BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 13 2013, pages 1638–1646
doi:10.1093/bioinformatics/btt246

Gene expression

Advance Access publication May 1, 2013

miMsg: a target enrichment algorithm for predicted miR–mRNA interactions based on relative ranking of matched expression data

Martin A. Rijlaarsdam[1], David J. Rijlaarsdam[2,3], Ad J. M. Gillis[1], Lambert C. J. Dorssers[1] and Leendert H. J. Looijenga[1,*]

[1]Department of Pathology, Erasmus MC-University Medical Center Rotterdam, Rotterdam, The Netherlands, [2]Department of Mechanical Engineering, Control Systems Technology, Eindhoven University of Technology, Eindhoven, The Netherlands and [3]Department of Fundamental Electricity and Instrumentation, Free University of Brussels (VUB), Brussels, Belgium

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Algorithms predicting microRNA (miR)–mRNA interactions generate high numbers of possible interactions, many of which might be non-existent or irrelevant in a certain biological context. It is desirable to develop a transparent, user-friendly, unbiased tool to enrich miR–mRNA predictions.

Results: The miMsg algorithm uses matched miR/mRNA expression data to enrich miR–mRNA predictions. It grades interactions by the number, magnitude and significance of misplacements in the combined ranking profiles of miR/mRNA expression assessed over multiple biological samples. miMsg requires minimal user input and makes no statistical assumptions. It identified 921 out of 56 262 interactions as top scoring and significant in an actual germ cell cancer dataset. Twenty-eight miR–mRNA pairs were deemed of highest interest based on ranking by miMsg and supported by current knowledge about validated interactions and biological function. To conclude, miMsg is an effective algorithm to reduce a high number of predicted interactions to a small set of high confidence interactions for further study.

Availability and Implementation: Matlab source code and datasets available at www.martinrijlaarsdam.nl/mimsg

Contact: l.looijenga@erasmusmc.nl (homepage)

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 4, 2012; revised on April 24, 2013; accepted on April 25, 2013

1 INTRODUCTION

MicroRNAs (miRs) are a subclass of a large group of non-coding RNAs that play an important role in post-transcriptional regulation of gene expression, crucially influencing development and physiology (Carthew and Sontheimer, 2009; He and Hannon, 2004; Mallory and Vaucheret, 2006; Pauli *et al.*, 2011; Taft *et al.*, 2010). In addition, miRs also contribute to initiation and progression of diseases, including cancer (Carthew and Sontheimer, 2009; Esquela-Kerscher and Slack, 2006; Krol *et al.*, 2010). Around 2000 miRs have been identified in mammals so far [miRBase, (Griffiths-Jones *et al.*, 2006)], each possibly regulating (protein levels of) hundreds of different mRNAs according to available prediction algorithms (Bartel, 2009). Much research has been done to elucidate the regulation of mRNA levels by miRs. Initial results showed that in plants mRNA degradation occurs by fully complementary binding of miRs to open reading frames (Huntzinger and Izaurralde, 2011). In animals, miRs were hypothesized to predominantly repress translation by binding of partially complementary sequences to their target's mRNAs 3′-untranslated region, hence not leading to mRNA degradation (Carthew and Sontheimer, 2009; Huntzinger and Izaurralde, 2011; Sobkowiak *et al.*, 2008; Voinnet, 2009). However, mRNA degradation as a consequence of miR binding has also been demonstrated in animals (Carthew and Sontheimer, 2009; Eulalio *et al.*, 2008; Filipowicz *et al.*, 2008; Huang *et al.*, 2007b; Huntzinger and Izaurralde, 2011; Lim *et al.*, 2005; Thomas *et al.*, 2010; Wu and Belasco, 2008). In general, induction of miR expression results in reduced levels of their mRNA targets, while depletion of a specific miR will lead higher target levels (Baek *et al.*, 2008; Guo *et al.*, 2010; Hendrickson *et al.*, 2009; Selbach *et al.*, 2008). Inactivation of the miR processing pathway also leads to higher overall levels of mRNA targets (Huntzinger and Izaurralde, 2011; Schmitter *et al.*, 2006). Therefore, the current dogma acknowledges that both translational repression and mRNA degradation are present in plants and animals (Carthew and Sontheimer, 2009; Eulalio *et al.*, 2008; Filipowicz *et al.*, 2008; Huang *et al.*, 2007b; Huntzinger and Izaurralde, 2011; Lim *et al.*, 2005; Rajewsky, 2006; Thomas *et al.*, 2010; Wu and Belasco, 2008). These effects on mRNA levels are also detectable in gene expression data (Huntzinger and Izaurralde, 2011; Lim *et al.*, 2005).

