Identifying Functional Cancer-Specific miRNA-mRNA Interactions in Testicular Germ Cell Tumor

Nafiseh Sedaghat¹, Mahmood Fathy¹, Mohmmad-H Modarressi² and Ali Shojaie³

¹ Computer Engineering School, Iran University of Science and Technology;

² Medical Genetics Department, Tehran University of Medical Science;

³ Department of Biostatistics, University of Washington.

May 20, 2016

Abstract

Testicular cancer is the most common cancer in men aged between 15 and 35 and more than 90% of testicular neoplasms are originated at germ cells. Recent research has shown the impact of microRNAs (miRNAs) in different types of cancer, including testicular germ cell tumor (TGCT). MicroRNAs are small non-coding RNAs which affect the development and progression of cancer cells by binding to mRNAs and regulating their expressions. The identification of functional miRNA-mRNA interactions in cancers, i.e. those that alter the expression of genes in cancer cells, can help delineate post-regulatory mechanisms and may lead to new treatments to control the progression of cancer. A number of sequence-based methods have been developed to predict miRNA-mRNA interactions based on the complementarity of sequences. While necessary, sequence complementarity is, however, not sufficient for presence of functional interactions. Alternative methods have thus been developed to refine the sequence-based interactions using concurrent expression profiles of miRNAs and mRNAs. This study aims to find functional cancer-specific miRNA-mRNA interactions in TGCT. To this end, the sequence-based predicted interactions are first refined using an ensemble learning method, based on two well-known methods of learning miRNA-mRNA interactions, namely, TaLasso and GenMiR++. Additional functional analyses were then used to identify a subset of interactions to be most likely functional and specific to TGCT. The final list of 13 miRNA-mRNA interactions can be potential targets for identifying TGCT-specific interactions and future laboratory experiments to develop new therapies.

Keywords- miRNA-mRNA interaction, Testicular germ cell tumor, Gene ontology, KEGG pathways, Disease ontology.

1 Introduction

Testicular germ cell tumor (TGCT) is the most common malignancy in males between 15 and 35, and comprises approximately 98% of all the testicular neoplasms. The American Cancer Society estimated that in 2015, ~8430 new cases of TGCT would be diagnosed and ~380 men would die from TGCT. Currently, about 1 of every 263 males develops testicular cancer at some point during their life. Confirmed risk factors of TGCT include family history, height (at adulthood), infertility and previous contralateral testicular cancer [1–3].

Apart from the contribution of these risk factors TGCT is *strikingly heritable*. Having a brother (father) diagnosed with TGCT results in an estimated 8 to 10 fold (4 to 6 fold) increased risk of TGCT for men. These numbers are significantly higher than the largest familial relative risk of adult cancers for first-degree relatives, which is estimated to 1.5 to 2.5 fold [1–3]. Various research groups have thus investigated the genetic risk factors of TGCT with the goal of identifying genes associated with TGCT [4–6]. In parallel, it

has been discovered that microRNAs (miRNAs) can act as oncogenes or tumor suppressors by regulating the expression of genes in various types of cancers, including TGCT [7–9]. miRNAs have also been used to develop new therapies based on control of the expression levels of target genes by modifying the expression levels of their regulating miRNAs [10–13].

miRNAs regulate the expression levels of their target genes, which are identified by partial complementarity of binding sites. miRNA to mRNA binding results in cleavage of the target mRNA and inhibits the translation of mRNA to protein. Regulation by miRNA can thus be key to many biological processes and diseases, especially cancers. It has been estimated that 30% of all genes in human are regulated by miRNAs [14]. Investigating miRNAs and their relationships with mRNAs can thus help delineate underlying epigenetic regulations and disease mechanisms.

While partial complementarity is necessary for miRNA-mRNA interaction, only a small fraction of hundreds, or thousands of possible DNA sequence matches result in *functional* miRNA-mRNA interactions, where the miRNA expression levels change the mRNA transcript levels of target genes. This is because occurrence of functional interactions depends on additional binding factors. Moreover, only the smaller subset of bindings plays a role in tumor development and progression [15, 16]. Sequence-based predicted targets should thus be further filtered to identify more probable functional miRNA-mRNA interactions that are involved in cancer initiation and progression, or, are *cancer-specific*. Identifying such functional cancer-specific miRNA-mRNA interactions in TGCT can shed light into novel mechanisms of TGCT initiation and progression and facilitate development of new therapies.

Various computational methods have been developed to refine sequence-based predicted interactions based on additional information, such as concurrent expression profiles of miRNAs and mRNAs on the same set of samples [17–21]. Correlation-based approaches, using, e.g., the Pearson or Spearman correlation between expression of miRNAs and mRNAs have been used to identify miRNA-mRNA interactions. Often only negative correlations are considered, following the general assumption that miRNAs repress the expression levels of mRNAs [17, 18]. A drawback of correlation coefficients is that they measure *linear* associations between two sets of observations (or their ranks) and are thus not suited for non-linear associations. Mutual Information (MI) can be alternatively used to identify associations between miRNAs and mRNAs. However, MI quantifies the common information between miRNA and mRNA expressions and is always non-negative. It thus fails to determine the direction of miRNA and mRNA effects [19].

Bayesian methods and regularized estimation procedures are two other classes of approaches for refining sequence-based interactions. Bayesian methods facilitate statistical inference by incorporating the prior knowledge into a probabilistic framework. *GenMiR++* [20] is a well-known Bayesian model that scores each miRNA-mRNA pair to determine if the expression of miRNA explains the expression of target mRNA. Higher scores in this framework indicate interactions that are more likely to occur. *miRNA-Target LASSO* (*TaLasso*) [21] is, on the other hand, a popular regularized estimation approach for identifying miRNA-mRNA interactions. TaLasso is based on the Least Absolute Shrinkage and Selection Operator (LASSO) [22] and couples the LASSO’s ℓ_1 penalty with a *non-positivity constraint*.

In this paper, we propose a computational framework for identifying functional, cancer-specific miRNA-mRNA interactions in TGCT. In the first step, we utilize miRNA and mRNA expression profiles to refine the sequence-based predicted interactions using a consensus-learning approach for discovering biological networks. More specifically, we consider consensus miRNA-mRNA interactions from TaLasso and GenMiR++. The rationale for considering these two estimation procedures is further discussed in Section 2. We then further screen out the identified interactions to find *condition-specific* miRNA-mRNA interactions, by considering interactions *not* identified in normal samples. Due to lack of normal testis samples, our validation step is based on normal samples from *prostate* tissues, which are expected to have close expression profiles to normal testis samples. The second step of our analysis entailed validation of identified interactions by investigating associations between genes in identified miRNA-mRNA interactions and known TGCT genes. To this end, we examine the enrichment of identified mRNAs from five perspectives: 1) gene functions by analyzing Gene Ontology (GO) terms, 2) disease association by analyzing Disease Ontology (DO) terms, 3) gene co-expression network, 4) physical interaction network, and 5) pathway enrichment analysis. Investigating the association between the identified genes and known TGCT genes helps identify *potential targets*

Figure 1: The proposed computational framework to identify functional cancer-specific miRNA-mRNA interactions in TGCT. The first step includes refining the sequence-based predicted interactions using both an ensemble learning approach based two existing approaches, namely, TaLasso and GenMiR++, as well as exclusion of interactions in normal samples. The second step includes investigating the association between gene targets of identified interactions from step 1 and known TGCT genes in different aspects of functionality e.g. GO/Do and pathway analyses.

of miRNAs in TGCT. We completed this step by examining the previous literature in search of evidence for the involvement of identified genes in TGCT.

The rest of the paper is organized as follows. In Section 2, the data set used in the analysis and preprocessing steps are described. We also provide details on steps of the proposed framework to identify functional cancer-specific miRNA-mRNA interactions, briefly described above. Section 3 contains the main results of the study. We conclude the paper with a discussion in Section 4.

2 Materials and Methods

2.1 Overview of the Proposed Computational Framework

Figure 1 outlines the main steps of the proposed computational framework. As shown in the figure and briefly discussed in the Introduction, the framework consists of two main steps: 1) identifying functional cancer-specific miRNA-mRNA interactions and 2) validating the identified interactions by associating the

gene targets of miRNA-mRNA interactions to known TGCT genes from various perspectives.

