and l denote the negative log-likelihood and ρ_λ be L-2 sparse regularizer , we consider the following model for our data

$$min_{B\varepsilon D}\, l(B;X)+\rho_\lambda(B) \qquad \textbf{(3)}$$

where, $D \square R^{p \times p}$ is the set of weighted adjacency matrix that represents a acyclic directed graphs s from all the three terms: the loss function l,the constraint D and the regularizer ρ_λ . The next stage after learning Bayesian network structure is estimating parameters of conditional distribution. In this study, to fit the model an array of penalties was decided and a particular penalty was selected which fits to the data using the minimax concave penalty (MCP) algorithm. Method of least squares regression was used to regress between node and its parent as the data is continuous to estimate the parameters.

Further to improve the estimation of parameter, let $B = est(\beta)$ as a weighted adjacent matrix like before, and use it to estimate the conditional variance by given formula:

$$Est(W_j)^2 = var(x_j - X(est(\beta_j))) \quad \textbf{(4)}$$

Apply $\Omega = diag(est(W_1)^2, \ldots, est(W_p)^2)$ as a variance matrix and combining $[est(B), est(\Omega)]$ to calculate the variance covariance matrix Σ. From the covariance matrix the accuracy for each parameter had calculated. A suitable convergence criteria (Error tolerance < 10^{-4}), precision value >=85% and a 5% level of significance will be considered for selection of parameters. Those parameters decide how larger the effect size (positive/negative) in between miRNA and RBP. The algorithm followed in BNA was performed using “Sparsebn”, "ccdr Algorithm" and "SparsebnUtils" R-packages and the basic steps followed in current approach are described in Figure 6.

Functional assessment of miRNA:RBP associations

The detailed work flow of different steps followed in the current study to form the miRNA:RBP interaction in miRNA biogenesis model is illustrated in Additional file 4 : Figure S18. From the result of BNA, miRNA:RBP combinations were collected. RBPs appearing responsible

(positive/negative association) in processing of pre-miRNA from pri-miRNA, and mature miRNA from pre-miRNA along with their respective back-chain in each experimental conditions were collected. The association of RBPs with miRNAs (pre-miRNA & mature miRNA) was verified based on the expression correlation coefficient of >=|0.6| as strong association. RBP expected to be involved in processing pre-miRNA from pri-miRNA would exhibit a positive correlation with the corresponding pre-miRNA. Similarly, the RBPs apparently involved in mature miRNA formation is supposed to display positive correlation with mature miRNA and negative correlation with its pre-miRNA.

The miRNA:RBP associations obtained in this study were validated across eight different independent normal tissues such as bladder, testis, brain, breast, lungs, pancreas, placenta and saliva with totally different source of data. The aim of this validation was to find out if the observed miRNA:RBP combinations existed in other tissues than those on which primary observations were made and if they hold universality. The above validation was performed only for RBP associations in mature miRNA processing from pre-miRNA, and not for pre-miRNA from pri-miRNA due unavailability of any well-established pre-miRNA or pri-miRNA expression data for these eight tissues. The mature miRNAs expression data were collected from miRmine database [37] and RNA-seq expression data were collected from GTEx [38], ARCHS4 [39] and Array Express [40]. Correlation analysis was performed between mature miRNA expression and RNA-seq expression data. Those combinations obtained for mature miRNA-RBP from the BNA were evaluated in eight independent normal tissues considering absolute correlation of 0.6 (>=|0.6|) as strong association. A correlogram plot was constructed using the corr-plot R package to visualized the miRNA and RBP association during mature miRNA processing from pre-miRNA combining all eight tissue expression data stated above.

To check the functional association of both miRNAs and their associated RBP, a functional enrichment analysis was performed. The mature miRNAs were clustered based on their expression data (combining the eight tissues) and those RBPs associated with the cluster of miRNAs were separated and also checked the number of RBPs common among the group of miRNAs in a particular cluster. Those mature miRNAs were considered for clustering where at least 50% samples exhibited expression. A functional enrichment (pathway, molecular function and biological process) was performed for both miRNA targets and those associated RBPs for each cluster of miRNAs. Experimentally validated miRNAs target was collected from miRTarBase [**54**] database. The common pathways, molecular functions, and biological processes were checked for each miRNA cluster and their associated RBPs.

RNA-seq based potential miRNome profile detection using XGBoost regression

The objective was to build predictive models of miRNA expression based on the gene expression data. For prediction of miRNA expression level the interaction network of RBP and its associated proteins obtained from miRNA biogenesis model based on the Bayesian network analysis were used. To built the predictive model XGBoost regression [5**1**] was used. XGBoost stands for extreme gradient boosting. For a given dataset with n samples and m features $D=\{(X_i,y_i)\}(|D|=n,X_i \in R^m,y_i \in R)$, a ensemble tree model uses K additive function to predict the output

$$estimate(y_i)=\Phi(X_i)=\sum_{k=1}^{K} f_k(X_i), f_k \in F \quad (5)$$

where $F=f_x=w_{q(x)}(R_m \rightarrow T, W \in R^T)$ is the space of regression trees. Here '***q*** ' defines the structure of each tree that maps its corresponding leaf index. T is the number of leaves in a tree. Each f_k corresponds to an independent tree structure q and leaf weight w. Unlike decision trees, each

regression tree contains a continuous score on each of leaf, we use w_i to represent score on *i*-th leaf.

To learn the set of functions used in the model, we minimize the following regularized objective

$$L(\Phi) = \sum_i \iota\left(estimate(y_i), y_i\right) + \sum_k \Omega(f_k) \tag{6}$$

where, $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \square w \square^2$

Here 'ι' is a differentiable convex loss function that measure the difference between the prediction $estimate(y_i)$ and target y_i . The second term, **Ω,** penalizes the complexity of the regression tree function. The additional regulizer term helps to smoothen the final learned weight to avoid over-fitting. The regularized objective finally selects a model applying simple and predictive functions.

The RNA-seq and sRNA-seq expression data were collected from TCGA database for seven different tissues (both normal and cancer). These are independent from the above described 47 experimental conditions used in the construction of miRNA biogenesis models. For further validation of the tool other eight different tissue condition data were used separately from TCGA. Details about these training, testing, and revalidatory data are given in Additional file 1 : Table S20. The predictive models were validated considering the following statistical measures:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{\left(Predicted(y_i) - observed(y_i)\right)^2}{n}} \tag{7}$$

To check the predictive accuracy of each model Relative mean absolute percentage error (RMAPE) was used. The RMAPE is widely used to validate forecast accuracy, which provides an indication of the average size of prediction error expressed as a percentage of the relevant observed value.

$$RMAPE = \frac{1}{n} \frac{\sum_{i=1}^{n} |(observed(y_i) - predicted(y_i))}{observed(y_i)} \times 100 \quad (8)$$

$$Model\ Accuracy = 100 - RMAPE \quad (9)$$

Application of the developed tool in discovery of SARS-CoV-2 host miRNome system

RNA-seq data of four covid19 patient data (Accession number:CRA002390) and six control samples were collected from BIG Data Center (https://bigd.big.ac.cn/). In this study transcriptome sequencing of the RNAs isolated from the bronchoalveolar lavage fluid (BALF) and peripheral blood mononuclear cells (PBMC) specimens of Covid19 patients was done [52]. Another RNA-seq dataset (GSE147507) was collected from GEO database for healthy individuals and Covid19 patients (Lungs biopsy) having two samples each [53]. Gene expression analysis was performed using the protocol described in the method section for both the datasets. By using the developed approach described in the previous sections, miRNAs and their expression level were predicted using the gene expression profile for each sample using their RNA-seq data. Those miRNAs expressed in at least 50%. Those miRNAs which were displaying commonality between both the studies were considered further. The SARS-CoV-2 specific up and down regulated miRNAs were decided based on the $\log_2$ fold-change values compared to control conditions. The experimentally validated targets of different miRNAs for both up and down were collected from mirTarbase database [54] separately. Those targets which displayed expression anti-correlation of 0.7 or more with their respective miRNAs in each dataset were only considered for further analysis. The enrichment analysis was performed for pathways, biological processes, and molecular function using Enrichr[55]. The enrichment analysis was performed for individual miRNAs and also by combining all miRNA targets together while considering significance p-value <= 0.05. The pathways were studied for three different databases such as KEGG, Wiki, and PantherDB. Different pathways were ranked based on the number of miRNA target and number of genes involved in that

particular pathway. The pathway maps were constructed for each pathways reported in this study. Pathways mapping was done using locally developed script, implemented in miRbiom server, to visually analyze the exact location and interactions of the gene in any given pathway.

shRNA Constructs and transformation

To validate and complement our computational findings, it was important to detect mature levels of miRNAs in the absence of DICER and AAR2. Based on literature and previous studies an experiment was designed to evaluate whether miRNA maturation is largely independent of DICER and dependent on AAR2. The computational data ascertained the plausible role of AAR2 in maturation of miRNAs. To validate it experimentally, we also evaluated the levels of the mature miRNAs post knockdown of AAR2 in a follow-up study as well. The study was conducted on two connected fronts: 1) generating stable DICER knockdown in cells using shRNA and detecting levels of mature miRNA after stable knockdown of the DICER and 2) the follow-up experiment of generating AAR2 deficient cells through knockdown and detection of the same set of miRNAs in $AAR2^{-/-}$ cells. Schematic representation of the study depicting the work-plan and consecutive steps are presented in Additional file 4: Figure S19.