Degradation of mRNAs by miRs can be detected in high-throughput expression data and can be successfully applied to enrich the many possible interactions suggested by sequence-based prediction algorithms (Barbato *et al.*, 2009), e.g. TargetScan (Lewis *et al.*, 2003), Miranda (John *et al.*, 2004) or PicTar (Krek *et al.*, 2005). In this article, a novel method (miMsg) will be presented to enrich predicted miR–mRNA interactions within a specific biological context. The algorithm uses matched high-throughput miR and mRNA expression data of biological samples. The miMsg algorithm is based on relative ranking (RR) profiles of the expression data and does not directly assess the absolute expression levels. Therefore, the algorithm can be applied to detect small effects of miRs on

*To whom correspondence should be addressed.

mRNA, inferring power by detecting patterns over multiple biological samples. In recent literature, various other computational methods have been applied to matched mRNA and miR expression data including regression, correlation analysis and Bayesian learning algorithms [detailed comparison and review of previously published related methods is provided in Supplementary data A and (Barbato *et al*., 2009)]. In comparison, miMsg is independent of the methods used to measure expression levels and the statistical properties of the data. It also only needs a minimal number of user-defined parameters and has built-in quality control. Finally, it yields a result that is directly related to the observed expression patterns and also supplies a visualization tool for individual interactions. miMsg is extensively benchmarked against other methods using 16 different (sub)sets of cancer-related expression data. Benchmarking shows that non-parametric correlation methods are inferior to the other algorithms tested. For the remaining algorithms, results vary strongly depending on the dataset, and overlap is minimal, identifying miMsg as a valuable alternative. As a proof of principle, miMsg is applied to germ cell tumor (GCT) data, investigating mRNA degradation, as most miR have an inhibitory effect on mRNA/protein levels. In this deregulated (cancer) context, miMsg identified validated relevant interactions and miRs/mRNAs.

2 SYSTEM AND METHODS

2.1 Biological samples, expression profiling, data normalization and statistics

The following matched GCT samples were included: five spermatocytic seminomas (SS, type III GCT), three seminomas (SE, type II GCT) and three embryonal carcinomas (EC, type II GCT) (Oosterhuis and Looijenga, 2005). mRNA expression data (Human Genome U133 Plus 2.0 Array) was acquired, pre-processed and normalized as described in (Looijenga *et al*., 2006). Matched miR expression data (multiplex qPCR, Applied Biosystems) were acquired as described in Supplementary Data M. Raw Ct values were normalized to RNU6b (results main text, Supplementary Table S1) or their global average [(Mestdagh *et al*., 2009), Supplementary Table S2, significant overlap with RNU6b-based normalization (30% overlap, $P<0.01$ based on simulation)]. Unless specified otherwise, a *t*-test was used to compare groups (Excel 2010/Matlab 2012a).

2.2 Annotation and target prediction matching

mRNA target predictions were downloaded from TargetScan's Web site (conserved miR families). The most recent annotation files for the Affymetrix GeneChip HG-U133-Plus2 were downloaded from the manufacturer's website. Data were carefully parsed to fit the miMsg format (Supplementary Data B). miR annotations were split and matched on miR family, class and member separately to prevent missed matches. Gene symbols were used to merge Affymetrix probe annotations with TargetScan's predictions. Only predictions for humans were included (speciesID = 9606) and only probes marked as specific for one transcript were selected (suffix '_at'). In the end, 37 956 mRNA probes and 293 miRs were available for analysis. Of these, 5919 mRNA probes and 148 miRs were predicted to interact (inhibition, 56 262 interactions, Table 1).

3 ALGORITHM

miMsg uses miRNA/mRNA expression and publicly available databases with predicted interactions to enrich miR–mRNA interactions within a specific biological context. The miMsg algorithm is based on the dogma that low mRNA target levels as a result of degradation are detectable in real targets of a specific miR and are absent in non-interacting predicted targets. miMsg is preeminently applicable to miR–mRNA interactions because the relevance of miR–mRNA interactions in a biologically coherent set of samples depends on how the samples interrelate (i.e. ranking) rather than individual expression levels. In an ideal interaction profile, high expression of a miR in a tissue should be coupled with low mRNA expression. By observing this over multiple tissues, miMsg creates a combined ranking profile of miR/mRNA expression for each interaction (RR, Fig. 1).

