Integration of WGCNA, PPI Networks and miRNA Prediction

Abigail Moore

December 2018

Abstract

MicroRNAs (miRNAs) are small noncoding RNAs that contribute to gene regulatory networks. To characterize network dynamics, bioinformatics tools have been developed for miRNA target prediction, protein-protein interactions, and gene co-expression. Presently, no comprehensive tool exists to construct miRNA-gene and gene-gene network dynamics from high-throughput miRNA and RNA-seq data. Thus, we sought to develop an R software package that supports such analyses by capitalizing on existing tools. Our package . . .

Introduction

MicroRNA (miRNA) are small non-coding RNA with an average length of 22 nucleotides. miRNA post-trancriptionally regulate gene expression most frequently by inhibiting expression, although, miRNA may promote expression (O'Brien, 2018). miRNA are active in many biological processes and are known to mediate cell-cell communications (Kloosterman, 2006; Gebert 2018). Aberrantly expressed miRNA are associated with numerous human diseases and are potential prognostic markers (Mendell, 2012; Schwarzenbach, 2014). When overexpressed, each miRNA may target over a hundred genes (Lim, 2005). Some genes are targeted by dozens of miRNA, resulting in miRNA target hubs (Lai, 2012). Further, RNA with miRNA binding sites may act as competitive endogenous RNAs (Ebert, 2012). These miRNA binding sites may also act as hubs that facilitate cross-talk between RNA (Salmena, 2011). Thus, it is interesting to identify aberrantly expressed miRNA, potential miRNA targets, and interacting RNAs. This may be accomplished by integrating weighted gene co-expression analysis (WGCNA), protein-protein interaction (PPI) networks and miRNA target predication.

WGCNA is a systems biology approach to analyzing large, high-dimensional gene expression profiles (Wang, 2017). WGCNA implements pairwise correlations among genes, which results in clusters of genes called modules. The R package called WGCNA (R/WGCNA) allows users to identify modules of significantly correlated genes, correlate modules to sample traits, determine module membership and gene significance, and identify intramodular hub genes. To identify modules, R/WGCNA offers users the option to conduct analyses via a single block or block-wise approach. Since the block-wise approach simply approximates the results produced by a single block, the single block approach is preferred when computational resources are available. The results of WGCNA may be substantiated by analyzing for protein-protein interaction networks to identify physical interactions with previously predicted significance.

Protein-protein interactions predict protein function, participate in many biological processes, and offer information important to drug discovery and development (Rao, 2014). PPI networks may be constructed via R packages like Path2PPI (Philip, 2016), which relies on the iRefIndex (Razick, 2008) database, or cisPath (Wang, 2015), which is supported by the PINA (Cowley, 2012), iRefIndex (Razick, 2008) and STRING (Szklarczyk, 2016) databases. Alternatively, users may directly access and submit queries to these databases via the web. One such database is InBio Map, which provides a scored human PPI network with over 600,000 interactions from eight databases (Wernersson, 2017). InBio Map is distinguished from PINA, iRefIndex, and STRING by its confidence scoring and ability to recapitulate known pathways (Wernersson, 2017). The results of both WGCNA and PPI network analysis may supply a more robust characterization

of molecular functions, especially when exploring the complex interactions found among mRNA and miRNA.

Integrating mRNA and miRNA expression often involves miRNA target prediction for which there are multiple computational tools (Chen, 2018). Tools vary by considering different miRNA-mRNA interaction features like sequence complementarity, free-energy of interactions, conservation of sequences, and binding accessibility (Witkos, 2011). To capitalize on these varied approaches, databases like mirDIP integrate target predictions from multiple platforms (Tokar, 2017). Specifically, mirDIP sources predictions from 30 databases and houses 152 million human miRNA–target predictions. As compared to the two largest integrative resources, mirDIP integrates more prediction tools and includes more predictions (Tokar, 2017). With both predicted miRNA targets and protein-protein interactions, a network may be constructed to describe both miRNA-mRNA and mRNA-mRNA interactions.