Four different shRNA expression clone constructs targeting DICER and three targeting AAR2 along with one scrambled control for each sets in psi-U6 vector were procured from Genecopoeia. Constructs had eGFP as reporter gene and puromycin as mammalian selection marker (Additional file 4 : Figure S20 and Table 3). All constructs were cloned in DH5-alpha competent cells (Invitrogen) using previously defined protocol (38). Plasmids were isolated and transformants were analyzed for identification of clones expressing the desired shRNA constructs (Additional file 4 : Figure S20).

Antibodies

Primary antibodies raised against eGFP (Thermo Fisher), DICER (Thermo Fisher) and Beta-tubulin (Santa Cruz biotechnology) were used in the analysis (Table-4). Anti-Mouse IgG-horseradish peroxidase (HRP) (Bio-Rad) and Anti-Rabbit IgG-HRP (Bio-Rad) raised in goat (1:3000 dilution) were the secondary antibodies used in the study.

Cell Culture

Human Head and Neck Cancer cell line (CAL 27) was obtained from ATCC (American Type Tissue Culture), USA. The cells were cultured under prescribed conditions in DMEM (Dulbecco's Modified Eagle's Medium) culture medium, supplemented with 10% of FBS (Fetal Bovine Serum; Gibco) and 1 % antibiotic-antimycotic solution (Invitrogen) at 37 degree C in 5% CO2 atmosphere. Cell line used in the experimentation was pre-authenticated through ATCC and checked for contamination free culture using MycoFluor™ Mycoplasma Detection Kit (Invitrogen) and Cell Culture Contamination Detection Kit (Invitrogen) before start of the experiment. Additionally, we analyzed the cell morphology and population doubling time before beginning of the experiment.

Transfection and shRNA mediated knockdown of DICER and AAR2

Transfection was performed with the aid of Attractene transfection reagent (Qiagen) using manufacturer's recommended traditional transfection protocol (Qiagen). Briefly, five different transfection complexes containing shRNAs and scrambled control were prepared. 1.0×10^6 CAL27 cells were seeded per 100 mm cell culture treated petri dish, cells were grown at 60% confluency and starved in FBS and antibiotic free DMEM medium for 1 hour. Post starvation, transfection complex consisting of 5 µg shRNA plasmid constructs diluted in DMEM medium and 12.5 µL Attractene transfection reagent were introduced to cells and incubated as per the manufactures protocol. Selection of positive transformants were performed using selection medium containing DMEM and Puromycin (1µg/ml puromycin, Sigma Aldrich). Post selection, transformants were further maintained in selection media for 21 days to obtain a stable transformed cell line. All

experiments were performed in triplicates. Validation of transfection and knockdown efficiency were performed using fluorescence microscopy and immunoblotting. Non-transfected CAL27 cells were used as control in the experiment.

Fluorescence microscopy

Fluorescence microscopic imaging (EVOS-FL Auto 2 Imaging System, Thermo Scientific) were performed for confirmation of knockdown of both the proteins and evaluation of knockdown efficiency of different shRNA plasmid constructs. Briefly, images of transfected and control cells were obtained at different timepoints before and during the process of transfection, post transfection (pre selection) and post selection. Images were acquired at different magnifications under bright-field and fluorescence channel to check the expression of eGFP protein inside transformant cells.

Immunoblotting

Protein expression of GFP and DICER was assessed by immunoblotting as per previously described protocol . Whole cell lysate of CAL27 cells were prepared in RIPA lysis buffer (Sigma Aldrich) followed by resolving through SDS-polyacrylamide gel electrophoresis (7% and 12% acrylamide) and blotted on nitrocellulose membrane. Primary antibodies raised against eGFP (Thermo Fisher), DICER (Thermo Fisher) and Beta-tubulin (Santa Cruz biotechnology) were used in the analysis (Table 4). Anti-Mouse IgG-horseradish peroxidase (HRP) (Bio-Rad) was the secondary antibody used in the study. Clarity™ Western enhanced chemiluminescence (ECL) Substrate (Bio-Rad) used to visualize the protein bands with ECL imager (Azure).

Total RNA Isolation, cDNA synthesis and quantitative Real time PCR (qPCR) for detection of mature miRNA levels

Cell to Cst -Quantimir kit (SBI- Systems Biosciences) used for the isolation and quantification of target miRNAs as per the manufacture's recommended protocol. Briefly, the whole process

involved isolation of the total RNA, tagging of targeted non-coding micro RNAs and cDNA synthesis and finally quantification of measurable cDNA using qPCR. Total RNA was isolated from CAL27 cultured cells of transfected (CAL27 sh3 for DICER and sh2 for AAR2) as well as non-transfected control / normal control group (CAL27 NT) using Cell to Cst-Quantimir kit (System Biosciences, SBI) as per the manufacturer's recommended protocol. Then the non-coding RNAs were tagged with poly-A tail and anchored with oligo-dT adapters. Subsequently, cDNA was synthesized to form pool of tagged non-coding miRNAs using the forward primers for our target miRNAs and universal reverse primer supplied with the kit, as per the manufacturers recommended protocol. cDNA was checked using end point PCR (Additional file 4: Figure S20) and was diluted 1:5 before proceeding to the final quantification step. qPCR was performed using SYBR Green Jump start Taq Ready Mix (Sigma Aldrich, USA) on ABI, USA instrument by means of default parameters. HPLC grade primers synthesized from Integrated DNA Technologies, Inc (IDT) were used for this step. U6 was used as an endogenous control in experiment to facilitate the relative quantification of generated qPCR data. Experiments were conducted in triplicates and statistical analysis was performed using GraphPad Prizm software version 7.0. (Protocol employed of the end point PCR, qPCR along with the list of primers used is provided in Table 5).

Acknowledgments

We are thankful to the Director, CSIR-IHBT, for his kind support in getting the experimental validation work carried out at CSIR-IHBT. We are thankful to Dr. Dharam Singh for his discussions. We are thankful to DBT for the funding support they gave for this project. UKP and PK are thankful to ICAR, New Delhi for providing support in Ph.D. PA is immensely thankful to Department of Science and Technology (DST), Government of India for the fellowship. UKP, PA, NKS, and PK are thankful to Academy of Scientific and Innovative Research (AcSIR) for their

Ph.D. enrollment. We are also thankful to Anoop Kumar for help in image generation for this MS. This MS has CSIR-IHBT MSID # 4643.

Funding

RS is thankful to Department of Biotechnology, Govt. of India for supporting this study through grant in Big Data analysis[Grant number: BT/PR16331/BID 17/589/2016 (GAP-0228)] to RS.

Conflict of interest statement. None declared.

Availability of data and materials

All secondary data used in this study are available in supplementary data files provided along with and also made available through the related server at https://scbb.ihbt.res.in/miRbiom.

Contributions

UKP carried out the computational and statistical part of the study. PA carried out the wet-lab experiments. PK developed the web-server. NS carried out the Covid19 research application part and participated in computational analysis and helped in server implementation. AK carried out the structural analysis of AAR2. RP was involved in designing the wet-lab experiments. YP designed and supervised the wet-lab experiment part and drafted the MS for the wet-experiments part along with PA. RS conceptualized, designed, analyzed and supervised the entire study. UKP and RS wrote the MS.

References

1. Lewis BP, Shih I -hung, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian

microRNA targets. Cell. 2003;115:787–98.

2. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 2006;34:D140-144.

3. Jha A, Panzade G, Pandey R, Shankar R. A legion of potential regulatory sRNAs exists beyond the typical microRNAs microcosm. Nucleic Acids Res. 2015;43:8713–24.

4. Kawaji H, Severin J, Lizio M, Waterhouse A, Katayama S, Irvine KM, et al. The FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. Genome Biol. 2009;10:R40.

5. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007;447:799–816.

6. Ghildiyal M, Zamore PD. Small silencing RNAs: an expanding universe. Nat Rev Genet. 2009;10:94–108.

7. Kiss T. Biogenesis of small nuclear RNPs. J Cell Sci. 2004;117:5949–51.

8. Hentze MW, Castello A, Schwarzl T, Preiss T. A brave new world of RNA-binding proteins. Nat Rev Mol Cell Biol. 2018;19:327–41.