The output of the miMsg algorithm is straightforward and focuses on biological interpretation. It detects interactions with similar or inverse RR depending on the predicted functional relation. The algorithm computes only three scores for each interaction. These are intuitively related to the observed RR and describe the quality and significance of the deviation from the ideal interaction profile. These scores are used to identify relevant and significant patterns, which can be further studied.

3.1 Definitions

As explained in Figure 1A, the miMsg algorithm uses two matched sets of expression data: A and B. CC contains the hypothesized interactions between A and B and their direction (inhibition for miR–mRNA interactions). The RR of the expression levels is presented in RR_j for each interaction j and used to calculate the following quantities: (i) ε_j as an estimate for the number of misplacements in RR_j (Definition 1), (ii) σ_j as a measure for the magnitude of these misplacements (Definition 2) and (iii) λ_j (effect size), which combines the ε_j and σ_j (Definition 3). Also, associated levels of (empirically derived) significance and false discovery rate (FDR) are presented. Next, these quantities are formally defined (also see Supplementary Data B–E).

Let n denote the total number of interactions between the two datasets and m the total number of biological samples per interaction. Furthermore, let $j=1,\ldots,n$ denote the interaction and $i=1,\ldots,m$ the biological sample.

DEFINITION 1: ε_j (RELATIVE NUMBER OF MISPLACEMENTS). *The relative number of misplacements ε_j is defined as follows:*

$$\varepsilon_j = \frac{\psi_j}{(m-1)} \in [0\ 1]$$

where $\psi_j = 0,1,\ldots,m$-1 denotes the number of misplacements in the RR of interaction j involving m samples. A misplacement is defined as a difference between adjacent ranks being unequal to 1, i.e. ordering differs from ideal profile. (Supplementary Data D).

The relative number of misplacements alone is an incomplete measure for how well the profile in RR_j correlates to the ideal profile. The magnitude of the difference between adjacent ranks is important as well. Small differences between adjacent ranks

Table 1. Number of miRs/mRNAs in miMsg results (GCT)[a]

(Sub)set	miRs[b]	Probes (genes)	Unique Interactions based on probes (gene symbol)	Validated[c]
All interactions	148	5919 (2762)	56 262 (23 429)	170
Significant interactions	146	2447 (1521)	6093 (5137)	51
Top scoring (FDR 5%, $\lambda \leq 0.8750$)	145	1891 (1248)	3473 (3176)	39
Top scoring (FDR 1%, $\lambda \leq 0.8000$)	137	682 (560)	921 (896)	10
Top-100 (FDR 0.1%, $\lambda \leq 0.6867$)	56	93 (90)	100 (99)	2

[a]Parameters: $\alpha = 0.05$, $u_{min} \geq 3$, $tol = 10^{-4}$.
[b]miR-29a was measured twice using two independent primers on the qPCR.
[c]Present in miRTarBase as validated interaction.

(small misplacements) are considered less serious than larger ones (Fig. 1D). Therefore, consider the following measure for the magnitude of the misplacements in RR_j.

DEFINITION 2: σ_j (MAGNITUDE OF MISPLACEMENTS). *The relative magnitude of the misplacements σ_j, measured by its relative variance, is defined as follows:*

$$\sigma_j = \frac{\sum_{i=1}^{m-1} \left| RR_{j,i+1} - RR_{j,i} - 1 \right|}{floor\left(\frac{m^2}{2}\right)} \in [0\ 1]$$

where $RR_{j,i}$ denotes the i^{th} biological sample in the j^{th} interaction, and $floor(m^2/2)$ is the maximum obtainable variance in a RR containing m biological samples.

To quantify the combined effect of the number and magnitude of the misplacements, a combined score (effect size) is defined.