The combined knowledge gained from WGCNA, PPI networks, and miRNA target prediction provides a multilevel -omics, systems biology approach to identifying candidate interactions that promote aberrant biological processes. While a variety of tools exist to conduct such an integrated analysis, to our knowledge, no tool exists that streamlines this process by allowing users to conduct analyses out of a single software tool. A few key inefficiencies include 1) Users who wish to employ R/WGCNA to identify modules in single block mode must carry out a multi-step process within R/WGCNA. In contrast, there is a one-step function, *blockwiseModules()*, to perform WGCNA in block-wise mode. 2) Prominent PPI network databases like InBio Map often are not immediately accessible within the R environment. Users must either submit queries via a website or download the database and independently write scripts to query the data. 3) Similarly, while the mirDIP website does offer scripts to perform queries in R, there is no formal R package available at present.

Thus, we sought to streamline the process of integrating WGCNA, PPI networks and miRNA prediction by developing an R package that offers 1) access to miRNA-seq data in public databases like The Cancer Genome Atlas (TCGA; Weinstein, 2013), 2) a one-step function for single-block module detection, 3) PPI network identification with InBio Map, 4) miRNA target prediction with mirDIP, and 5) visualizations, including network export to Cytoscape (Shannon, 2003). The resulting package allows users to capitalize on multiple resources while remaining within the R environment. Within our package, users may identify significant miRNA-mRNA and mRNA-mRNA interactions by submitting normalized mRNA and miRNA expression to obtain gene expression modules, PPI networks, and predicted miRNA targets. These results can be used to construct gene regulatory networks that characterize the presence of miRNA. In sum, this package capitalizes on existing tools to facilitate the integration of high-throughput, multilevel omics data, which could supply new knowledge of miRNA in human disease.

Results

Benchmark findings

To develop benchmark results, we sought to construct a network consisting of miRNA-gene and gene-gene interactions from the normal and tumor tissue of breast cancer patients. We identified 101 patients with mRNA and miRNA raw counts from both normal and tumor tissue within the TCGA breast invasive carcinoma study (BRCA; Methods). We processed these samples through the pre-existing, independent tools R/WGCNA, InBio Map and mirDIP to obtain co-expression modules, protein-protein interactions, and miRNA targets. Further, we visualized results within R and Cytoscape.

After variance stabilizing mRNA counts, we identified 17 co-expression modules (Figs. 1 2). Further, we checked for module preservation within a test dataset consisting of unpaired mRNA counts from the tumor tissue of 59 patients with TCGA BRCA, and mRNA counts from the normal adjacent tumor (NAT) tissue of 59 patients from GTEx (GTEx Consortium, 2013). Preservation of all modules was confirmed despite the test dataset being unpaired, and known gene expression differences between the GTEx NAT tissue and the

TCGA normal tissue (Aran, 2017). Preservation was defined by a Zsummary score (Methods). The midnight blue and magenta modules showed moderate evidence of preservation, while all other modules showed strong evidence of preservation (Fig. 3).