In the first step, predicted sequence-based miRNA-mRNA are filtered to identify functional interactions by examining the miRNA and mRNA expression levels. To this end, we use a consensus-based approach by considering the intersection of miRNA-mRNA interactions from TaLasso and GenMiR++. GenMiR++ and TaLasso are two popular approaches that span two broad classes of approaches for reconstructing miRNA-mRNA interactions. GenMiR++ identifies interactions based on *marginal* associations between miRNA and mRNA expression levels; specifically GenMiR++ uses a Bayesian model to assess the significance of the correlation between expression levels of pairs of miRNA and mRNA, without adjusting for other miRNA or mRNA levels. On the other hand, TaLasso utilizes a penalized regression framework to identify miRNA-mRNA interactions based on *conditional* associations; more specifically, it selects for large (negative) partial correlations between miRNA-mRNA after adjusting for the expression levels of other miRNA levels. [23] recently investigated advantages and shortcomings of network reconstruction methods based on marginal and conditional associations and found that consensus learning of interactions can improve the reliability of identified interactions. The underlying hypothesis in this step is that miRNA-mRNA interactions identified by both GenMiR++ and TaLasso are more likely functional.

Given the consensus miRNA-mRNA interactions, the algorithm then seeks to find cancer-specific interactions by filtering out interactions that are identified in normal samples. As pointed out in the Introduction, due to lack of normal samples from testis, in this paper we use publicly available normal prostate samples, which are expected to have similar expression profiles to normal testis samples. The second step of the algorithm consists of five different enrichment analyses to identify genes in functional cancer-specific miRNA-mRNA interactions associated with known TGCT genes. Detailed description for each of the above steps, as well as data preprocessing procedures are presented in the remainder of this section.

2.2 Data Pre-processing

Matched miRNA and mRNA RNA-Seq files were downloaded in level 3.0 format for TGCT from The Cancer Genome Atlas (TCGA) database¹. RNA-Seq files in level 3.0 provide *read_counts*. Level 3.0 data from TCGA have been carefully pre-processed and checked for quality using various metrics, such as base quality per cycle and average coverage per human gene². Thus, following [24], we did not perform any additional data pre-processing. However, prior to network analysis, we standardized the *Read_counts* in each sample, and applied $\log_2$ -transformation on $(read_counts + 1)$. The resulting data matrices for miRNA and mRNA were of dimensions $[1046 \times 156]$ and $[20531 \times 156]$, respectively; the first dimension in the above matrices shows the number of miRNAs and mRNAs, respectively, and the second dimension shows the number of common samples.

Normal prostate samples were downloaded from TCGA database in order to perform validation analysis and were similarly processed. The dimension of miRNA and mRNA matrices are $[1046 \times 52]$ and $[20531 \times 52]$, respectively. The miRNA and mRNA data from normal prostate tissue are based on identical platforms to TGCT, which justifies the use of the same preprocessing steps.

2.3 Refinement of Sequence-based Predicted Interactions

Sequence-based predicted interaction matrix was constructed based on putative interaction matrix available through the TaLasso homepage³. The most recent *validated* interaction data were downloaded from *miRWalk 2.0*⁴, and were used as the experimentally validated interaction matrix.

After collecting the required data, sequence-based intersections between genes and miRNAs were obtained. Then, miRNAs that did not target any mRNAs were removed from the data sets. Similarly, mRNAs

¹<https://tcga-data.nci.nih.gov/tcga/>

²For more information visit <https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp>.

³talasso.cnb.csic.es

⁴<http://zmf.umm.uni-heidelberg.de/apps/zmf/mirwalk2/holistic.html>

that were not targeted by miRNAs were also removed. The number of the remaining miRNAs and mRNAs in TGCT and normal prostate tissues were 192 and 11339, respectively. *TaLasso* and *GenMiR++* codes were run on the resulting preprocessed data to refine the sequence-based predicted interactions.

An *enrichment analysis* was then performed to assess the interactions identified using either method. This analysis determines the extent to which identified interactions corroborate with validated miRNA-mRNA interactions and consists of the following steps:

- For $n \in \{100, 200, \dots, 1000\}$, select the top n ranked interactions and determine the number of experimentally validated interactions among them, denoted by p .
- For $i = 1, \dots, B$, select n interactions randomly from all possible interactions and find the number of experimentally validated interactions among them; denote this number by $p^i, i = \{1, 2, \dots, 100\}$.
- Calculate the enrichment p-value as follows:

$$P\text{-value} = \frac{1}{B} \sum_{i=1}^B I(p \geq p^i) \quad (1)$$

where $I(\cdot)$ is the indicator function.

$B = 100$ random samples was used in our analyses. The p-value from the enrichment analysis evaluates the significance of the result from each method. However, it should be noted that experimentally validated interactions may not be cancer-specific. Thus, while the p-values resulting from this analysis can be used to compare methods of learning functional miRNA-mRNA interactions, the results should be interpreted cautiously when assessing cancer-specific interactions.

To obtain cancer-specific miRNA-mRNA interactions, identified interactions using the proposed consensus approach were next screened for specificity to TGCT. In short, this is achieved by excluding miRNA-mRNA interactions that are also present in *normal* samples. As mentioned, due to unavailability of matched miRNA and mRNA data from normal testis samples, normal prostate samples were used in this paper. Prostate tissue was chosen because both testis and prostate are male-specific organs involved in reproduction. Thus, the normal testis and prostate tissues are expected to have similar epigenetics, including miRNA and mRNA expression profiles. In fact, previous research has identified a number of genes expressed in both tissues [25–34]. Normal prostate samples can thus be used in lieu of normal testis samples to identify cancer-specific interactions in TGCT. The process of identifying normal miRNA-mRNA interactions is similar to that of TGCT interactions; specifically, common interactions between top-ranked interactions predicted by *TaLasso* and *GenMiR++* using data from normal prostate samples were obtained. The list of TGCT interactions were then contrasted with interactions in normal cells to exclude interactions present in both tissue types.

2.4 GO/DO Term Enrichment Analysis

An important post-genomic analysis is the study of functional relationship between genes. Relationships between genes can be analyzed from various perspectives and various types of relationships exists between genes.

The Gene Ontology (GO) consortium [35] aims to annotate consistent descriptions of gene products via the structured gene ontology. The ontology includes three main groups: *cellular components*, *molecular function*, and *biological process*. *Cellular component* terms determine the component of the cell where the gene product is expressed, such as “proteasome” and “ribosome”. *Molecular function* terms describe molecular-level activities of a gene product, such as “binding activity”. *Biological process* terms, on the other hand, include a series of events with a defined start point and end point such as “signal transduction”. Each group contains many terms with unique names and identifiers which have been organized into a directed acyclic graph (DAG). The hierarchy of nodes in this graph represents the generality of the terms, with terms

higher in the hierarchy being more general. The edges of this graph represent various relationships between the GO term.

Similar to GO, Disease Ontology (DO) [36] provides a formal gene-disease association to bridge the gap between biological research and clinical applications as a post-genomic analysis. DO provides better understanding of disease states by placing heritable disorders in the context of other related diseases. Investigation of DO terms for genes thus helps delineate whether genes are likely to be associated with a specific disorder/disease. DO is also helpful in integrating disparate data through disease terminologies. Similar to GO, disease concepts in DO are organized as a DAG, such that the terms near to the root are more general terms and those farther away from the root are more specific.

To perform GO/DO term analysis, common GO/DO terms were sought between two lists of genes. To determine if the overlaps were statistically significant, enrichment analyses were performed for each common GO/DO term. In this regard, a list of genes equal to the number of query genes (13 genes, as described in Section 3) was randomly selected, from all genes included in human organism. Then, GO/DO terms were obtained for the list of randomly selected genes and p-value was calculated as

$$P\text{-value}(T) = \frac{1}{B} \sum_{i=1}^B I(T \in \text{list}_i) \quad (2)$$

where T is the GO/DO term for which p-value is calculated for, and list_i is the list of GO/DO terms for the i^{th} randomly selected set of genes. Also, $I(\cdot)$ is the indicator function, which is 1 if $T \in \text{list}_i$ and 0 otherwise. The results in Section 3 are based on $B = 100$ random samples.

In this analysis we used the GOSemSim [37] and DOSE [38] R-packages to analyze GO and DO terms, respectively. GOSemSim measures functional similarities among genes based on GO terms based on a *semantic* comparison. DOSE similarly provides semantic similarity measurements of DO terms for gene products.