DEFINITION 3: λ_j (COMBINED EFFECT SIZE). *The number and magnitude of misplacements (Definition 1, 2) are combined in λ_j using the following equation:*

$$\lambda_j = \varepsilon_j + \sigma_j - (\varepsilon_j \cdot \sigma_j) \in [0\ 1]$$

λ is a monotonically increasing function of ε and σ, which approaches zero if and only if both ε and σ approach zero (ε:[0 1] and σ:[0 1]). It is a measure for the combined effect size of ε and σ and is used to easily rank selected interactions. Low values of λ indicate a better match to the ideal RR (Fig. 1D).

Finally, a statistical test is required to show relevance and statistical significance of the scores.

3.1.1. Significance and relevance The final steps of the algorithm determine (i) which interaction patterns are statistically unlikely to occur by chance and (ii) which values of λ should be considered relevant (= sufficiently low) (Fig. 1C). These independent steps will be discussed below.

3.1.1.1 Significance In the preceding computations, the absolute values of ε, σ and λ that result for each interaction may be coincidental. Hence, a measure of statistical significance is required. As no a priori knowledge about the statistical properties of the data is present, an empirically derived 2D cumulative distribution function (cdf) is generated to assess the probability of coincidental (ε, σ) combinations. This cdf is iteratively approximated using permuted (shuffled) versions of the original expression data. After each iteration, the cdf is evaluated at all possible combinations (ε_k, σ_m) present in the real data, $k = 1,2,\ldots,K$ and $m = 1,2,\ldots,M$. This yields a cdf $\rho(\varepsilon, \sigma)$ that can be compared between iterations at (ε, σ). The process of adding (ε, σ) combinations from permuted data is repeated until the difference between the current and previous iteration is neglectable (i.e. <tol). This difference is defined as the relative root mean square error (rrε):

$$rr\varepsilon_i = \sqrt{\frac{\sum_{k=1}^{K} \sum_{m=1}^{M} (\rho_i(\varepsilon_k, \sigma_m) - \rho_{i-1}(\varepsilon_k, \sigma_m))^2}{\sum_{k=1}^{K} \sum_{m=1}^{M} \rho_i^2(\varepsilon_k, \sigma_m)}}$$

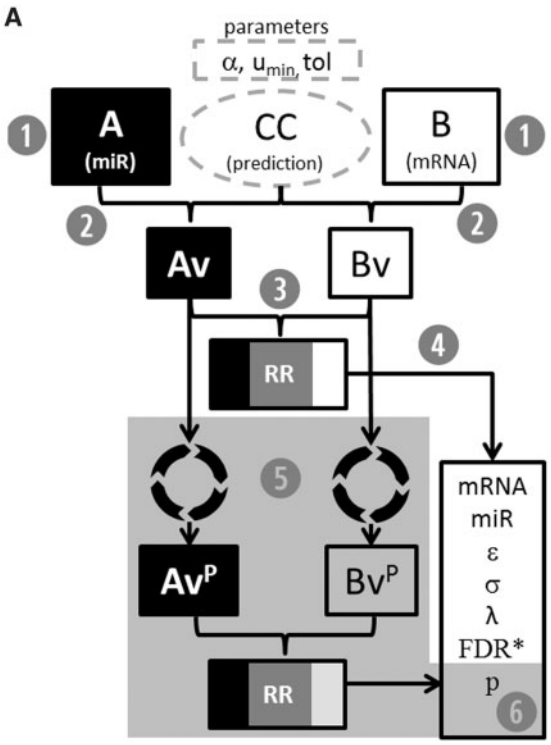
After convergence, the resulting $\rho_{def}(\varepsilon, \sigma)$ is evaluated at all combinations from the real data (ε_k, σ_m) to assess their respective *P*-values (Supplementary Data E). $\rho_{def}(\varepsilon, \sigma)$ is uniquely generated for each group of interactions with the same number of unique biological samples, as this number defines the theoretical variability of ε and σ [i.e. the shape of $\rho_{def}(\varepsilon, \sigma)$]. (Supplementary Data F for more information about repetitive values.) Formally, an interaction is deemed significant if the following holds.