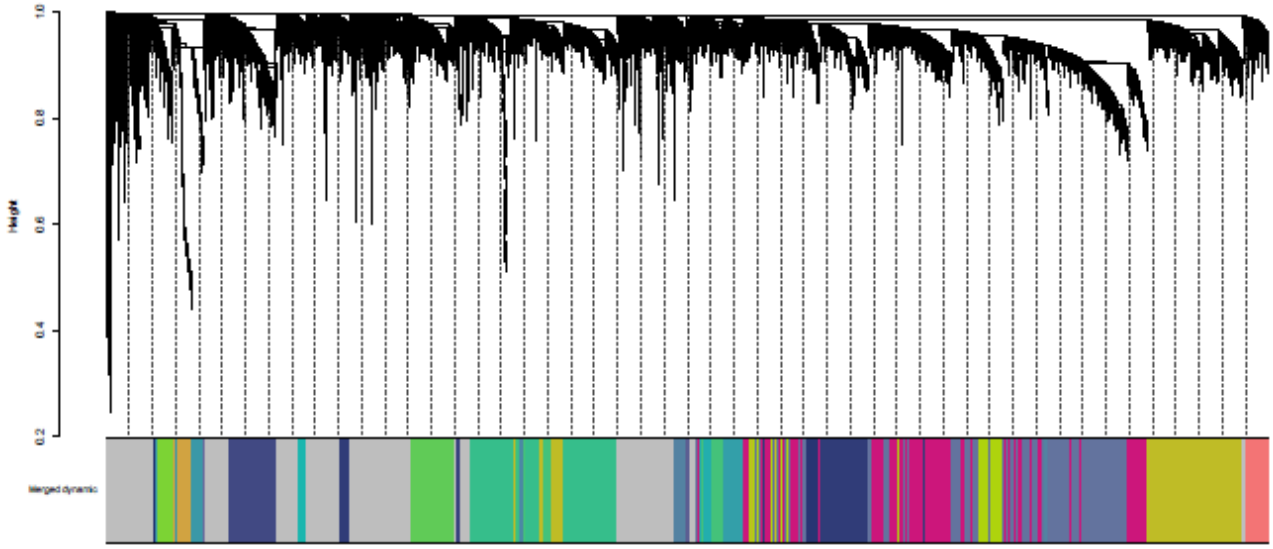

Figure 1. Dendrogram showing gene-gene coexpression modules. Color bars represent each module, exception of grey, which indicates no module assignment. (need to make labels legible and maybe make plot narrower)

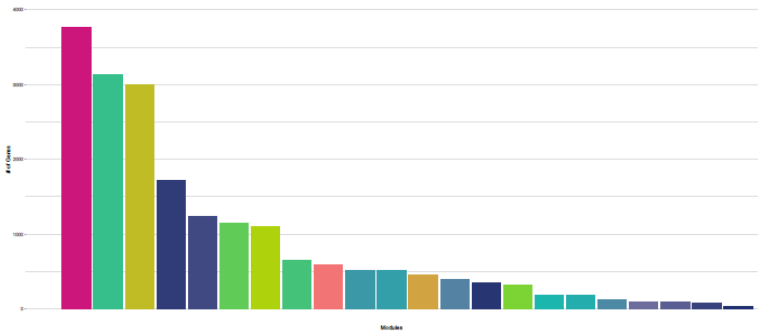

Figure 2. Bar plot representing the number of genes per module. Color bars represent each module.

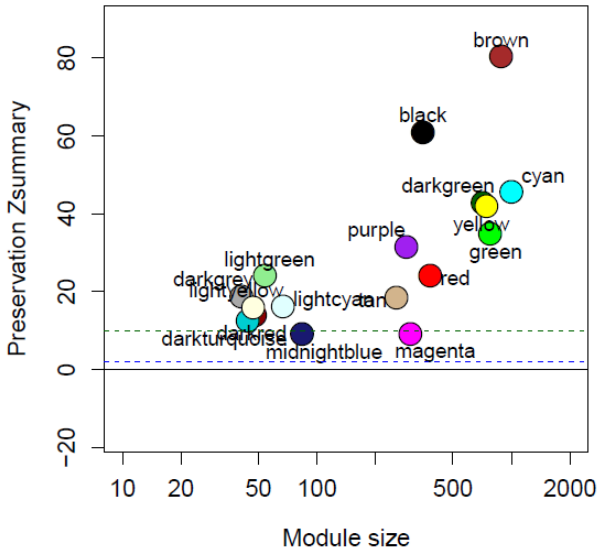

Figure 3. Preservation of all modules defined by Zsummary scores. Colors represent each module. Blue, dashed line represents no evidence of preservation. Black dashed line represents strong evidence of preservation. (need to remove text labels within plot and prevent points from overlapping)