2.5 Co-expression and Physical Interaction Analysis

Biological interactions can be characterized in many ways. Two important notions are captured in *physical interaction networks* and *gene co-expression networks* (GCN). Physical interaction networks represent binding affinity of genes and their protein products, based on DNA sequences and protein. They thus provide insight into *potential* interactions that govern cellular functions. GCNs, on the other hand, capture similarities in activities of functions in a given biological condition. They thus provide valuable information functional similarities of genes, sharing of transcription factors, or membership in biological pathways [39, 40]. The complementary information about genetic interactions from these two networks can help predict biological functions of a given gene set and potential involvement in disease mechanisms.

We used GeneMANIA⁵ [41, 42] to construct both physical and co-expression networks in TGCT samples. GeneMANIA utilizes information extracted from well-known databases such as Gene Expression Omnibus (GEO)⁶, Pathway Commons⁷, and BioGRID⁸. It can also add functionally related genes to a gene list, to provide a well-informed network.

2.6 Pathway Over-representation Analysis

Pathway analysis is another popular post-genomic approach, which helps investigate the combined effect of genes in a gene set or pathway. One of the earlier approaches for gene set analysis is *over-representation analysis* (ORA) [43]. Given a target set of genes T , ORA tests whether members of gene sets/pathways of

⁵<http://www.genemania.org>

⁶<http://www.ncbi.nlm.nih.gov/geo/>

⁷www.pathwaycommons.org

⁸thebiogrid.org

interest are over-represented in the target list. To this end, ORA uses a *counting* strategy and calculates the probability of observing genes in a pathway by chance under the *hypergeometric* distribution. The p-value for each gene set/pathway P is thus calculated as

$$P\text{-value}(P) = 1 - \sum_{l=0}^x \frac{\binom{M}{l} \binom{N-M}{K-l}}{\binom{N}{K}}, \quad (3)$$

where N is the total number of genes in all the pathways, M is the number of genes in pathway P , K is the number of genes in the target set, and x is the number of genes present in both the target set T and the pathway of interest P .

ORA has a number of limitations and more advanced methods have been recently proposed for pathway enrichment analysis [44]. However, the hypergeometric test in ORA is particularly suitable in settings where the target set T is obtained from external information and is fixed, which is the case in our analysis. In particular, we perform ORA to identify biological pathways that are over-represented in the list of genes identified based on miRNA-mRNA interactions, as well as those over-represented in known TGCT genes. In this analysis, we assess the over-representation of known TGCT genes, as well as genes identified from miRNA-mRNA interactions in pathways from the *Kyoto Encyclopedia of Genes and Genomes* (KEGG). Here, P is a given KEGG pathway, and the target set T is either the set of known TGCT genes or genes from identified miRNA-mRNA interactions. A small p-value from (3) then indicates a smaller chance of over-representation of members of the KEGG pathway P in one of the target sets T . The KEGG pathways and their members were obtained using the R-package *graphite* [45].

3 Results

3.1 Refinement of Sequence-based Predicted Interactions

Figure 2(a) shows the number of common interactions between validated miRNA-mRNA interactions and the top n interactions identified by TaLasso and GenMiR++ for $n \in \{100, 200, \dots, 1000\}$. Enrichment analysis of the results for both methods and for all values of $n \in \{100, 200, \dots, 1000\}$ results in small p-value (≤ 0.01), confirming that both methods are statistically more accurate than random guessing.

The number of shared interactions between TaLasso and GenMiR++, as well as the number of validated interactions included in them are shown in Figure 2(b). It can be seen that, roughly, one-tenth of the number of top n predicted interactions are found using both estimation methods. Moreover, the growth rates of the curves corresponding to the number of shared interactions and the fraction of validated shared interactions are not the same. These results suggest that many of the shared interactions are not yet validated and there is a potential to identify new functional miRNA-mRNA interactions.

The results in Figure 2 suggest that for $n \leq 500$ top ranked pathways, the number of validated interactions from TaLasso and GenMiR++ are the same. Of the top 500 interactions identified by both methods, 49 interactions are in common; 12 of these interactions are *experimentally* validated, and the remaining 37 interactions are novel⁹. Figure 3 shows the network of these 37 interactions, created using the Cytoscape platform [46]. The 37 identified interactions occur among 31 miRNAs and 13 genes, which we henceforth refer to as *query genes*. We limit our further analyses to these 37 interactions and 13 genes.

To validate these interactions, the same procedure was also run on *normal prostate* tissue samples from TGCA. In this set of samples, 52 interactions were common between the top 500 interactions from TaLasso and GenMiR++ interactions¹⁰. Interestingly, no common miRNA-mRNA interaction were found between TGCT and normal prostate samples. The interactions from normal prostate samples involve 7 genes, namely,

⁹The list of interactions is available in Supp. 1 and 2.

¹⁰The list of interactions is available in Supp. 3.

Figure 2: (a) Number of validated interactions among top-ranked interactions; (b) Number of common interactions among top-ranked interactions between *TaLasso* and *GenMiR++*.

Figure 3: The common interactions in top-ranked 500 interactions between *TaLasso* and *GenMiR++*. Green nodes represent mRNAs/genes and yellow nodes are miRNAs.

EEF1A1, *ACPP*, *KLK3*, *RDH11*, *SLC45A3*, *HSPA1A*, and *AZGP1*. Only *EEF1A1* is common between this list and our query genes. These results suggest that the identified miRNA-mRNA interactions are specific to TGCT.

It is believed that miRNAs do not play an *independent* role in cell processes, but act only through their target genes. Thus, to examine the relationship between the obtained interactions and TGCT we primarily focus on gene targets of identified TGCT interactions. To link our query genes to TGCT, we investigated

Table 1: TGCT and query genes. The first nine TGCT genes are from *DGA* database and remains are from *MalaCards* database.

TGCT Genes	Query Genes
ATF7IP	GAPDH
DMRT1	ENO1
ESR1	EEF2
ESR2	RPL8
KITLG	LDHB
LHCGR	EEF1A1
POLG	FN1
TERT	RPS17
BCL10	RPS18
KIT	RPS19
STK10	RPLP0
STK11	RPL19
FGFR3	HSP90AB1

known TGCT genes with proven association with the disease. Using two databases that describe the association between genes and diseases, namely, *Disease and Genes Annotations (DGA)*¹¹ and *MalaCards*¹², we identified 13 genes that have been declared to be associated with TGCT in both databases¹³. Table 1 lists TGCT and query genes.

3.2 GO/DO Term Enrichment Analysis

Figure 4(a) shows the common GO terms between our query and known TGCT genes, along with the name of the GO terms and their corresponding p-values. Only 5 GO terms were significantly enriched for both sets with a $p\text{-value} < 0.01$ using the permutation test in (2). We next measured the similarity between our query genes and those related to TGCT based on their common GO terms using the method of *Wang* [47], which is implemented in the *GOSemSim* R-package. Wang’s method calculates the similarity between two GO terms by considering their location in the GO graph and their relationships with their ancestors. The resulting similarity matrix is depicted in Figure 4(b).

In addition to GO terms, the common DO terms between the two lists of genes were also investigated. Similar to the GO term analysis, the association of DO terms were assessed using the permutation test in (2). However, two query genes and one TGCT gene were not included in the DO database and were hence removed from the analysis. Figure 5(a) shows the common DO IDs along with their names and corresponding p-values. The similarity matrix of DO terms is depicted in Figure 5(b).

¹¹<http://dga.nubic.northwestern.edu/>

¹²<http://www.malacards.org/>

¹³The TGCT genes in DGA correspond to the id DOID:5557, “testicular germ cell cancer”, and TGCT genes in MalaCards correspond to “testicular germ cell tumor” disease.

GO:0042803: Protein homodimerization activity(0.13)
GO:0042802: Identical protein binding(0.07)
GO:0002020: Protease binding(**0.01**)
GO:0003700: Sequence-specific DNA binding transcription factor activity(0.13)
GO:0000049: tRNA binding(**0.01**)
GO:0005515: Protein binding(0.85)
GO:0003723: RNA binding(0.08)
GO:0017134: Fibroblast growth factor binding (**0.01**)
GO:0030235: Nitric-oxide synthase regulator activity (**0.01**)
GO:0019900: Kinase binding(**0.01**)
GO:0019901: Protein kinase binding(0.07)

(a)

(b)

Figure 4: (a) The network illustrates the shared GO terms between query genes and TGCT genes. Yellow nodes are GO terms, green nodes are TGCT genes, and blue nodes are query genes. The name of GO terms are in the below of the network. The numbers in parenthesis show their p-values. (b) GO term similarity between TGCT genes and query genes. TGCT genes and query genes are in rows and in columns, respectively.