9. Proudfoot NJ. Ending the message: poly(A) signals then and now. Genes Dev. 2011;25:1770–82.

10. Baralle FE, Giudice J. Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol. 2017;18:437–51.

11. Shav-Tal Y. RNA localization. Journal of Cell Science. 2005 ;118:4077–81.

12. Barreau C, Paillard L, Osborne HB. AU-rich elements and associated factors: are there unifying principles? Nucleic Acids Res. 2005;33:7138–50.

13. Gregory RI, Yan K-P, Amuthan G, Chendrimada T, Doratotaj B, Cooch N, et al. The Microprocessor complex mediates the genesis of microRNAs. Nature. 2004;432:235–40.

14. Dueck A, Meister G. Assembly and function of small RNA - argonaute protein complexes. Biol Chem. 2014;395:611–29.

15. Zisoulis DG, Kai ZS, Chang RK, Pasquinelli AE. Autoregulation of microRNA biogenesis by let-7 and Argonaute. Nature. 2012;486:541–4.

16. Westholm JO, Lai EC. Mirtrons: microRNA biogenesis via splicing. Biochimie. 2011;93:1897–904.

17. Newman MA, Thomson JM, Hammond SM. Lin-28 interaction with the Let-7 precursor loop mediates regulated microRNA processing. RNA. 2008;14:1539–49.

18. Morlando M, Dini Modigliani S, Torrelli G, Rosa A, Di Carlo V, Caffarelli E, et al. FUS stimulates microRNA biogenesis by facilitating co-transcriptional Drosha recruitment. EMBO J. 2012;31:4502–10.

19. Abdelmohsen K, Tominaga-Yamanaka K, Srikantan S, Yoon J-H, Kang M-J, Gorospe M. RNA-binding protein AUF1 represses Dicer expression. Nucleic Acids Res. 2012;40:11531–44.

20. Cifuentes D, Xue H, Taylor DW, Patnode H, Mishima Y, Cheloufi S, et al. A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. Science. 2010;328:1694–8.

21. Jha A, Mehra M, Shankar R. The regulatory epicenter of miRNAs. J Biosci. 2011;36:621–38.

22. Nussbacher JK, Yeo GW. Systematic Discovery of RNA Binding Proteins that Regulate MicroRNA Levels. Mol Cell. 2018;69:1005-1016.e7.

23. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116:281–97.

24. Bartel DP, Chen C-Z. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. Nat Rev Genet. 2004;5:396–400.

25 Treiber T, Treiber N, Plessmann U, Harlander S, Daiß J-L, Eichner N, et al. A Compendium of RNA-Binding Proteins that Regulate MicroRNA Biogenesis. Mol Cell. 2017;66:270-284.e13.

26. Keene JD. RNA regulons: coordination of post-transcriptional events. Nat Rev Genet. 2007;8:533–43.

27. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING

v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43:D447-452.

28. Niaz S, Hussain MU. Role of GW182 protein in the cell. Int J Biochem Cell Biol. 2018;101:29–38.

29. Zhang, L. Kong, S. Guo, M. Bu, Q. Guo, Y. Xiong, N. Zhu, C. Qiu, X. Yan, Q. Chen, H. Zhang J, Kong L, Guo S, Bu M, Guo Q, Xiong Y, et al. hnRNPs and ELAVL1 cooperate with uORFs to inhibit protein translation. Nucleic Acids Res. 2017 ;45:2849–64.

30. Auxilien S, Guérineau V, Szweykowska-Kulińska Z, Golinelli-Pimpaneau B. The human tRNA m5C methyltransferase Misu is multisite-specific. RNA Biol. 2012;9:1331–8.

31. Sakita-Suto S, Kanda A, Suzuki F, Sato S, Takata T, Tatsuka M. Aurora-B Regulates RNA Methyltransferase NSUN2. Mol Biol Cell. 2007;18:1107–17.

32. Zhong X, Yu J, Frazier K, Weng X, Li Y, Cham CM, et al. Circadian Clock Regulation of Hepatic Lipid Metabolism by Modulation of m6A mRNA Methylation. Cell Rep. 2018;25:1816-1828.e4.

33. Xiang Y, Laurent B, Hsu C-H, Nachtergaele S, Lu Z, Sheng W, et al. RNA m6A methylation regulates the ultraviolet-induced DNA damage response. Nature. 2017;543:573–6.

34. Du Y, Hou G, Zhang H, Dou J, He J, Guo Y, et al. SUMOylation of the m6A-RNA methyltransferase METTL3 modulates its function. Nucleic Acids Res. 2018;46:5195–208.

35. Alarcón CR, Lee H, Goodarzi H, Halberg N, Tavazoie SF. N6-methyladenosine marks primary microRNAs for processing. Nature. 2015;519:482–5.

36. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science. 2005;308:523–9.

37. Panwar B, Omenn GS, Guan Y. miRmine: a database of human miRNA expression profiles. Bioinformatics. 2017;33:1554–60.

38. Amore personal view of human-gene regulation . Nature News.2017; 550: 157.

39. Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, et al. Massive mining of publicly available RNA-seq data from human and mouse. Nat Commun. 2018;9:1366.

40. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, et al. ArrayExpress--a public database of microarray experiments and gene expression profiles. Nucleic Acids Res. 2007;35:D747-750.

41. Kooshapur H, Choudhury NR, Simon B, Mühlbauer M, Jussupow A, Fernandez N, et al. Structural basis for terminal loop recognition and stimulation of pri-miRNA-18a processing by hnRNPA1. Nature Communications. 2018;9.

42. Alarcón CR, Goodarzi H, Lee H, Liu X, Tavazoie S, Tavazoie SF. HNRNPA2B1 is a mediator of m6A-dependent nuclear RNA processing events. Cell. 2015;162:1299–308.

43.Wu H, Sun S, Tu K, Gao Y, Xie B, Krainer AR, et al. A splicing-independent function of SF2/ASF in microRNA processing. Mol Cell. 2010;38:67–77.

44. Zhao L, Mao Y, Zhao Y, He Y. DDX3X promotes the biogenesis of a subset of miRNAs and the potential roles they played in cancer development. Sci Rep. 2016;6:32739.

45. Hubé F, Ulveling D, Sureau A, Forveille S, Francastel C. Short intron-derived ncRNAs. Nucleic Acids Res. 2017;45:4768–81.

46. Li H-K, Mai R-T, Huang H-D, Chou C-H, Chang Y-A, Chang Y-W, et al. DDX3 Represses Stemness by Epigenetically Modulating Tumor-suppressive miRNAs in Hepatocellular Carcinoma. Scientific Reports. 2016;6:28637.

47. Scott MS, Avolio F, Ono M, Lamond AI, Barton GJ. Human miRNA Precursors with Box H/ACA snoRNA Features. PLOS Computational Biology. 2009;5:e1000507.

48. Dueck A, Ziegler C, Eichner A, Berezikov E, Meister G. microRNAs associated with the different human Argonaute proteins. Nucleic Acids Res. 2012;40:9850–62.

49. Plante I, Davidovic L, Ouellet DL, Gobeil L-A, Tremblay S, Khandjian EW, et al. Dicer-derived microRNAs are utilized by the fragile X mental retardation protein for assembly on target RNAs. J Biomed Biotechnol. 2006;2006:64347.

50. Zhang T, Wu Y-C, Mullane P, Ji YJ, Liu H, He L, et al. FUS Regulates Activity of MicroRNA-Mediated Gene Silencing. Mol Cell. 2018;69:787-801.e8.

51. T.Chen and C.Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(2016), pages

785–794.

52. Xiong Y, Liu Y, Cao L, Wang D, Guo M, Jiang A, et al. Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients. Emerging Microbes & Infections. 2020;9:761–70.

53. Blanco-Melo D, Nilsson-Payant BE, Liu W-C, Møller R, Panis M, Sachs D, et al. SARS-CoV-2 launches a unique transcriptional signature from in vitro, ex vivo, and in vivo systems. Microbiology; 2020.

54.Huang H-Y, Lin Y-C-D, Li J, Huang K-Y, Shrestha S, Hong H-C, et al. miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. Nucleic Acids Res. 2020;48:D148–54.

55. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44:W90–7.

56. Mi H, Muruganujan A, Huang X, Ebert D, Mills C, Guo X, et al. Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). Nature Protocols. 2019 ;14:703–21.

57. Gahlan P, Singh HR, Shankar R, Sharma N, Kumari A, Chawla V, et al. De novo sequencing and characterization of Picrorhiza kurrooa transcriptome at two temperatures showed major

transcriptome adjustments. BMC Genomics. 2012;13:126.

58. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.

59. Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics. 2009;25:1026–32.

60. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. 2012;40:37–52.

61. Aragam B, Zhou Q. Concave Penalized Estimation of Sparse Gaussian Bayesian Networks. :56.

62. Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. Journal of the American Statistical Association. 2001;96:1348–60.

63. Zhang C-H. NEARLY UNBIASED VARIABLE SELECTION UNDER MINIMAX CONCAVE PENALTY. The Annals of Statistics. 2010;38:894–942.

List of Tables

Table 1: Experimentally validated miRNA: RBP associations collected from various literature. Majority of these already reported combinations were discovered by the BNA presented in the present study.

RBP	**miRNAs (Pri-miRNA to Pre-miRNA)**	**Association**	**References**
HNRNPA1	hsa-mir-18a	positive	(41)
FUS	hsa-let-7a, hsa-let-7f, hsa-mir-149, hsa-mir-186, hsa-mir-5001, hsa-mir-636, hsa-mir-652, hsa-let-7b, hsa-mir-103, hsa-mir-106a, hsa-mir-124, hsa-mir-132, hsa-mir-135b, hsa-mir-143, hsa-mir-146a, hsa-mir-149, hsa-mir-16, hsa-mir-182, hsa-mir-18a, hsa-mir-191, hsa-mir-192, hsa-mir-194, hsa-mir-197, hsa-mir-199a, hsa-mir-19a, hsa-mir-19b, hsa-mir-218, hsa-mir-21, hsa-mir-22, hsa-mir-23b	Positive	(18)
HNRNPA2B1	hsa-mir-103a, hsa-mir-3651, hsa-mir-6516	Positive	(42)
METTL3	hsa-mir-4284	Positive	(42)
SRSF1	hsa-mir-7,hsa-mir-221	Positive	(43)
DDX3X	hsa-mir-20a, hsa-mir-1, hsa-mir-141, hsa-mir-145, hsa-mir-19b, hsa-mir-34a	Positive	(44)
RBP	**miRNAs (Pre-miRNA to mature miRNA)**	**Association**	**References**
ILF3	hsa-miR-144	Positive	(22)
RBFOX2	hsa-miR-144, hsa-miR-18a, hsa-miR-126	Positive	(22)
SF3B4	hsa-miR-4745, hsa-miR-6753	Positive	(22)
AUF1/HNRNPD	hsa-miR-122	Negative	(19)
CPSF1	hsa-miR-17, hsa-miR-19a, hsa-miR-20a, hsa-miR-19b	Positive	(45)
DDX3X	hsa-miR-122	Positive	(46)
DKC1	hsa-miR-664	Positive	(47)
AGO1	hsa-miR-30a, hsa-miR-182, hsa-miR-21, hsa-miR-183, hsa-let-7f, hsa-let-7a, hsa-miR-30a, hsa-miR-	Positive	(48)

	378a, hsa-miR-140, hsa-miR-92a		
AGO2	hsa-miR-451, hsa-miR-30a, hsa-miR-21, hsa-miR-92a, hsa-miR-99b, hsa-miR-183, hsa-miR-27a, hsa-miR-182, hsa-miR-25	Positive	(48)
AGO3	hsa-miR-30a, hsa-miR-21, hsa-miR-182, hsa-miR-30d, hsa-miR-378a, hsa-miR-30a, hsa-miR-183, hsa-miR-92a, hsa-miR-99b, hsa-miR-151a	Positive	(48)

Table 2: This table descries the different RNA-seq and sRNA-seq experimental data collected in the current study. It provides the accession ID, study ID and different experimental conditions considered in this study.

RNA-seq		**sRNA-seq**		**Experiment**	**Conditions**
Accession no.	**Study id**	**Accession no.**	**Study id**		
GSE56862	SRP041228	GSE56862	SRP041228	PolyA independent deep sequencing of the chromatin-isolated RNA fraction.	Cervical normal
GSE63511	SRP050087	GSE63511	SRP050087	Thyroid tissue	Thyroid tumor , Thyroid norml
GSE68631	SRP058087	GSE68631	SRP058087	HEK293 cells in 6-well plate were transiently transfected with 400 ng plasmids of control, shRNA, shRNA-LC, siRNA-RZ for 48 hr.	HEK293 cell
GSE69446 & GSE66209	SRP058953 & SRP055438	GSE69446 & GSE66209	SRP058953 & SRP055439	Healthy controls and Crohn's patients	Monocytes_chrone, Monocytes_normal
GSE69787	SRP059380	GSE69787	SRP059380	Ossification of the posterior longitudinal ligament (OPLL) tissue and normal posterior longitudinal ligament (PLL) tissue	OPLL and PLL.
GSE86491	SRP090091	GSE86491	SRP090091	Endmometrial tissue from two time points of the menstrual cycle	Endmometrial_TP1, Endmometrial_TP2
GSE37764	SRP012656	GSE37764	SRP012656	Primary non-small cell lung adenocarcinoma tumors and normal tissues.	Lungs Normal , Lungs Tumor
GSE67491	SRP056784	GSE67491	SRP056785	Whole blood samples personalized medicine study	Caucasian male blood,,Caucasian female blood, African_america_female_blood

GSE70666	SRP060566	GSE70666	SRP060565	Oral squamous cell carcinoma	Oral_cancer
GSE85145	SRP080859	GSE85145	SRP080860	Primary Human Astrocytes Infected with Borrelia burgdorferi	Astrocytes_infection
ERS327013	ERP003613	SRP023531	GSE47720	Human pancreatic islets and enriched beta-cells	Pancreatic beta-cells
GSE71336	SRP061565	GSE71336	SRP061566	LNCaP cells expressing the wild-type androgen receptor (AR-WT) or the ligand-independent AR-V7 splice variant	LncaP cell
GSE65515	SRP053046	GSE65515	SRP053046	To reveal dynamic changes in networks of gene expression and epigenetic regulation during healthy human T cell aging.	Newborn, Middle-aged,long-lived
GSE92876	SRP095604	GSE92874	SRP095605	Examination, identification and comparision of mRNA expression profliles in control and schizophrenia npc.	Disease and control
GSE79032 & GSE62830	SRP071331 & SRP 049391	GSE79032 & GSE62830	SRP071334 & SRP049389	RNA Sequencing Reveals that Kaposi Sarcoma-Associated Herpesvirus Infection Mimics Hypoxia Gene Expression Signature & Differential Expression Profiles of miRNA-mRNA Target Pairs in KSHV-Infected Cells	Infected, KSHV-infected
GSE73502	SRP064515	GSE73502	SRP064235	Widespread shortening of 3' untranslated regions and increased exon inclusion characterize the human macrophage response to infection	Salmonella, Non-infected, Listeria (each condition in 2hrs and 24 hrs)
GSE78169	SRP070657	GSE78169	SRP070659	An integrative transcriptomics approach identifies miR-503 as a candidate master regulator of the estrogen response in MCF-7 breast cancer cells	Breast cancer at 10 different time point.
GSE46224	SRP021193	GSE46224	SRP021193	Dynamic regulation of myocardial noncoding RNAs in failing human heart and remodeling with mechanical circulatory support	Heart_ischemic, Heart_nonfailing, Heart_nonischemic
ERS327018	ERP003613	SRX262197	SRP018255	Placenta normal tissue	Placenta
ERS326957	ERP003613	SRX271415	SRP021475	Testis normal tissue	Testis

Table 3: List of shRNA constructs

S.No.	Clone Name	Target	Location	Length	Target Sequence	Designation
1	CS-HSH061144-CU6-01-a(OS527105)	DICER1	398	21	GCAGCTCTGGATCATAATACC	sh1
2	CS-HSH061144-CU6-01-b(OS527106)	DICER1	2162	21	CCAAGTGATCCGTTTACTCAT	sh2
3	CS-HSH061144-CU6-01-c(OS527107)	DICER1	3833	21	GGAAATCAGCTAAATTACTAC	sh3
4	CS-HSH061144-CU6-01-d(OS527108)	DICER1	4293	21	GCAACTGTAATCTGTATCGCC	sh4
5	CSHCTR001-1-CU6(OSNEG20)	None	NA	19	GCTTCGCGCCGTAGTCTTA	SC
6	HSH064128-CU6-b(OS718159)	AAR2	798	21	GCCAGCTGAGATAACCAAGCA	sh1
7	HSH064128-CU6-c(OS718160)	AAR2	930	21	GAATGTGTACGAGGCATTTGA	sh2
8	HSH064128-CU6-d(OS718161)	AAR2	1192	21	GCTCACCTGACCAAGAAGTTC	sh3
9	CSHCTR001-1-CU6(OSNEG20)	None	NA	19	GCTTCGCGCCGTAGTCTTA	SC

Table 4 : List and information of antibodies used

S. No.	Antibody	Dilution	Raised in	Make
1	**GFP (GF28R)**	1:1000	MOUSE	THERMO
2	**DICER (MA5-27595)**	1:1000	MOUSE	THERMO
3	**β tubulin (sc-58882)**	1:500	MOUSE	SANTA CRUZ

Table 5: List and information of Primers

S.No.	miRNA	Sequence	Length (bp)
1	**hsa-miR-25-3p_81(F)**	CATTGCACTTGTCTCGGTCTGA	22
3	**hsa-miR-206_462(F)**	TGGAATGTAAGGAAGTGTGTGG	22

Figure legends

Figure 1: Correlogram based clustering identifies shared function by the cluster members. Commonly expressing miRNAs have common RBPs, common control system, common biological roles. The miRNAs clustered in common group were found to share more than 80% common RBPs which bound to their pre-miRNAs. This diagram represent different clusters of miRNAs along with the members in different cluster. The rectangle boxes display the common pathways, biological process and molecular functions between miRNAs and their associated RBPs belonging to same cluster.