To identify significant miRNA-gene interactions, we calculated the correlation between each gene expression and miRNA expression. We filtered these results to only interactions with a p-value < 0.05 followed by an interquartile range (IQR) filter of < 0.25. With the resulting 12,799 miRNA-gene interactions, we filtered for only miRNA-gene interactions predicted with very high confidence in mirDIP, which yielded 807 interactions. Of these interaction, 64 miRNA interacted with genes in 11 modules (Fig. 4).

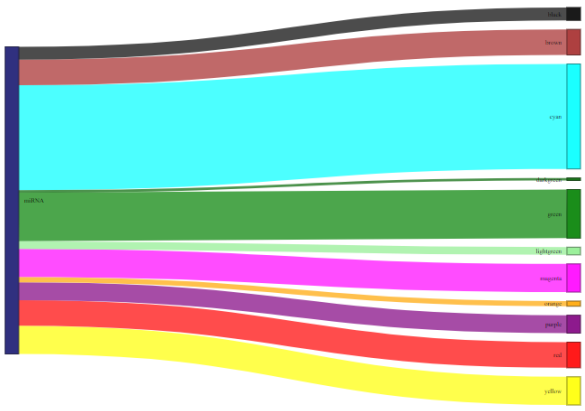

Figure 4. Sankey plot of miRNA with expression that are highly correlated to gene expression within 11 modules. Edge width corresponds to the number of miRNA-gene interactions within a given module. (need to fix color scheme, make labels more legible and add number of genes per node)

Further, we identified significant protein-protein interactions by filtering all gene-gene interactions to only those with a p-value < 0.05 followed by an interquartile range (IQR) filter of > 0.75. This yielded 4,809 gene-gene interactions, which were filtered to only include interactions with a confidence threshold of 0.1 within the InBio Map database. We merged the 2,738 gene-gene interactions and the 807 miRNA-gene interactions to construct a network in Cytoscape (Fig. 5).