DEFINITION 4: STATISTICAL SIGNIFICANCE. *Consider the j^{th} interaction, the corresponding measures ε_j and σ_j (Definition 1, 2) and a user-defined level of significance α. Moreover, consider the cdf $\rho_{def}(\varepsilon,\sigma)$ obtained using the procedure discussed in Section 3.1.1.1. Then, the j^{th} interaction is defined to be significant if $\rho_{def}(\varepsilon,\sigma) \leq \alpha$.*

3.1.1.2 Relevance Apart from the significance of an interaction pattern, it is required to determine which values of λ should be considered low enough (i.e. relevant). High values of λ might still be statistically significant, but do not constitute relevant interaction patterns. By iteratively using permuted expression data as input for miMsg, a cdf of λ is generated, denoted by $\tau(\lambda)$ (analogous to 3.1.1.1). Using τ, the FDR for all significant values of λ is determined. The researcher can determine a maximal acceptable FDR (β, usually $\leq 1\%/5\%$).

See also Supplementary Data E. Formally, an interaction is deemed relevant if the following holds.

DEFINITION 5: RELEVANCE OF THE SIZE OF λ. *Consider the j^{th} interaction, the corresponding measure λ_j (Definition 3) and a user-defined maximal acceptable FDR β. Moreover, consider the cdf $\tau(\lambda)$ obtained using the procedure discussed in Section 3.1.1.2. Then, the j^{th} interaction is defined to be relevant if $\tau_{def}(\lambda) \leq \beta$.*



Original			Sort according to Bv_j (desc)			RR
Tissue	Av_j	Bv_j	Tissue	Av_j	Bv_j	Rank Av_j
T1	100	400	T5	500	500	5
T2	300	200	T1	100	400	1
T3	400	300	T3	400	300	4
T4	200	100	T2	300	200	3
T5	500	500	T4	200	100	2

3.2 Output

The final output of miMsg is a table with all significant interactions and associated values of ε, σ and λ as well as the associated *P*-values (example: Supplementary Table S1) and FDRs. Also, an overview of the raw expression levels is supplied. A detailed runfile with settings and properties of the input/output is generated as well as a visualization of the quality of the results. An additional tool is supplied to visualize individual significant interactions (example: Fig. 2; Supplementary Data G).

Fig. 1. Flowchart of miMsg. (**A**) (User-defined parameters) α, the desired level of significance (usually $\alpha \leq 0.05$); u_{min}, the minimal number of unique samples required in an interaction to qualify for analysis ($u_{min} \geq 3$); tol, the accepted variation at convergence of the empirically derived cdf of $(\varepsilon,\sigma)/\lambda$ (default: $\leq 10^{-4}$). The following steps are executed: (1) Define two sets of measurements: A and B. Columns represent matched biological samples and rows contain measurements for miRs (A) or mRNAs (B). (2) A and B are linked using table CC containing predicted miR–mRNA interactions. This results in sets Av and Bv. (3) Using Av and Bv, the RR for each interaction is calculated. This process is depicted graphically in Figure 1B. (4) Using RR, the measures ε, σ and λ (Definition 1–3) are calculated, describing the deviation of RR from the ideal interaction profile for each interaction. (5) An empirical 2D cdf for (ε,σ) is computed using iterative permuted versions of Av and Bv (Av^P and Bv^P, gray area). This cdf is then used to derive the significance of the scores for each interaction. (6) λ, the associated level of significance and the FDR (Definition 3–5) are used to identify biologically relevant and significant interactions. (Asterisk)An empirically derived cdf of λ is generated by permuting A and B. Using this cdf, the FDR of λ is calculated for all significant interactions to determine which values of λ are sufficiently low and thus represent relevant interaction profiles. (**B**) Example of the calculation of RR. Five tissues (T1–T5) are analyzed for a single interaction j (hypothesis = inhibition). Values in (A) and (B) are equal to their rank × 100. Av_j/Bv_j are sorted according to Bv_j. The ranks of Avj in this sorted set constitute RR_j and are used to compute ε_j, σ_j and λ_j: $\Psi_j = \{3\text{-}2 = 1;\ 4\text{-}3 = 1;\ 1\text{-}4 \neq 1;\ 5\text{-}1 \neq 1\} = 2$, $\varepsilon_j = 2/(5\text{-}1) = 0.5$; $\sigma_j = (|3\text{-}2\text{-}1| + |4\text{-}3\text{-}1| + |1\text{-}4\text{-}1| + |5\text{-}1\text{-}1|)/\text{ceil}(5^2/2) = 7/13$, $\lambda_j = 0.5 + 7/13\text{-}[0.5*(7/13)] = 0.8$ (Supplementary Data B–G). (**C**) Graphical representation of the selection of interaction patterns that are both significant and relevant (black diamond). ALL = all interactions. Areas are not to scale. (**D**) Examples of two RRs (10 tissues) with larger (dashed black) or smaller (dashed, gray) displacements. Dotted gray line represents the ideal RR. $\varepsilon_{gray} = \varepsilon_{black} = 2/(10\text{-}1) = 0.22$; σ_{gray} $(|8\text{-}5\text{-}1| + |6\text{-}10\text{-}1|)/\text{ceil}(10\hat{}2/2) = 7/50 = 0.14 \rightarrow \lambda_{gray} = 0.3229$; $\sigma_{black} = (|1\text{-}7\text{-}1| + |8\text{-}3\text{-}1|)/\text{ceil}(10\hat{}2/2) = 11/50 = 0.22 \rightarrow \lambda_{black} = 0.3616$. This illustrates definitions 2 and 3: larger displacements are penalized more severely. It also shows that miMsg is tailored to detect linear patterns with multiple 'line segments', only penalizing discontinuities at their boundaries. Non-parametric correlation methods (Spearman and Kendall) identify the whole discontinuity as a disturbance of the overall pattern and repeatedly penalize all ranks involved (Supplementary Data L). Because groups of biological samples can behave differently but still correlate well per group, this leads to underappreciated patterns