Figure 2: Validation of shRNA mediated knockdown of DICER and AAR2 and levels of mature miRNAs in respective cells. **a)** Through visual imaging of GFP expression through EVOS FL Auto 2 Imaging system (Thermo Scientific Fisher). Images were captured at 10X objective (220X magnification), 20X objective (440X magnification) in trans, GFP and merged format. Representative scale bar denotes 500µm (10X), 200µm (20X). **b)** Western blot assay performed to assess DIECR expression levels. Post knockdown, we observed a clear decrease in the expression levels of DICER in sh1, sh2 and sh3. **c)** Detection of mature miRNA expression levels post DICER knockdown (sh3 shRNA construct) was performed in CAL27 cell line (Normal non-transfected control - $Dicer^{+/+}$ and DICER knockdown- $Dicer^{-/-}$) through qPCR analysis. Histogram depicts upregulation or basal level expression of the miRNA as compared to non-transfected control. The relative fold expression levels of mature miRNAs were obtained through normalization with endogenous U6 control and determining the threshold cycle (Ct) difference between non-transfected control (CAL27 NT) and transfected group (CAL27 sh3) through the $2^{-\Delta\Delta Ct}$ method. All the qPCR assays were conducted in triplicate. Results expressed in terms of means ±standard deviation. d) Through visual imaging of GFP expression through EVOS FL Auto 2 Imaging system (Thermo Scientific Fisher). Images were captured at 10X objective

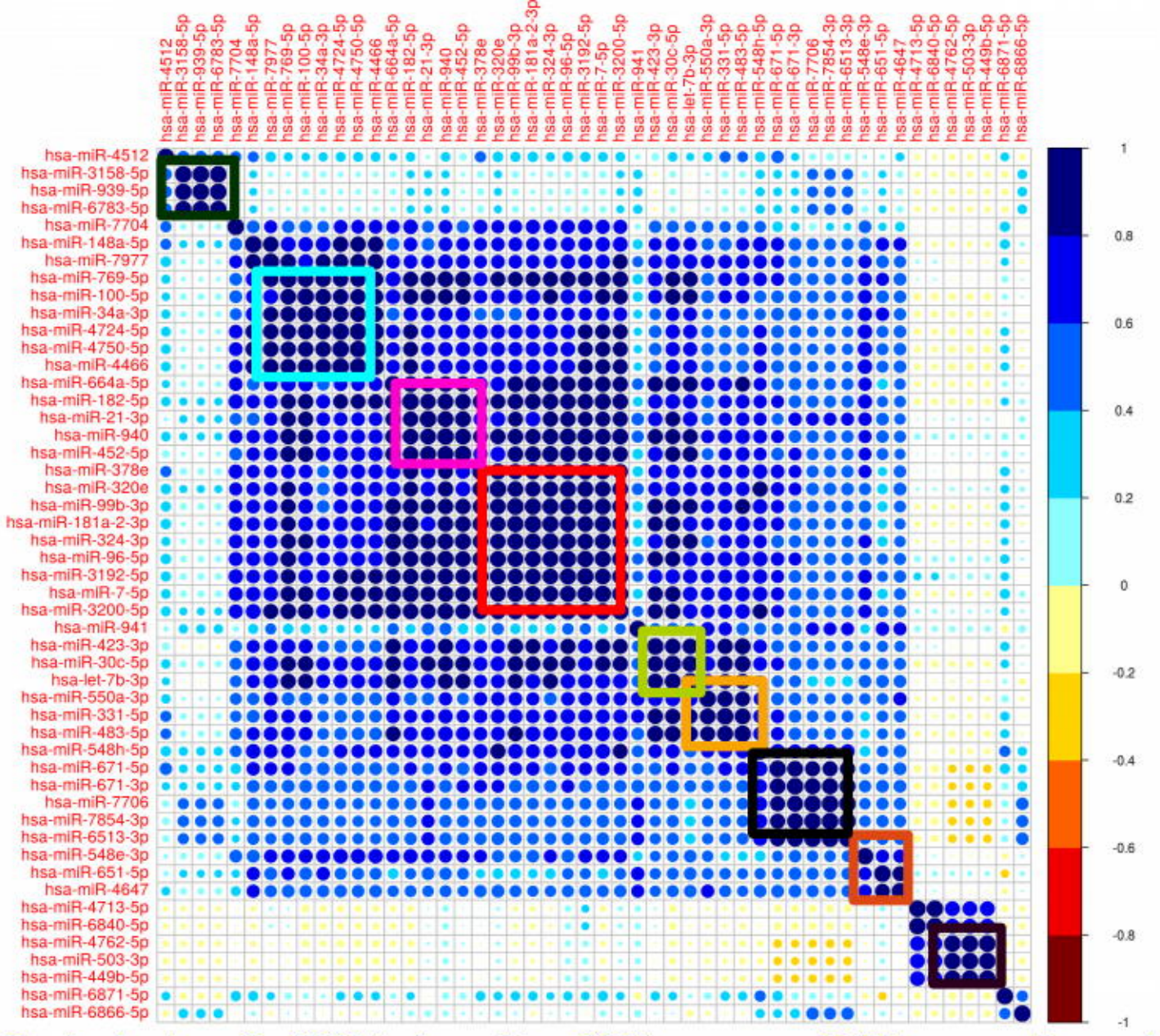
(220X magnification) and 40X objective in trans, GFP and merged format. Representative scale bar denotes 500μm (10X), 100μm (40X). e) Detection of mature miRNA expression levels post AAR2 knockdown (sh3 shRNA construct) was performed in CAL27 cell line (Normal non-transfected control – $AAR2^{+/+}$ and AAR2 knockdown- $AAR2^{-/-}$) through qPCR analysis. Histogram depicts upregulation or basal level expression of the miRNA as compared to non-transfected control. The relative fold expression levels of mature miRNAs were obtained through normalization with endogenous U6 control and determining the threshold cycle (Ct) difference between non-transfected control (CAL27 NT) and transfected group (CAL27 sh2) through the $2^{-\Delta\Delta Ct}$ method. All the qPCR assays were conducted in triplicate. Results are expressed in terms of means ±standard deviation. ****, $p < 0.001$ represent the statistical significance of the expression level as compared to control.

Figure 3: Combination of RBPs decide the fate of miRNAs. There exists both cooperative and competitive association between RBPs in different steps of miRNA biogenesis. Above figure illustrates the cooperative and competitive associations between RBPs during different steps of miRNA biogenesis.

Figure 4: Covid19 specific miRNAome and their important pathways. (A) Figure showing overlap between up and down miRNA set from two study (B) Top 20 miRNAs were selected on the basis of average expression among all fourteen samples, bar plot showing log2 value of average expression, similarly their number of targets are also depicted in same plot (C) miRNAs were ranked on the basis of number of enriched pathways among three different databases (Kegg, Wiki and PantherDB), top 20 ranked miRNAs were selected and number of enriched significant pathways are depicted here. Up miRNAs targeted anti-correlated genes enrichment showing top 20 pathways on the basis of p-value among three different databases Kegg (D) WikiPathways(E) and PantherDB(F). Protein family classification(G) were also done for Up miRNAs targeted anticorelated genes using PantherDB.