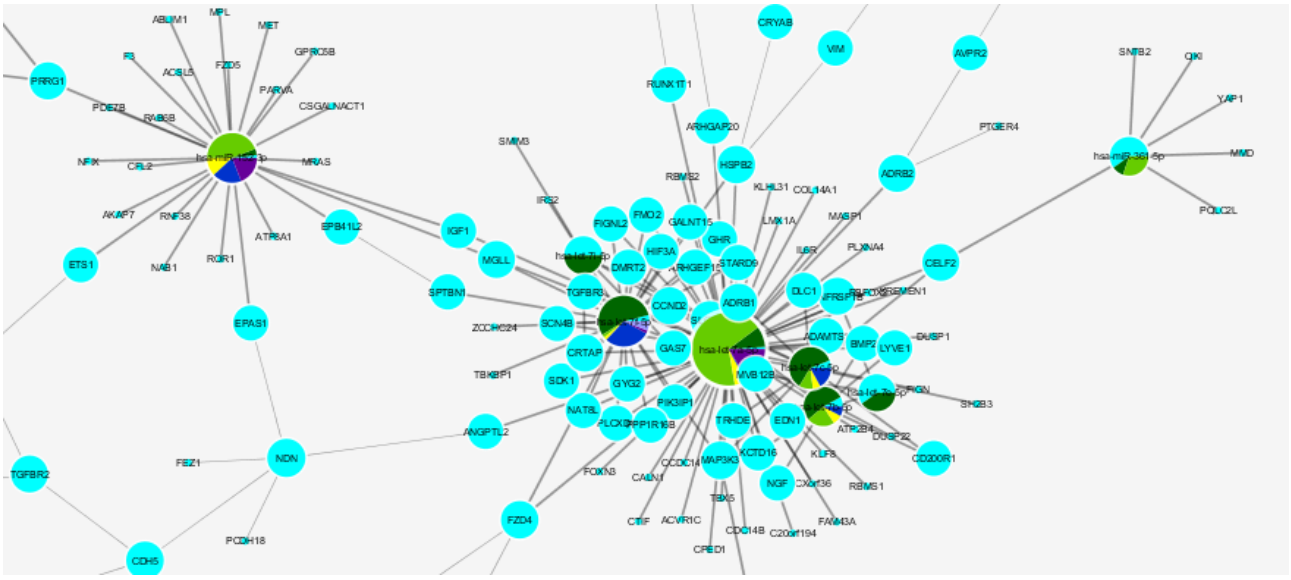

Figure 5. A portion of the network for the cyan module. Nodes represent miRNA or genes. Pie charts represent percentage of connections a given miRNA or gene has within each module. Colors represent each module. (need to fix color scheme to match WGCNA, keep nodes from overlapping, and make labels legible.)

Discussion

Methods

Discovery data

To obtain the discovery dataset, the TCGAbiolinks package was used to query the TCGA database with the following parameters, *GDCquery(project = "TCGA-BRCA", data.category = "Transcriptome Profiling", data.type = "Gene Expression Quantification", workflow.type="HTSeq - Counts", legacy = FALSE)*. The query results were subsetted to only patients with paired mRNA-miRNA, matched normal tissue-primary tumor tissue data. This included 101 female and one male patient from which we selected only female patients. Instances of multiple count datasets for a single patient, tumor type, and extracted molecule combination were filtered to include only the most advantageous aliquots. This resulted in four dataset IDs per patient, including counts from normal tissue mRNA and miRNA, and primary tumor tissue mRNA and miRNA for a total of 408 datasets (Table X). These datasets were retrieved from the recount2 resource (Collado-Torres, 2017) to minimize differences in bioinformatics techniques between the discovery and test datasets.

Test data

To confirm the preservation of modules identified in the discovery dataset, we constructed a test dataset from the recount2 resource. We retrieved all GTEx RNA-seq counts obtained from NAT tissue of females, which included 59 individuals. Additionally, the TCGA database was queried as previously described. The results were filtered to exclude any patients used within the discovery dataset and include only samples retrieved from the primary tumor tissue of females. From these results, we randomly selected 59 patients and retrieved RNA-seq counts from recount2 for the most advantageous aliquots. Both the GTEx and TCGA datasets were merged before confirming module preservation (Table X).

Expression value normalization and adjustment of covariates

Raw counts were filtered for low expression with a threshold of > 0 expression in at least 10% of patients. Data was normalized with DESeq2 (Love, 2014) functions *estimateSizeFactors()*, *estimateDispersions()*, and finally *varianceStabilizingTransformation()*. Batch effects were investigated with R package *ExpressionNormalizationWorkflow* (Murugesan, 2018). (See supplement for details.) No surrogate variables were removed

from the data.

Weighted gene co-expression network analysis (WGCNA)
Signed co-expression networks were developed with R/WGCNA. Pairwise correlations between genes were calculated with biweight midcorrelation. Scale-free topology fit was used to determine a power of five, which was then used to calculate pairwise topological overlap. Coexpression modules were identified by the *cutreeDynamic* function with the following parameters minClusterSize = 40, cutHeight = 0.99. Each module was assigned a module eigengene, and modules with significantly correlated eigengenes were merged with the *mergeCloseModules()* function. The biweight midcorrelation of each gene to the module eigengene defined the module membership of each gene. Module preservation was determined by the Zsummary score, which is defined by summarizing the Z score,

$$Z = \frac{observed - mean_{permutated}}{sd_{permutated}},$$

from each measure. We used a permutation value of 200. Evidence of module preservation was defined as strong, weak to moderate, or no preservation based on Zsummary > 10, $2 >$ Zsummary < 10, or Zsummary < 2, respectively.