DOID:684: Hepatocellular carcinoma(0.16) DOID:5223: Infertility(0.02)
 DOID:114: Heart disease(0.08) DOID:4971: Myelofibrosis(0.02)
 DOID:848: Arthritis(0.13) DOID:225: Syndrome(0.32)
 DOID:4: Disease(0.3) DOID:5419: Schizophrenia(0.17)
 DOID:1909: Melanoma(0.15) DOID:2355: Anemia(0.04)
 DOID:5844: Myocardial infarction(0.06) DOID:7148: Rheumatoid arthritis(0.12)
 DOID:3963: Thyroid carcinoma(0.03) DOID:3858: Medulloblastoma(0.02)
 DOID:3310: Atopic dermatitis(0.03) DOID:1936: Atherosclerosis(**0.01**)
 DOID:326: Ischemia(0.07) DOID:3459: Breast carcinoma(0.11)
 DOID:1240: Leukemia(0.16) DOID:3905: Lung carcinoma(0.03)
 DOID:3347: Osteosarcoma(0.05) DOID:1749: Squamous cell carcinoma(0.14)
 DOID:162: Cancer(0.53) DOID:10652: Alzheimer's disease(0.11)
 DOID:1168: Familial hyperlipidemia(**0.01**) DOID:3275: Thymoma(**0.01**)
 DOID:5082: Liver cirrhosis(**0.01**) DOID:3908: Non-small cell lung carcinoma(0.02)

(a)

(b)

Figure 5: (a) The network illustrates the shared DO terms between query genes and TGCT genes. Yellow nodes are DO terms, green nodes are TGCT genes, and blue nodes are query genes. The name of DO IDs are provided in the below of the network. The numbers in parenthesis show their p-values. (b) DO term similarity between TGCT genes and query genes. TGCT genes are in rows and query genes are in columns.

Table 2 summarizes the findings of GO and DO analyses in Figures 4 and 5. The figure shows the query and TGCT genes mapped to enriched GO/DO terms. The genes included are those mapped to significantly enriched GO/DO terms ($p\text{-value} < 0.01$), but have been filtered to only include genes from either TGCT or query genes with relatively large similarity (> 0.4) with members of the other set.

The joint mapping of TGCT and query genes to GO/DO terms highlight common functionality of the query genes with known TGCT genes. For instance, GO:0017134 is believed to “interact with a fibroblast growth factor” [35]. Fibroblast Growth Factors (FGFs) are known as hormonal factors and play a key role in regulating growth and development of several reproductive organs, including the testis [48]. Similarly, GO:0030235 is related to the *activity of nitric oxide synthase*. Nitric Oxide (NO) is a reactive radical gas mediating many biological functions. Of the different forms of Nitric Oxide Synthase (NOS) the endothelial form (eNOS) has a role in modulating sexual and reproductive function. It has also been suggested that NO might be involved in different testicular abnormalities, i.e. in inhibiting human sperm motility, in germ cell degeneration and in stress-impaired testicular steroidogenesis [49–51]. Finally, GO:0019900 is known to “interact with kinases, i.e., enzymes that catalyze the transfer of a phosphate group.” The protein kinase gene family is the most frequently mutated in human cancer and the role of mutated protein kinases in development of TGCT has been previously reported [52]. Of the reported DO terms, DOID:1168, or familial hyperlipidemia, is an inherited disorder that causes high cholesterol and high levels of triglycerides in the blood; previous studies have shown a relationship between male fertility and hyperlipidemia [53, 54].

Table 2: Significant GO/DO terms.

GO/DO Term	Query Genes	TGCT Genes	Similarity
GO:0002020	FN1	POLG, BCL10	[0.6-0.8]
GO:0017134	RPS19	FGFR3	[0.6-0.8]
GO:0019900	HSP90AB1	BCL10	[0.6-0.8]
GO:0000049	EEF1A1	TERT	[0.4-0.6]
GO:0030235	HSP90AB1	ESR1	[0.4-0.6]
DOID:3275	GAPDH	ESR1	[0.4-0.6]
DOID:1168, DOID:5082, DOID:1936	FN1	ESR1, ESR2, TERT	[0.4-0.6]

3.3 Co-expression and Physical Interaction Analysis

Figure 6(a) shows the gene co-expression network for query and TGCT genes. One of the TGCT genes, *FGFR3*, is absent in the network due to the unavailability of co-expression information about it in GeneMANIA. Interestingly, about 50% of the query genes have direct co-expression relationship with known TGCT genes. Figure 6(b) shows the physical interaction network between query and TGCT genes. Only 8 of 13 TGCT genes have known physical interactions in the GeneMANIA database; the 5 TGCT genes with no physical interactions are excluded from the network.

Table 3: Direct relationships between Query and TGCT genes in GCN and physical interaction network.

Query Genes	TGCT genes with direct co-expression relation	TGCT genes with direct physical interaction
GAPDH	-	ESR1
ENO1	-	ESR1, TERT
EEF2	STK11, KIT	ESR1
LDHB	KIT, POLG, BCL10	-
EEF1A1	ESR2	ESR1
FN1	LHCGR, TERT	-
RPS17	-	ESR1
RPS18	ESR2	ESR1, ESR2
RPS19	-	ESR1
RPLP0	-	ESR1, ESR2
RPL8	TERT	ESR1, FGFR3, ATF7IP
RPL19	KITLG	ESR1
HSP90AB1	-	ESR1, FGFR3, STK11

The observations from co-expression and direct physical interactions are summarized in Table 3. It can be seen that all query genes, except *LDHB* and *FN1*, have direct physical interactions with at least one of the Estrogen Receptor (*ESR*) genes, *ESR1* and *ESR2*. Among the query genes, *RPL8* and *HSP90AB1* have the highest number of direct interactions in the physical interaction network. On the other hand, *LDHB* and *FN1* have the highest number of direct relationships in GCN. However, these genes do not have any direct physical relationship with TGCT genes.

(a) Gene co-expression network based on GeneMANIA database.

(b) Physical interaction network based on GeneMANIA database.

Figure 6: Networks of genes with direct co-expression (a) or physical (b) interactions with the query genes or TGCT genes. In both networks, green nodes represent known TGCT genes, blue nodes represent query genes, and yellow nodes are genes with direct interactions with the former two groups based on information from GeneMANIA. The orange edges in this figure show direct relationships between query genes and TGCT genes.

Figure 7: Pathway over-representational analysis. (a) Bar chart of common pathways which query and TGCT genes hit them. Blue and green bars correspond to query genes and to TGCT genes, respectively. Each bar shows the number of genes which are included in the pathway. Label of bars show the obtained p-value based on Eq. 3; (b) The name of each pathway and genes which are included in it.

3.4 Pathway Over-representation Analysis

Pathway analysis was performed using the ORA method, and p-values were separately calculated for each KEGG pathway based on query and TGCT genes¹⁴ [55]. 28 and 32 pathways out of 295 KEGG pathways have common genes with query and TGCT genes, respectively and they are all significantly enriched (p-value < 0.01). Further, 6 KEGG pathways are common between query and TGCT genes¹⁵. Figure 7 summarizes these results.

Aside from “*pathways in cancer*”, which is a general pathway, direct impacts of “*PI3K-AKT signaling pathway*”, “*Proteoglycans in cancer*” and “*Estrogen signaling pathway*” on testicular have been previously reported [56–58]. Given that PI3K-AKT pathway is dysregulated in several cancers and has a cross talk with WNT-CTNNB1 signaling pathway, Boyer *et al.* [56] have shown the involvement and synergy of the PI3K-AKT and WNT-CTNNB1 pathways in the formation of sex cord tumors of the testis. The role of proteoglycans in various cancer processes, such as proliferation and metastasis, has also been observed. Furthermore, it has been observed that proteoglycans such as glypican (-1, -3) may interact with growth factors, cytokines, morphogens and enzymes, also leading to tumor growth and invasion. In [57], expression of glypican-3 (GPC3) has been observed in all samples of neoplastic cells of yolk-sac tumor. According to [58] estrogens, the archetype of female hormones, contribute to control of male germ cell proliferation and hypofertility. The authors have also shown that estrogens contributed to human testicular germ cell cancer proliferation by rapid activation of ERK1/2 and PKA through a membrane nonclassical ER.