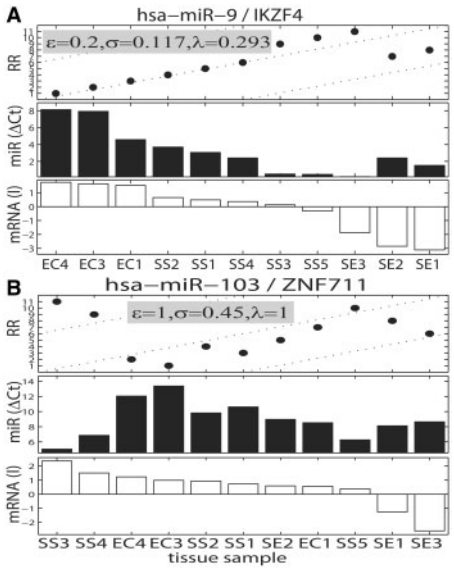


Fig. 2. Visual representation of the RR and miR/mRNA expression levels. (**A**) Top-scoring significant interaction. (**B**) Worst scoring significant interaction. (RR plot) The RR is plotted on the *y*-axis (see also Fig. 1B and D). A straight ascending line is expected when there is a perfect inverse relation between miR and mRNA expression. High values of λ (B) are associated with a noisier RR, not necessarily with inverse ranking patterns. (miR/mRNA expression plots) Bars represent relative mRNA levels (I, mean centered intensity) or miR expression (ΔCt values, high value = low expression). Ideally, identical descending patterns would be observed. *x*-axis: biological samples. Suffix of sample name is arbitrary

4 APPLICATION AND BENCHAMARKING

A selection of histologically diverse GCTs was analyzed to illustrate the applicability and functionality of miMsg. Firstly, the performance of miMsg was studied by assessing the response to permuted or random data. Secondly, a selection of top-scoring and highly significant interactions was further investigated. Finally, extensive benchmarking was done including 16 datasets/subsets and 5 different algorithms for the selection of miR–mRNA interactions.

4.1 GCT data

4.1.1 Response to permuted/random data and noise To assess the robustness of the miMsg algorithm and the relevance of the patterns detected, various tests with random permuted and modified tests were run. When compared with 800 runs with permuted or random expression data, miMsg identified significantly more and better scoring interactions in the real biologically relevant data ($P < 0.01$). This suggests detection of relevant patterns in biological data (Supplementary Data H). Replacement of a fraction of the predicted interactions with random interactions did not prevent miMsg from identifying the originally top-scoring interactions. Moreover, random or targeted reduction of the dataset did not change miMsg's identification and grading of originally selected interactions. This shows that miMsg is relatively insensitive to (introduced) noise and the number of predicted targets (Supplementary Data I).