Prediction of miRNA targets
Prediction of miRNA targets was conducted with the scored miRNA-gene interactions downloaded from mirDIP Explore v.4 (http://ophid.utoronto.ca/mirDIP/download.jsp). This included four confidence classes, "very high", "high", "medium", and "low" that defined individual miRNA target predictions based on integrated scores inferred from prediction confidence measures across 30 databases. To select miRNA for target prediction, miRNA-gene interactions with p-value < 0.05 were identified. From this subset miRNA in the $<$ 0.25 interquartile range were selected. With these miRNA, targets with a confidence class of "very high" in mirDIP were predicted.

Protein-protein interaction (PPI) analysis
PPI analysis was performed with the InBio Map core database downloaded from https://www.intomics.com/inbio/api/data/m
To select genes for PPI analysis, gene-gene interactions with p-value < 0.05 were identified. From this subset genes in the > 0.75 interquartile range were selected. With these genes, protein-protein interactions with a confidence score of 0.1, as recommended by InBio Map, were predicted.

Network construction
To setup data for network construction, all miRNA-gene interactions were filtered to include only interactions predicted by mirDIP. And all gene-gene interactions were filtered to only those predicted by InBio Map. After merging these results, we visualized the network within Cytoscape. Topological parameters like node degree were used to analyze the network with the Cytoscape plug-in Network Analyzer (Assenov, 2008). In each network, a gene or miRNA is represented as a node and the interactions between nodes are defined as edges. Degree represents the number of edges connected to a given node, and was used to define the size of nodes. The weight of edges that connect genes were kept constant, while the edges between miRNA and genes were weighted to reflect strength of correlation; thicker edges represent stronger negative correlations. Node colors reflect the percentage of connections a given node has within each module.

Gene set enrichment analysis
GO enrichment in biological processes, molecular functions, and cellular components was conducted using g:Profiler (Reimand, 2016). Enrichment analysis included a statistical background of all genes obtained after filtering for low counts and normalization. To obtain p-values, a hypergeometric distribution was used and the default method, g:SCS, was used to correct for multiple testing.

References

Wang, W., Jiang, W., Hou, L., Duan, H., Wu, Y., Xu, C., ... Zhang, D. (2017). Weighted gene co-expression network analysis of expression data of monozygotic twins identifies specific modules and hub genes related

to BMI. BMC Genomics, 18, 872. http://doi.org/10.1186/s12864-017-4257-6

Li, T., Wernersson, R., Hansen, R. B., Horn, H., Mercer, J., Slodkowicz, G., ... Lage, K. (2017). A scored human protein–protein interaction network to catalyze genomic interpretation. Nature Methods, 14(1), 61–64. http://doi.org/10.1038/nmeth.4083

Philipp, O., Osiewacz, H. D., Koch, I. (2016). Path2PPI: an R package to predict protein–protein interaction networks for a set of proteins. Bioinformatics, 32(9), 1427–1429. http://doi.org/10.1093/bioinformatics/btv765

Razick, S., Magklaras, G., Donaldson, I. M. (2008). iRefIndex: A consolidated protein interaction database with provenance. BMC Bioinformatics, 9, 405. http://doi.org/10.1186/1471-2105-9-405

Wang, L., Yang, L., Peng, Z., Lu, D., Jin, Y., McNutt, M., Yin, Y. (2015). cisPath: an R/Bioconductor package for cloud users for visualization and management of functional protein interaction networks. BMC Systems Biology, 9(Suppl 1), S1. http://doi.org/10.1186/1752-0509-9-S1-S1

Cowley, M. J., Pinese, M., Kassahn, K. S., Waddell, N., Pearson, J. V., Grimmond, S. M., ... Wu, J. (2012). PINA v2.0: mining interactome modules. Nucleic Acids Research, 40(Database issue), D862–D865. http://doi.org/10.1093/nar/gkr967

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., Jensen, L. J., ... von Mering, C. (2016). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic acids research, 45(D1), D362-D368.