Moreover, *actin cytoskeleton* has been shown to have a role in cancer cell migration and invasion [59]. Finally, *AMP-activated protein kinase* (AMPK) in “*AMPK signaling pathway*”, can act as a physiological cellular energy sensor and strongly suppresses cell proliferation in both nonmalignant and tumor cells [60].

¹⁴The list of pathways and their involved genes is available in Supp. 4.

¹⁵The list of pathways and their p-values is available in Supp. 5.

3.5 Towards Finding Effective miRNA-mRNA Interactions

Investigation of existing literature reveals that five of the query genes, namely, *HSP90AB1*, *EEF1A1*, *LDHB*, *FN1*, and *RPS19* have been associated with spermatogenesis, TGCT, yolk sac tumors¹⁶, and seminomas¹⁷. Interestingly, these genes, except for *LDHB*, are involved in *enriched pathways* and/or *common significant GO terms* with TGCT genes. It emphasizes on the power of the performed computational analyses. In the following, we briefly discuss the evidence supporting each of these genes.

- *HSP90AB1* and *EEF1A1* play key roles in regulation of spermatogenesis [61]. Our analysis finds that *HSP90AB1* is involved in three enriched pathways that also include known TGCT genes; it has a common significant GO term with *BCL10* and *ESR1*, has three direct physical interactions with *ESR1*, *STK11*, and *FGFR3*. *EEF1A1*, on the other hand, is included in common significant GO term with *TERT*; it also directly interacts with *ESR1* and *ESR2* in GCN and physical interaction networks.
- *LDHB* is over-expressed in TGCT and *FN1* is differentially expressed in yolk sac tumors [62, 63]. Based on our analysis, *LDHB* has direct interactions with known TGCT genes in GCN. *FN1*, on the other hand, is involved in four enriched pathways and a common significant GO term with known TGCT genes; it also has direct relationships in GCN with TGCT known genes.
- *RPS19* is a ribosomal protein and has been recently found to be associated with sex development [64]. Specifically, overexpression of *RPS19* has been detected in seminoma in comparison with normal testicular samples [65]. Interestingly, there is a difference between *RPS19* and other query genes whose names start with ‘RP’. None of the other “RP” genes are involved in any enriched pathways and all of them have direct physical interaction with estrogen genes, except for *RPS19* which has a common significant GO term, *GO:0017134* with *FGFR3*, one of the TGCT genes. *GO:0017134* concerns ‘fibroblast growth factor (FGF) binding’; FGFs are a family of growth factors, which play important roles in the processes of differentiation and proliferation of different tissues. The relationship between FGFs and testis has already been observed in [66].
- Finally, *EEF2* is involved in an enriched pathway, “AMPK signaling pathway”, with *STK11*. It also has direct interaction with *STK11* and *KIT* in GCN and direct physical interaction with *ESR1*. *EEF2* has been found to be over-expressed as an antigen in various cancers [67]; however, it has not yet been associated with TGCT or related diseases.

No strong evidence was found in support of the other query genes based on a review of the literature or in this study. Thus, while these genes represent good potential for future research on TGCT, we henceforth focus on the genes described above, namely *HSP90AB1*, *EEF1A1*, *LDHB*, *FN1*, *EEF2*, and *RPS19*, as potential targets of miRNAs in TGCT. Figure 8 shows the identified miRNA-mRNA interactions for the above genes; it contains 17 interactions involving 16 different miRNAs.

To further delineate the effect of miRNAs on TGCT, we used the expression data for miRNAs in testis and sperm tissues from the miRmine¹⁸ database. Table 4 shows the expression values of miRNAs in log₂ value of Reads Per Million (RPM); data on both 3p and 5p strands were available for some miRNAs, and have been reported. The data suggest that *hsa-miR-137*, *hsa-miR-559*, and *hsa-miR-802* are not expressed in testis or sperm tissues, and can be excluded from further considerations. The above three miRNAs have been marked with an orange color in the network of TGCT-specific miRNA-mRNA interactions in Figure 8.

A number of identified miRNAs have been previously found to be associated with TGCT. Specifically,

- *hsa-miR-892b* is member of *miR-888* gene family that has restricted expression in adult testicular germ cells. They are known as Cancer-testis (CT) antigens which are clustered near the end of the

¹⁶Yolk sac tumor is a kind of germ cell tumor cancers.

¹⁷ Testicular seminomas are the most common testicular tumors and account for approximately 45% of all primary testicular tumors.

¹⁸<http://guanlab.ccmb.med.umich.edu/mirmine>, released July 2015.

Figure 8: The final inferred miRNA-mRNA interaction network after removing genes not found to be strongly related to TGCT using the functional analyses of Sections 3.2-3.4. Orange nodes show miRNAs with zero expression values in testis and sperm tissues, based on the data from miRmine.

Table 4: Expression of miRNAs in testis and sperm. The unit of values is $\log_2(RPM)$.

miRNAs	Testis	Sperm
hsa-miR-130a-5p	0	0
hsa-miR-130a-3p	9.2	7.1
hsa-miR-137	0	0
hsa-miR-146a-5p	9.8	13.4
hsa-miR-146a-3p	0	0
hsa-miR-148b-5p	4.2	6.1
hsa-miR-148b-3p	10.2	11.2
hsa-miR-152-5p	4.8	2.1
hsa-miR-152-3p	11.6	11.2
hsa-miR-181c-5p	7.3	7.7
hsa-miR-181c-3p	4.7	4.4
hsa-miR-210-5p	1.4	0.7
hsa-miR-210-3p	6.4	8.4
hsa-miR-23a-5p	0	1.1
hsa-miR-23a-3p	12.6	11.3
hsa-miR-23b-5p	3.8	3
hsa-miR-23b-3p	12.7	10.1
hsa-miR-33b-5p	0	3.8
hsa-miR-33b-3p	2	0
hsa-miR-559	0	0
hsa-miR-618	3.5	3.6
hsa-miR-632	1	0
hsa-miR-760	5.8	6
hsa-miR-802	0	0
hsa-miR-892b	5.8	7.3

long arm of the X chromosome (Xq27-Xq28) and contain several reported miRNAs like hsa-miR-888, hsa-miR-890, hsa-miR-891a, hsa-miR-891b, hsa-miR-892a, and hsa-miR-892b [68–70];

- *hsa-miR23a* and *hsa-miR-23b* are involved in epididymal maturation of sperm and express in epididymides [71];
- *hsa-miR-33**—including both *hsa-miR-33a* and *hsa-miR-33b*—has been shown to be differentially expressed in testes [72]; moreover, according to [73] hsa-miR-33a-5p directly targets *EEF1A1*;
- *hsa-miR-148a* and *hsa-miR-148b* have been found to affect the modification of susceptibility to oligozoospermia [74];
- by comparing fertile and infertile populations of men *hsa-miR-152-3p* has been found to be correlated with sperm concentration [75];
- *miR-181b* and *miR-181c* were found up-regulated in adult in mouse testis tissue [76], and are involved in transcriptional regulation in haploid germ cells by targeting *rsbn1*, a gene postulated to be involved in transcriptional regulation in haploid germ cells;
- *miR-155* and *miR-146a* have been found to be present in male serum [77] and to be correlated with each other; however, while miR-155 is associated with male sub-fertility independent of Low-Grade Systemic Inflammation (LGSi) or androgens, miR-146a is only weakly associated with sub-fertility and LGSi;
- by comparing miRNA expression patterns for abnormal semen from infertile males and normal semen from healthy males, *hsa-miR-130a* has been found to be significantly under-expressed in the abnormal semen compared with the normal semen [78].

4 Discussion

miRNAs play key roles in many complex diseases, including various cancers. Identification of miRNAs involved in a disease, such as TGCT, can facilitate the development of new therapies to control and treat the disease by designing *Anti-miRNA Oligonucleotides (AMOs)*. AMOs are synthetically designed molecules to neutralize miRNA function in cells in order to achieve the desired response. In cases where over/under expression of miRNAs leads to disease initiation or progression, AMOs can thus be used to develop therapies by controlling regulation of miRNAs [10–13, 79, 80]. As an example, miR-21 is known as an *oncomir* in several cancers. Si *et al.* [79] and Sicard *et al.* [80] have successfully used *anti-miR-21* to control tumor growth in breast and pancreatic cancer, respectively. Further biological investigations to assess the therapeutic potential of miRNAs identified in this study and designing AMOs to control functional miRNAs can thus lead to future discoveries.