Figure-5: miRNAs regulate several genes critically involved in the cross talk of pathways related to IFN-Gamma signaling, Insulin/IGF/ATK/mTOR signaling, and associated Ubiquitin-Proteasomal pathways in SARS-CoV-2 patients**.** IFN-Gamma appears very important molecule here. It regulates Insulin regulated aminopeptidase gene, IRAP. IRAP is very critical for glucose metabolism as it induces translocation of GLUT-4. IRAP in influence of IGN-gamma performs its peptidase function to generate antigens from viral peptides and associates with MHC Class-I to present the peptide to CD-8 T-cells, causing immune response. Besides this, IRAP is also blocked by Angiotensin IV which was found elevated in Covid19 infection. IFN-Gamma system also interacts with IGF/RTK/IR/PI3K systems whose impact is wide, influencing MAPK pathway as well as Insulin/Glucose signaling through PI3K/AKT/mTOR signalling pathways. PI3K/AKT system is also helps in unduction and protein targeting for IRAP-GLUT-4 vesciles for ER/Golgi complex through Ubiquinylation/deubiquitinylation cycles. Tankyrase is a critical protein here which directly interacts with IRAP/GLUT-4 to translocate them. Tankyrase itself if controlled by series of genes invovled in Ubiquitinylation-proteasomal pathways. The same system also regulates production of MHC Class-I and associated immune response. IFN-gamma system is also involved here with JAK/STAT system where it forms complex with its receptor R1 and STAT and represses the transcription of HIF gene, as well as enhances the PHD genes which degrades HIF-alpha and restores normal oxygen conditions. This IFN-Gamma/R1/STAT complex also induces transcription of critical genes involved in immunity, and induces MHC-I and IRAP expression. Very interestingly, most of these critical genes were found downregulated by number of identified miRNAs, starting form IFN-gamma, its receptor R1, RTK, PI3K, AKT, IRAP, MHC-I, Tankyrase, PHD, and 114 critical genes of Ubiquitin-Proteasomal pathways, capturing the observed pathophysiological signs of Covid10 infection. This axis emerges as a very promising one for therapeutic interventions against SARS-CoV-2.

Figure 6: Bayesian network approach in construction of miRNA biogenesis model. Different steps were followed as: 1) Estimation of DAGs between RBP and miRNA, RBP and its PPI partners based on the expression data and prior information. 2) Identification of significant DAGs from total DAGs considering a suitable convergence criteria. 3) The significant DAGs obtained in previous step were directly modeled via a generalized linear model, assuming a multivariate Gaussian distribution. 4) Parameter estimation between RBP and miRNA, RBP and its PPI partners based on the significant DAGs using a sparse regularized penalty. 5) Selection of optimum sparse regularized penalty from an array of penalties that fit best to data. 6) Final estimation of parameters using the optimum penalty and selection of parameters considering certain suitable criteria.



Pathways

mi	C	R
69	48	59
34	23	29
31	21	34
30	19	29
37	24	31
48	27	37
39	18	26
14	9	17
39	26	35

Biological Process

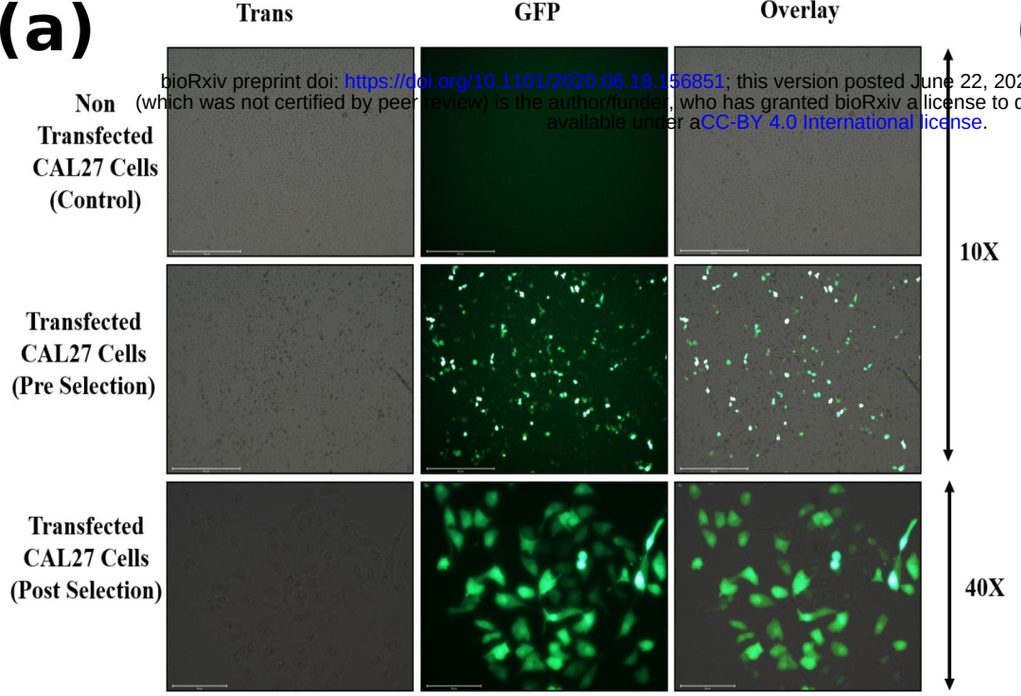
mi	C	R
335	123	166
115	89	193
164	126	334
102	87	274
174	149	178
134	117	265
493	144	186
180	1138	323
152	92	111

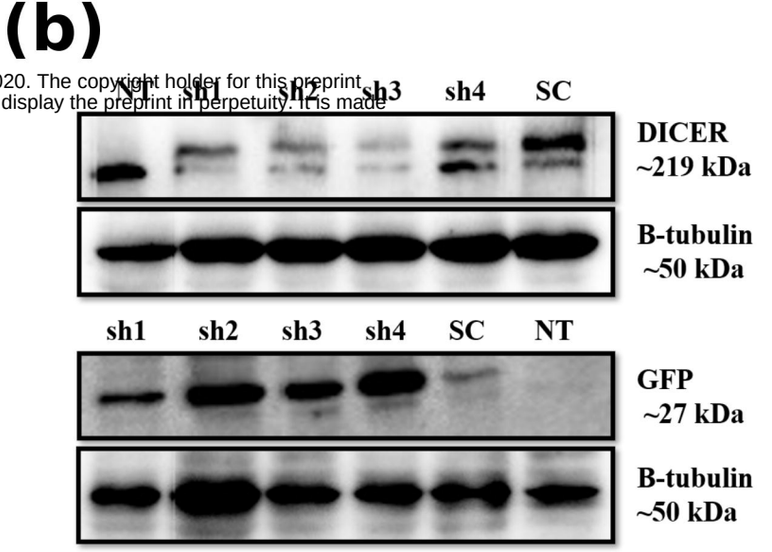
Molecular function

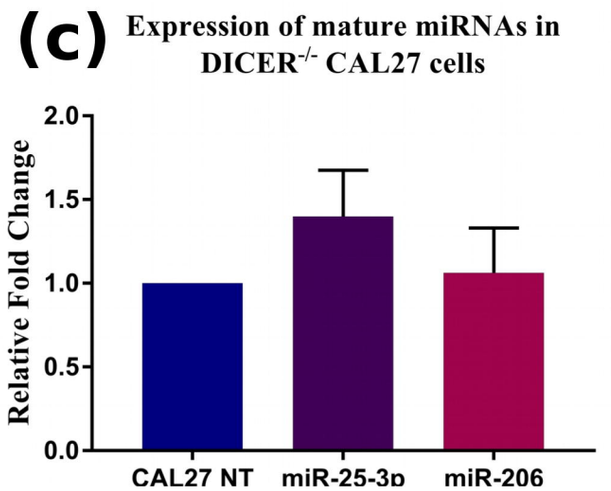
mi	C	R
77	57	84
25	15	98
38	31	142
33	21	107
33	28	77
55	44	120
54	41	97
50	37	140
44	32	50