Liang Chen, Liisa Heikkinen, Changliang Wang, Yang Yang, Huiyan Sun, Garry Wong; Trends in the development of miRNA bioinformatics tools, Briefings in Bioinformatics, , bby054, https://doi.org/10.109x3/bib/bby054

Tokar, T., Pastrello, C., Rossos, A., Abovsky, M., Hauschild, A. C., Tsay, M., Lu, R., ... Jurisica, I. (2017). mirDIP 4.1-integrative database of human microRNA target predictions. Nucleic acids research, 46(D1), D360-D370.

O'Brien, J., Hayder, H., Zayed, Y., Peng, C. (2018). Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Frontiers in endocrinology, 9, 402. doi:10.3389/fendo.2018.00402

Lim, L. P. et al. (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature 433, 769–773.

Kloosterman WP, Plasterk RH. The diverse functions of microRNAs in animal development and disease, Dev. Cell , 2006, vol. 11 (pg. 441-450)

Gebert LFR, MacRae IJ (2018) Regulation of microRNA function in animals. Nature reviews.Molecular cell biology.

Mendell, J. T., Olson, E. N. (2012). MicroRNAs in stress signaling and human disease. Cell, 148(6), 1172-87.

Schwarzenbach, H., Nishida, N., Calin, G. A. Pantel, K. (2014) Clinical relevance of circulating cell-free microRNAs in cancer. Nat. Rev. Clin. Oncol. 11, 145–156.

Shalgi, R., Lieber, D., Oren, M., Pilpel, Y. (2007). Global and local architecture of the mammalian microRNA-transcription factor regulatory network. PLoS computational biology, 3(7), e131.

Lai, X., Schmitz, U., Gupta, S. K., Bhattacharya, A., Kunz, M., Wolkenhauer, O., Vera, J. (2012). Computational analysis of target hub gene repression regulated by multiple and cooperative miRNAs. Nu-

cleic acids research, 40(18), 8818-34.

Ebert, M. S., Sharp, P. A. (2012). Roles for microRNAs in conferring robustness to biological processes. Cell, 149(3), 515-24.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., Pandolfi, P. P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?. Cell, 146(3), 353-8.

Rao, V. S., Srinivas, K., Sujini, G. N., Kumar, G. N. (2014). Protein-protein interaction detection: methods and analysis. International journal of proteomics, 2014, 147648.

Witkos, T. M., Koscianska, E., Krzyzosiak, W. J. (2011). Practical Aspects of microRNA Target Prediction. Current molecular medicine, 11(2), 93-109.

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., ... Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature genetics, 45(10), 1113-20.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., ... Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research, 13(11), 2498-504.

GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nature genetics, 45(6), 580-5.

Aran, D., Camarda, R., Odegaard, J., Paik, H., Oskotsky, B., Krings, G., Goga, A., Sirota, M., ... Butte, A. J. (2017). Comprehensive analysis of normal adjacent to tumor transcriptomes. Nature communications, 8(1), 1077. doi:10.1038/s41467-017-01027-z.

L. Collado-Torres, A. Nellore, K. Kammers, S. E. Ellis, M. A. Taub, K. D. Hansen, A. E. Jaffe, B. Langmead, J. T. Leek (2017), Reproducible RNA-seq analysis using recount2. Nat. Biotechnol. 35, 319–321. doi:10.1038/nbt.3838pmid:28398307.

Assenov Y, Ramirez F, Schelhorn SE, Lengauer T, Albrecht M (2008) Computing topological parameters of biological networks. Bioinformatics 24: 282–284.

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., Vilo, J. (2016). g:Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic acids research, 44(W1), W83-9.

Love, M. I., Huber, W., Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology, 15(12), 550.

Murugesan K (2018). ExpressionNormalizationWorkflow: Gene Expression Normalization Workflow. R package version 1.6.0.