Despite the promise of miRNAs as potential therapeutic targets, the identification of functional disease-specific miRNAs is challenging. The primary approach is to first find genes that are associated with the diseases, and then find miRNAs that target these genes, i.e., *miRNA-mRNA interactions*. Finding such interactions in the laboratory is a time and cost consuming process. Hence, prior to conducting laboratory experiments, it is often preferred to predict and identify more likely functional interactions using computational methods and narrow down the primitive list of possible interactions. Such an approach can reduce the number of possible interactions to a few candidate interactions, resulting in more efficient and amenable experiments. In this study, we employed such a strategy in the setting of TGCT and developed a computational framework to narrow down the potential list of functional miRNA-mRNA interactions in TGCT to a small set of 13 interactions. This list can provide an ideal candidate miRNAs for *in vivo* validation experiments and designing AMOs to control the activity of miRNAs and their target genes.

A potential limitation of our analysis is the use of normal prostate samples for validation of TGCT-specific interactions, due to unavailability of matched miRNA and mRNA expression profiles from normal testis samples. While the similarity of gene expression profiles in normal prostate and testis samples justifies

our analysis, improved results may be obtained if the method is used with expression profiles from cancer and normal testis tissue.

While the current study focused on TGCT, the proposed approach can be similarly applied to other diseases, including cancers. The availability of an automated pipeline can thus facilitate future research on the role of miRNAs in complex diseases.

Few possible extensions can improve the performance and reliability of the proposed framework. First, in the proposed framework, the primary list of interactions were identified using an ensemble learning approach, by applying two existing method of interaction prediction based on expression profiles to filter sequence-based predicted interactions. Ensemble methods can generate more accurate and reliable results than their individual building blocks and are hence preferable [81, 82]. It may thus be fruitful to include additional miRNA-mRNA interaction learning methods in the proposed ensemble learning approach. Moreover, in the proposed framework, candidate interactions are primarily narrowed down by filtering the genes, and information from miRNAs is only incorporated at the last step. As new miRNA expression repositories and knowledge bases emerge, it may be beneficial to incorporate the existing miRNA expression profiles and annotations in earlier steps of the framework.

Acknowledgement

The authors would like to thank the editor, the associate editor and the anonymous referees for helpful comments that led to improvements in the manuscript. Ali Shojaie would also like to acknowledge the support from the U.S. National Science Foundation (grant DMS-1161565).

5 Supplementary material

- Supp. 1: The list of top-ranked 500 interactions using TaLasso and GenMiR++ in TGCT.
- Supp. 2: The list of common interactions among top-ranked 500 interactions between TaLasso and GenMiR++ in TGCT.
- Supp. 3: The list of common interactions among top-ranked 500 interactions between TaLasso and GenMiR++ in prostate.
- Supp. 4: List of pathways and involved genes.
- Supp. 5: List of pathways and p-values.
- Supp. 6: mRNA and miRNA data files as ‘*.RData’.
- Supp. 7: A short note on the required preprocessing steps of level 3.0 TCGA data.
- Code for reproducing results of experiments at GitHub repository, ‘FCS-miRNA-mRNA-Net’¹⁹.

References

- [1] Testicular cancer. *American Cancer Society*, 2015.
- [2] Victor E Reuter. Origins and molecular biology of testicular germ cell tumors. *Modern Pathology*, 18:S51–S60, 2005.

¹⁹<https://github.com/NaSed/FCS-miRNA-mRNA-Net>

- [3] Mary L McMaster, Ketil R Heimdal, Jennifer T Loud, Janet S Bracci, Philip S Rosenberg, and Mark H Greene. Nontesticular cancers in relatives of testicular germ cell tumor (TGCT) patients from multiple-case tgct families. *Cancer medicine*, 2015.
- [4] Elizabeth A Rapley, Gillian P Crockford, Dawn Teare, Patrick Biggs, Sheila Seal, Rita Barfoot, Sandra Edwards, Rifat Hamoudi, Ketil Heimdal, Sophie D Fosså, et al. Localization to Xq27 of a susceptibility gene for testicular germ-cell tumours. *Nature Genetics*, 24(2):197–200, 2000.
- [5] Clare Turnbull, Elizabeth A Rapley, Sheila Seal, David Pernet, Anthony Renwick, Deborah Hughes, Michelle Ricketts, Rachel Linger, Jeremie Nsengimana, Panagiotis Deloukas, et al. Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer. *Nature Genetics*, 42(7):604–607, 2010.
- [6] Leendert HJ Looijenga, Hans Stoop, and Katharina Biermann. Testicular cancer: biology and biomarkers. *Virchows Archiv*, 464(3):301–313, 2014.
- [7] Carlo M Croce and George A Calin. miRNAs, cancer, and stem cell division. *Cell*, 122(1):6–7, 2005.
- [8] Marilena V Iorio, Manuela Ferracin, Chang-Gong Liu, Angelo Veronese, Riccardo Spizzo, Silvia Sab-bioni, Eros Magri, Massimo Pedriali, Muller Fabbri, Manuela Campiglio, et al. MicroRNA gene expres-sion deregulation in human breast cancer. *Cancer Research*, 65(16):7065–7070, 2005.
- [9] George Adrian Calin and Carlo Maria Croce. MicroRNA-cancer connection: the beginning of a new tale. *Cancer Research*, 66(15):7390–7394, 2006.
- [10] Stine B Thorsen, Susanna Obad, Niels F Jensen, Jan Stenvang, and Sakari Kauppinen. The therapeutic potential of microRNAs in cancer. *The Cancer Journal*, 18(3):275–284, 2012.
- [11] Jeppe Vinther, Jakob Lewin Rukov, and Noam Shomron. MicroRNAs and their antagonists as novel therapeutics. In *From Nucleic Acids Sequences to Molecular Medicine*, pages 503–523. Springer, 2012.
- [12] Eva van Rooij and Sakari Kauppinen. Development of microRNA therapeutics is coming of age. *EMBO Molecular Medicine*, 6(7):851–864, 2014.
- [13] Jan Stenvang, Andreas Petri, Morten Lindow, Susanna Obad, and Sakari Kauppinen. Inhibition of microRNA function by antimiR oligonucleotides. *Silence*, 3(1):1, 2012.
- [14] Ramiro Garzon, Guido Marcucci, and Carlo M Croce. Targeting microRNAs in cancer: rationale, strategies and challenges. *Nature Reviews Drug Discovery*, 9(10):775–789, 2010.
- [15] David P Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–233, 2009.
- [16] Anders Jacobsen, Jiayu Wen, Debora S Marks, and Anders Krogh. Signatures of RNA binding proteins globally coupled to effective microRNA target sites. *Genome Research*, 20(8):1010–1019, 2010.
- [17] Ilse Van der Auwera, R Limame, P Van Dam, PB Vermeulen, LY Dirix, and SJ Van Laere. Integrated miRNA and mRNA expression profiling of the inflammatory breast cancer subtype. *British Journal of Cancer*, 103(4):532–541, 2010.
- [18] Huiqing Liu, Angela R Brannon, Anupama R Reddy, Gabriela Alexe, Michael W Seiler, Alexandra Arreola, Jay H Oza, Ming Yao, David Juan, Louis S Liou, et al. Identifying mRNA targets of microRNA dysregulated in cancer: with application to clear cell renal cell carcinoma. *BMC Systems Biology*, 4(1):51, 2010.
- [19] Gabriele Sales, Alessandro Coppe, Andrea Bisognin, Marta Biasiolo, Stefania Bortoluzzi, and Chiara Romualdi. MAGIA, a web-based tool for miRNA and genes integrated analysis. *Nucleic Acids Research*, page gkq423, 2010.
- [20] Jim C Huang, Quaid D Morris, and Brendan J Frey. Bayesian inference of microRNA targets from sequence and expression data. *Journal of Computational Biology*, 14(5):550–563, 2007.