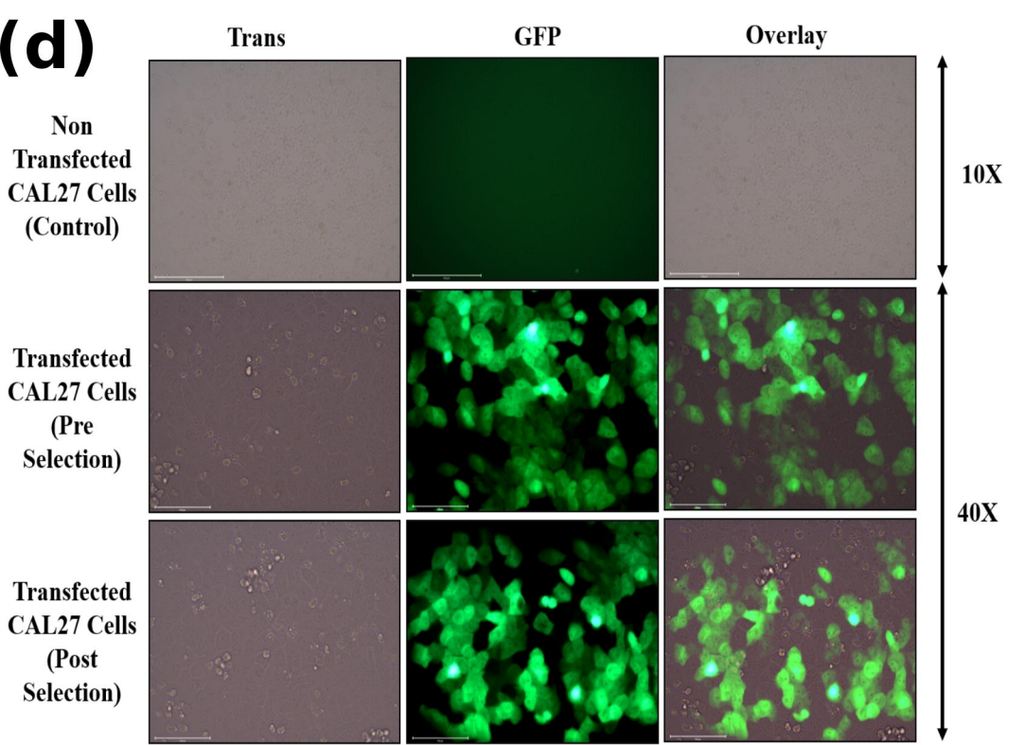
Each cluster of miRNA shared >= 80% common RBP among themselves

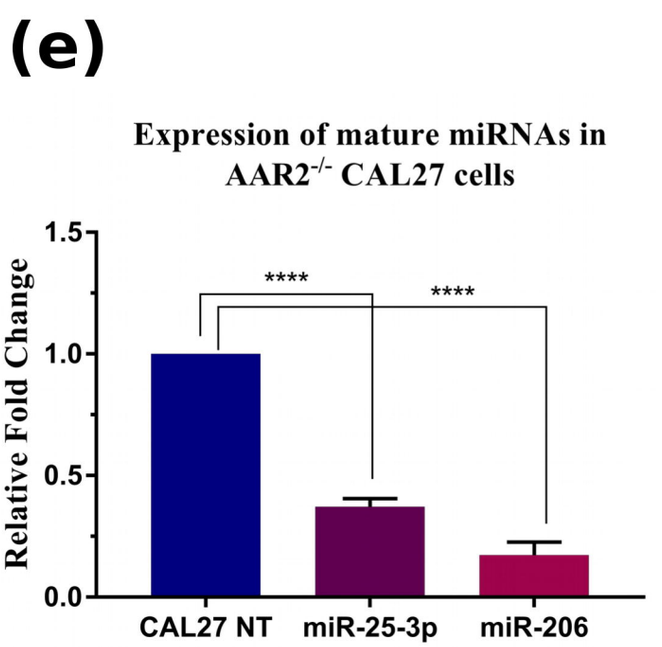
mi-miRNA
C-common
R-RBP



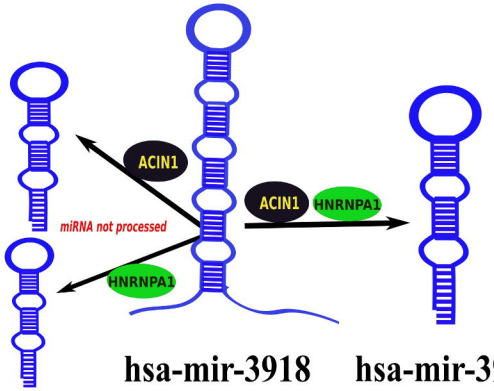








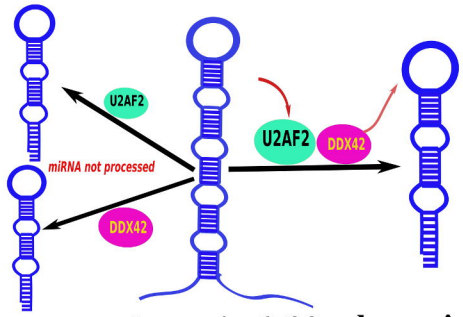
Co-operative model with mutual association



hsa-mir-3918 hsa-mir-3918

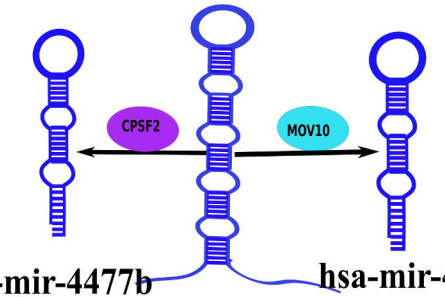
Co-operative model without mutual association

(Pri-miRNA to Pre-miRNA processing)



hsa-mir-6723 hsa-mir-6723

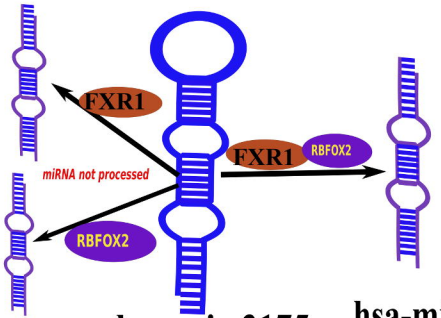
Competative model



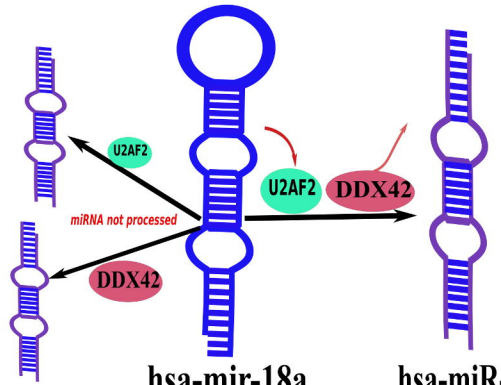
hsa-mir-4477b hsa-mir-4477b

hsa-mir-4477b

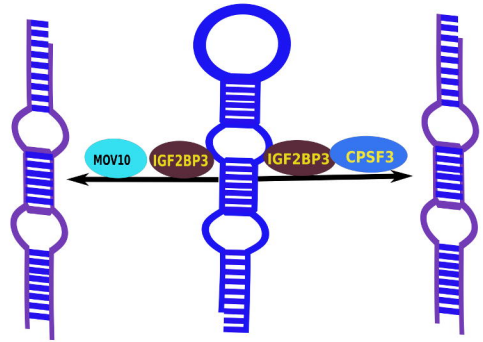
(Pre-miRNA to mature miRNA processing)