- [21] Ander Muniategui, Rubén Nogales-Cadenas, Miguél Vázquez, Xabier L Aranguren, Xabier Agirre, Aernout Luttun, Felipe Prosper, Alberto Pascual-Montano, and Angel Rubio. Quantification of miRNA-mRNA interactions. *PloS One*, 7(2):e30766–e30766, 2012.
- [22] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- [23] Nafiseh Sedaghat, Takumi Saegusa, Timothy Randolph, and Ali Shojaie. Comparative study of computational methods for reconstructing genetic networks of cancer-related pathways. *Cancer Informatics*, 13(Suppl 2):55, 2014.
- [24] Anders Jacobsen, Joachim Silber, Girish Harinath, Jason T Huse, Nikolaus Schultz, and Chris Sander. Analysis of microRNA-target interactions across diverse cancer types. *Nature structural & molecular biology*, 20(11):1325–1332, 2013.
- [25] Soudeh Ghafouri-Fard, Zahra Ousati Ashtiani, Bareto Sabah Golian, Seyyed-Mohammad Hasheminasab, and Mohammad Hossein Modarressi. Expression of two testis-specific genes, SPATA19 and LEMD1, in prostate cancer. *Archives of Medical Research*, 41(3):195–200, 2010.
- [26] Andrew John Riggan and Mohummad Minhaj Siddiqui. Development of intermediate and high-risk prostate cancer after testicular cancer. *Journal of Clinical Oncology (Meeting Abstracts)*, 33(7):suppl 177, 2015.
- [27] Yasin Aydogmus, Muhammet Fatih Kilinc, Mucahit Kabar, Elif Ozer, and Berkan Resorlu. Metastasis of prostate adenocarcinoma to the testis. *Urology*, 86(1):206, 2015.
- [28] Henrik Møller. Trends in incidence of testicular cancer and prostate cancer in denmark. *Human Reproduction*, 16(5):1007–1011, 2001.
- [29] Senthil R Kumar, Jeffrey Bryan, and Magda Esebua. Epigenetic changes in testis specific Y-like 5 gene in human prostate carcinoma: Gene expression analysis and its potential as a biomarker. *Cancer Research*, 74(19 Supplement):398–398, 2014.
- [30] Tapan K Bera, Rangan Maitra, Carlo Iavarone, Giuliana Salvatore, Vasantha Kumar, James J Vincent, BK Sathyanarayana, Paul Duray, BK Lee, and Ira Pastan. PATE, a gene expressed in prostate cancer, normal prostate, and testis, identified by a functional genomic approach. *Proceedings of the National Academy of Sciences*, 99(5):3058–3063, 2002.
- [31] Tapan Kumar Bera, Yoonsoo Hahn, Byungkook Lee, and Ira H Pastan. TEPP, a new gene specifically expressed in testis, prostate, and placenta and well conserved in chordates. *Biochemical and Biophysical Research Communications*, 312(4):1209–1215, 2003.
- [32] M Margalit, L Yogev, H Yavetz, O Lehavi, R Hauser, A Botchan, S Barda, F Levitin, M Weiss, I Pastan, et al. Involvement of the prostate and testis expression (PATE)-like proteins in sperm-oocyte interaction. *Human Reproduction*, page des064, 2012.
- [33] Yun-Fai Chris Lau and Jianqing Zhang. Expression analysis of thirty one Y chromosome genes in human prostate cancer. *Molecular Carcinogenesis*, 27(4):308–321, 2000.
- [34] Vijay K Dasari, Reza Z Goharderakhshan, Geetha Perinchery, Long-Cheng Li, Yuichiro Tanaka, Judy Alonzo, and Rajvir Dahiya. Expression analysis of Y chromosome genes in human prostate cancer. *The Journal of Urology*, 165(4):1335–1341, 2001.
- [35] Gene Ontology Consortium et al. The gene ontology project in 2008. *Nucleic Acids Research*, 36(suppl 1):D440–D444, 2008.
- [36] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, 40(D1):D940–D946, 2012.

- [37] Guangchuang Yu, Fei Li, Yide Qin, Xiaochen Bo, Yibo Wu, and Shengqi Wang. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7):976–978, 2010.
- [38] Guangchuang Yu, Li-Gen Wang, Guang-Rong Yan, and Qing-Yu He. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, page btu684, 2014.
- [39] Matthew T Weirauch. Gene coexpression networks for the analysis of dna microarray data. *Applied Statistics for Network Biology: Methods in Systems Biology*, pages 215–250, 2011.
- [40] Swarup Roy, Dhruba K Bhattacharyya, and Jugal K Kalita. Reconstruction of gene co-expression network from microarray data using local expression patterns. *BMC Bioinformatics*, 15(Suppl 7):S10, 2014.
- [41] David Warde-Farley, Sylva L Donaldson, Ovi Comes, Khalid Zuberi, Rashad Badrawi, Pauline Chao, Max Franz, Chris Grouios, Farzana Kazi, Christian Tannus Lopes, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38(suppl 2):W214–W220, 2010.
- [42] Jason Montojo, Khalid Zuberi, Harold Rodriguez, Gary D Bader, and Quaid Morris. GeneMANIA: fast gene network construction and function prediction for Cytoscape. *F1000Research*, 3, 2014.
- [43] Sorin Draghici, Purvesh Khatri, Adi Laurentiu Tarca, Kashyap Amin, Arina Done, Calin Voichita, Constantin Georgescu, and Roberto Romero. A systems biology approach for pathway level analysis. *Genome Research*, 17(10):1537–1545, 2007.
- [44] Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Computational Biology*, 8(2):e1002375, 2012.
- [45] Gabriele Sales, Enrica Calura, and Chiara Romualdi. *graphite: GRAPH Interaction from pathway Topological Environment*, 2015. R package version 1.16.0.
- [46] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.
- [47] James Z Wang, Zhidian Du, Rapeeporn Payattakool, S Yu Philip, and Chin-Fu Chen. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23(10):1274–1281, 2007.
- [48] Xin Jiang, Melissa Skibba, Chi Zhang, Yi Tan, Ying Xin, and Yaqin Qu. The roles of fibroblast growth factors in the testicular development and tumor. *Journal of Diabetes Research*, 2013, 2013.
- [49] JB Weinberg, E Doty, J Bonaventura, and AF Haney. Nitric oxide inhibition of human sperm motility. *Fertility and Sterility*, 64(2):408–413, 1995.
- [50] G Santoro, C Romeo, P Impellizzeri, R Ientile, G Cutroneo, F Trimarchi, S Pedale, N Turiaco, and C Gentile. Nitric oxide synthase patterns in normal and varicocele testis in adolescents. *BJU International*, 88(9):967–973, 2001.
- [51] William R Defoor, Chia-Yi Kuan, Malinda Pinkerton, Curtis A Sheldon, and Alfor G Lewis. Modulation of germ cell apoptosis with a nitric oxide synthase inhibitor in a murine model of congenital cryptorchidism. *The Journal of Urology*, 172(4):1731–1735, 2004.
- [52] Graham Bignell, Raffaella Smith, Chris Hunter, Philip Stephens, Helen Davies, Chris Greenman, Jon Teague, Adam Butler, Sarah Edkins, Claire Stevens, et al. Sequence analysis of the protein kinase gene family in human testicular germ-cell tumors of adolescents and adults. *Genes, Chromosomes and Cancer*, 45(1):42–46, 2006.
- [53] AE Samir Bashandy. Effect of fixed oil of *Nigella sativa* on male fertility in normal and hyperlipidemic rats. *International Journal of Pharmacology*, 3(1):27–33, 2007.

- [54] Hameed N Bataineh and Mohamad K Nusier. Effect of cholesterol diet on reproductive function in male albino rats. *Saudi Medical Journal*, 26(3):398–404, 2005.
- [55] Minoru Kanehisa and Susumu Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.
- [56] Alexandre Boyer, Marilène Paquet, Marie-Noëlle Laguë, Louis Hermo, and Derek Boerboom. Dysregulation of WNT/CTNNB1 and PI3K/AKT signaling in testicular stromal cells causes granulosa cell tumor of the testis. *Carcinogenesis*, 30(5):869–878, 2009.
- [57] Satoshi Ota, Michiyo Hishinuma, Naoko Yamauchi, Akiteru Goto, Teppei Morikawa, Tetsuya Fujimura, Tadaichi Kitamura, Tatsuhiko Kodama, Hiroyuki Aburatani, and Masashi Fukayama. Oncofetal protein glypican-3 in testicular germ-cell tumor. *Virchows Archiv*, 449(3):308–314, 2006.
- [58] Adil Bouskine, Marielle Nebout, Baharia Mograbi, Francoise Brucker-Davis, Cyril Roger, and Patrick Fenichel. Estrogens promote human testicular germ cell cancer through a membrane-mediated activation of extracellular regulated kinase and protein kinase A. *Endocrinology*, 149(2):565–573, 2008.
- [59] Hideki Yamaguchi and John Condeelis. Regulation of the actin cytoskeleton in cancer cell migration and invasion. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1773(5):642–652, 2007.
- [60] Hyeon-Jin Kim, Sang-Ki Kim, Byeong-Soo Kim, Seung-Ho Lee, Young-Seok Park, Byung-Kwon Park, So-Jung Kim, Jin Kim, Changsun Choi, Jong-Suk Kim, et al. Apoptotic effect of quercetin on HT-29 colon cancer cells via the AMPK signaling pathway. *Journal of Agricultural and Food Chemistry*, 58(15):8643–8650, 2010.
- [61] Joseph S Tash, Ramappa Chakrasali, Sudhakar R Jakkaraaj, Jennifer Hughes, S Kendall Smith, Kaori Hornbaker, Leslie L Heckert, Sedide B Ozturk, M Kyle Hadden, Terri Goss Kinzy, et al. Gamendazole, an orally active indazole carboxylic acid male contraceptive agent, targets HSP90AB1 (HSP90BETA) and EEF1A1 (eEF1A), and stimulates Il1a transcription in rat sertoli cells. *Biology of Reproduction*, 78(6):1139–1152, 2008.
- [62] S Rodriguez, O Jafer, H Goker, BM Summersgill, G Zafarana, AJM Gillis, RJHLM van Gurp, JW Oosterhuis, YJ Lu, R Huddart, et al. Expression profile of genes from 12p in testicular germ cell tumors of adolescents and adults associated with i (12p) and amplification at 12p11. 2–p12. 1. *Oncogene*, 22(12):1880–1891, 2003.
- [63] Dejan Juric, Sanja Sale, Robert A Hromas, Ron Yu, Yan Wang, George E Duran, Robert Tibshirani, Lawrence H Einhorn, and Branimir I Sikic. Gene expression profiling differentiates germ cell tumors from other cancers and defines subtype-specific signatures. *Proceedings of the National Academy of Sciences of the United States of America*, 102(49):17763–17768, 2005.
- [64] J Guo, P Zhu, C Wu, L Yu, S Zhao, and X Gu. In silico analysis indicates a similar gene expression pattern between human brain and testis. *Cytogenetic and Genome Research*, 103(1-2):58–62, 2003.
- [65] Tanja Pascale Neuvians, Isabella Gashaw, Christian Georg Sauer, Christian von Ostau, Sabine Kliesch, Martin Bergmann, Axel Häcker, and Rainer Grobholz. Standardization strategy for quantitative PCR in human seminoma and normal testis. *Journal of Biotechnology*, 117(2):163–171, 2005.
- [66] SB Winge, J Nielsen, A Jørgensen, S Owczarek, KA Ewen, JE Nielsen, A Juul, V Berezin, and E Rajpert-De Meyts. Biglycan is a novel binding partner of fibroblast growth factor receptor 3c (FGFR3c) in the human testis. *Molecular and Cellular Endocrinology*, 399:235–243, 2015.
- [67] Yusuke Oji, Naoya Tatsumi, Mari Fukuda, Shin-Ichi Nakatsuka, Sayaka Aoyagi, Erika Hirata, Isamu Nanchi, Fumihiko Fujiki, Hiroko Nakajima, Yumiko Yamamoto, et al. The translation elongation factor eEF2 is a novel tumor-associated antigen overexpressed in various types of cancers. *International Journal of Oncology*, 44(5):1461–1469, 2014.

- [68] Adriann M Hovey, Eric J Devor, Patrick J Breheny, Sarah L Mott, Donghai Dai, Kristina W Thiel, and Kimberly K Leslie. miR-888: a novel cancer-testis antigen that targets the progesterone receptor in endometrial cancer. *Translational oncology*, 8(2):85–96, 2015.
- [69] Shannon M Hawkins, Gregory M Buchold, and Martin M Matzuk. Minireview: the roles of small rna pathways in reproductive medicine. *Molecular Endocrinology*, 25(8):1257–1279, 2011.
- [70] Clémence Belleannée, Ezéquiél Calvo, Véronique Thimon, Daniel G Cyr, Christine Légaré, Louis Garneau, and Robert Sullivan. Role of micrnas in controlling gene expression in different segments of the human epididymis. *PLoS One*, 7(4):e34996, 2012.
- [71] Jinsong Zhang, Qiang Liu, Wei Zhang, Jianyuan Li, Zheng Li, Zhongyi Tang, Yixue Li, Chunsheng Han, Susan H Hall, and Yonglian Zhang. Comparative profiling of genes and miRNAs expressed in the newborn, young adult, and aged human epididymides. *Acta biochimica et biophysica Sinica*, 42(2):145–153, 2010.
- [72] Leigh Goedeke, Frances M Vales-Lara, Michael Fenstermaker, Daniel Cirera-Salinas, Aranzazu Chamorro-Jorganes, Cristina M Ramírez, Julie A Mattison, Rafael de Cabo, Yajaira Suárez, and Carlos Fernández-Hernando. A regulatory role for microRNA 33* in controlling lipid metabolism gene expression. *Molecular and cellular biology*, 33(11):2339–2352, 2013.
- [73] Zheng Chen, Jing Ye, Usama Ashraf, Yunchuan Li, Siqi Wei, Shengfeng Wan, Ali Zohaib, Yunfeng Song, Huanchun Chen, and Shengbo Cao. MicroRNA-33a-5p modulates Japanese encephalitis virus replication by targeting eukaryotic translation Elongation Factor 1a1. *Journal of virology*, 90(7):3722–3734, 2016.
- [74] Miaofei Xu, Yufeng Qin, Jianhua Qu, Chuncheng Lu, Ying Wang, Wei Wu, Ling Song, Shoulin Wang, Feng Chen, Hongbing Shen, et al. Evaluation of five candidate genes from GWAS for association with oligozoospermia in a Han Chinese population. *PloS One*, 8(11):e80374, 2013.
- [75] Albert Salas-Huetos, Joan Blanco, Francesca Vidal, Anna Godo, Mark Grossmann, Maria Carme Pons, F Silvia, Nicolás Garrido, Ester Anton, et al. Spermatozoa from patients with seminal alterations exhibit a differential micro-ribonucleic acid profile. *Fertility and sterility*, 104(3):591–601, 2015.
- [76] Naihong Yan, Yilu Lu, Huaqin Sun, Dachang Tao, Sizhong Zhang, Wenying Liu, and Yongxin Ma. A microarray for microRNA profiling in mouse testis tissues. *Reproduction*, 134(1):73–79, 2007.
- [77] Christos Tsatsanis, Johannes Bobjer, Hamideh Rastkhani, Erini Dermitzaki, Marianna Katrinaki, Andrew N Margioris, Yvonne Lundberg Giwerzman, and Aleksander Giwerzman. Serum miR-155 as a potential biomarker of male fertility. *Human Reproduction*, page dev031, 2015.
- [78] Te Liu, Weiwei Cheng, Yongtao Gao, Hui Wang, and Zhixue Liu. Microarray analysis of microRNA expression patterns in the semen of infertile men with semen abnormalities. *Molecular medicine reports*, 6(3):535–542, 2012.
- [79] ML Si, S Zhu, H Wu, Z Lu, F Wu, and YY Mo. miR-21-mediated tumor growth. *Oncogene*, 26(19):2799–2803, 2007.
- [80] Flavie Sicard, Marion Gayral, Hubert Lulka, Louis Buscail, and Pierre Cordelier. Targeting miR-21 for the therapy of pancreatic cancer. *Molecular Therapy*, 21(5):986–994, 2013.
- [81] Giovanni Seni and John F Elder. Ensemble methods in data mining: improving accuracy through combining predictions. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 2(1):1–126, 2010.
- [82] Pengyi Yang, Yee Hwa Yang, Bing B Zhou, and Albert Y Zomaya. A review of ensemble methods in bioinformatics. *Current Bioinformatics*, 5(4):296–308, 2010.