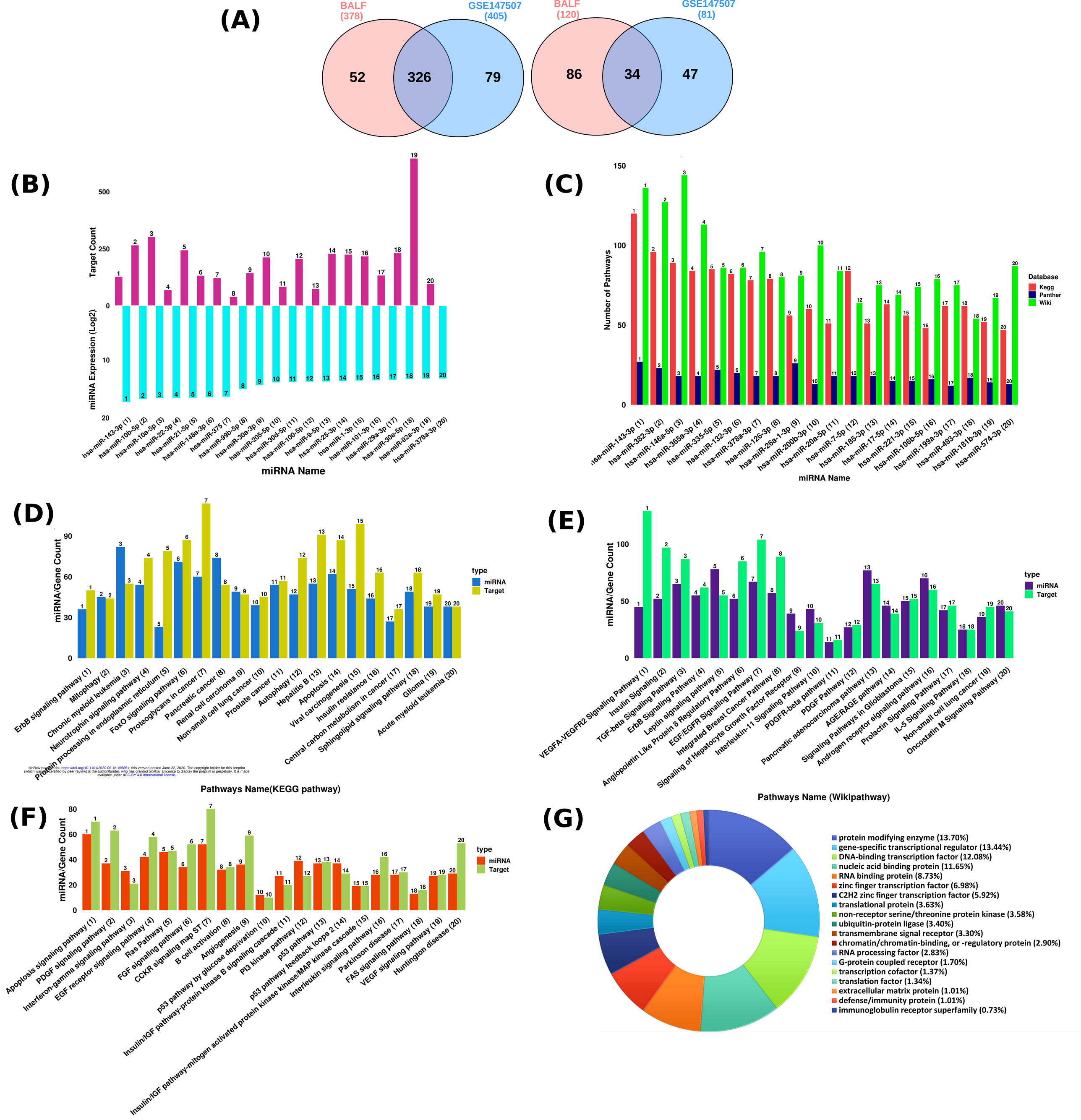
hsa-mir-3175 hsa-miR-3175



hsa-mir-18a hsa-miR-18a-3p



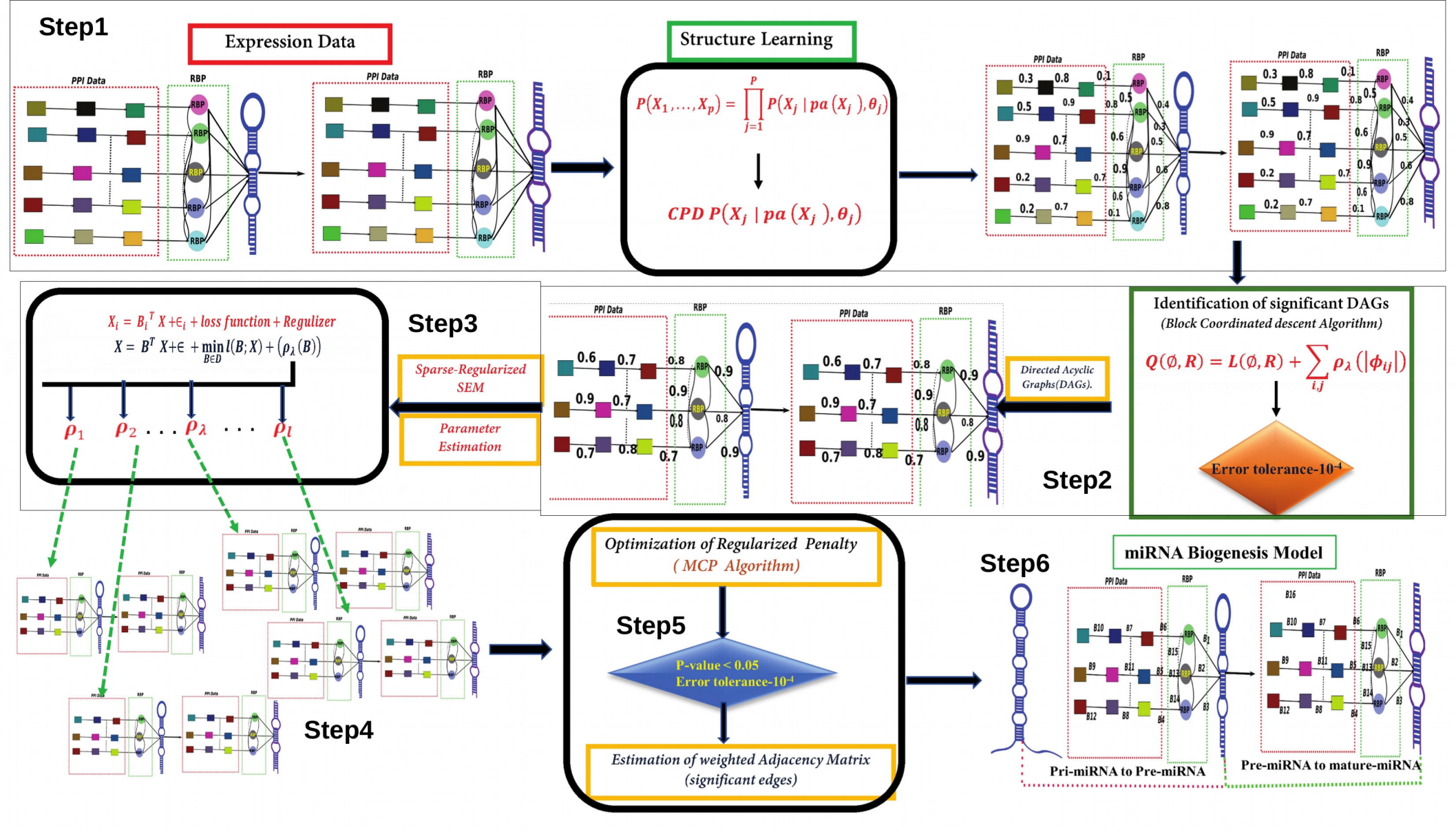
hsa-miR-619-5p hsa-mir-619 hsa-miR-619-5p



IGF
RTK
Insulin Receptor
IRS
IRS
IFNγ
2 miRNAs
IFNγ Receptor
Glucose
Ang-IV
Zn^{2+}
IRAP
IFNγ
GLUT4-mediated glucose uptake
43 miRNAs
23 miRNAs
4 miRNAs
MEK
PI3K
JAK/STAT Pathway
6 miRNAs
ERK
AKT
39 miRNAs
MNK
GSK-3
FOXO1
MAPK Pathway
9 miRNAs
23 miRNAs
Grb2
SOS
Shp2
2 miRNAs
9 miRNAs
21 miRNAs
2 miRNAs
IFNγR1
STAT
IFNγ
HIF
PHD
10 miRNAs
PI3K
AKT
mTOR signaling Pathway
23 miRNAs
39 miRNAs
RNA
8 miRNAs
12 miRNAs
MHC1
TCR
T-cell
Endoplasmic reticulum
Golgi body
Ub-Proteasome Cycle
Tankyrase
3 miRNAs
190 miRNAs
Activation of ubiquitine-proteasome pathway by IFN-G/R1/STAT complex
IFN-G/R1/STAT complex also induces more than 200 genes including MHC1 and IRAP
IFN-G/R1/STAT complex also repress HIF-1 gene
Inhibition
Down-regulated genes

WDTC1	CHFR	BRCA1	
UBE4A	CAND1	UBE2N	UBE2W
DCAF6	ANAPC10	HUWE1	UBE2D2
DCAF11	TOPORS	UBE2E3	APAC1
UBE4B	FANCL	UBE2D1	UBE2S
RNF14	UBE2K	TRIM27	CDC27
NAE1	CDC20	UBE2V2	CDC23
CUL4B	RAD18	UHRF1	FBXW7

Ubiquitin-proteasome pathway related 114 downregulated genes targeted by 190 miRNAs, were found associated with BRCA1 and WDTC1


Step1
Expression Data
PPI Data
RBP
Structure Learning
$P(X_1,\ldots,X_p)=\prod_{j=1}^{P}P(X_j \mid pa(X_j),\theta_j)$
$CPD\ P(X_j \mid pa(X_j),\theta_j)$
Step2
Identification of significant DAGs
(Block Coordinated descent Algorithm)
$Q(\emptyset,R)=L(\emptyset,R)+\sum_{i,j}\rho_\lambda(|\phi_{ij}|)$
Error tolerance-10^{-4}
Directed Acyclic Graphs(DAGs).
Step3
$X_i = B_i{}^T X + \epsilon_i$ + loss function + Regulizer
$X = B^T X + \epsilon + \min_{B \in D} l(B;X) + (\rho_\lambda(B))$
ρ_1 ρ_2 . . . ρ_λ . . . ρ_l
Sparse-Regularized SEM
Parameter Estimation
Step4
Step5
Optimization of Regularized Penalty
(MCP Algorithm)
P-value < 0.05
Error tolerance-10^{-4}
Estimation of weighted Adjacency Matrix
(significant edges)
Step6
miRNA Biogenesis Model
Pri-miRNA to Pre-miRNA
Pre-miRNA to mature-miRNA