4.1.2 Detection, selection and verification of interactions High quality of the RRs was identified by both a low *P*-value (significant) and a low value of λ (relevant profile, Fig. 2). A selection of 6093 significant interactions was identified from the original pool of 56 262 predicted interactions (Supplementary Table S1). Top-scoring sets with 1%/5% FDR and a top-100 (FDR 0.1%) were defined based on the accumulated data of 400 runs with permuted expression data (Table 1). These top-scoring and highly significant subsets were matched with miRTarBase to identify experimentally validated interactions. miRTarBase is a literature-based database of 3969 experimentally validated interactions between 625 miRs and 2433 genes in 14 species (version: April 15, 2011) (Hsu *et al.*, 2011). Only human interactions were included (n = 2817). The fraction of validated interactions in all predicted interactions was relatively small as most of the predicted miR–mRNA interactions used as input are not yet experimentally validated. However, enrichment of validated interactions (%) in the selected subset was of interest as a quality measure (rather than the a priori % of validated interactions). Still, enrichment might not be present because interactions validated in non-GCT systems are not necessarily active in GCTs. However, from all algorithms tested in the benchmarking, miMsg showed the highest enrichment for validated interactions in the GCT dataset (37%, Supplementary Data L: Supplementary Fig. S22).

4.2 Benchmarking

miMsg was compared with existing algorithms for miR–mRNA interaction analysis using common non-parametric correlation methods (Spearman and Kendall), a method using Bayesian learning [GenmiR++ (Huang *et al.*, 2007a,b)] and a LASSO-based regression algorithm (Lu *et al.*, 2011) (Supplementary Data A). Twelve dataset/subsets were investigated. Three were obtained from the original LASSO publication: *Madison* (nasopharyngeal cancer), *Broad* (various types and normal tissues, also used in the GenmiR publication) and MSKCC (prostate cancer). The other datasets included parallel miR–mRNA data from different genetic subtypes of multiple myeloma [MM; (Gutierrez *et al.*, 2010)] and our primary *GCT* dataset. Several subsets of the GCT and MM sets were analyzed. To start with, random subsets of interactions were constructed that were comparable with the (LASSO) datasets from (Lu *et al.*, 2011) with regard to the number of (validated) interactions. Also, biologically coherent subsets were created including only validated interactions or only genes differentially expressed between tumor types [analogous to (Lu *et al.*, 2011)].

Algorithms were benchmarked based on several criteria (Table 2, Fig. 3). For all algorithms, highly variable performance and little overlap between results were observed between datasets. Overall, common non-parametric correlation methods (Spearman/Kendall) are clearly outperformed by miMsg, LASSO and GenmiR (also see Fig. 1D). Between miMsg, LASSO and GenmiR there is no obvious best choice. Although we favor miMsg because of the limited number of interactions selected and the unbiased distribution-free approach, application of different algorithms can aid the selection of relevant interactions. (Table 2, Fig. 3, Benchmarking in detail: Supplementary Data L)

Table 2. Results of benchmarking

Criterion	Motivation/result
Fraction of interactions selected, and enrichment for validated interactions	**Motivation:** Selection of a high fraction (= high number) of interactions is not feasible when applying the results to research. In addition, if validated interactions are active in the studied tissues, there should be a higher fraction of validated interactions in the selected interactions as compared with the complete dataset.
	Result: Different algorithms select constant fractions from almost all datasets. LASSO selects ca. 50% of all interactions; the other algorithms select ≈5%. Enrichment for validated interactions varies greatly between datasets with no clear trend toward one optimal algorithm (correlation = worst). In the large datasets (GCT and MM), miMsg outperforms the other algorithms by achieving 37% and 33% enrichment for validated interactions ($\alpha \leq 0.05$, FDR ≤ 0.05).
Performance • Of all/selected interactions against permuted data • Of (selected) validated interactions against permuted interactions	**Motivation:** Unpermuted data in general should perform better than random data. Also, validated interactions should outperform permuted data (ROC analysis).
	Result: The total set of (validated) interactions most likely contains a high number of not interacting miR–mRNA pairs/validated interactions that are not active. These are not expected to perform better than random data. For most datasets, the algorithms therefore have difficulty outperforming permuted data when considering all (validated) interactions. GenmiR performs well on some datasets followed by miMsg; correlation analyses show the worst overall performance. When only the selected interactions are considered LASSO, miMsg and to a lesser extend GenmiR perform much better in assigning higher scores to unpermuted data and validated interactions.
Overlap between selected interactions	**Motivation:** Algorithms using different approaches, but identifying many of the same interactions might be more trustworthy.
	Results: There is limited overlap between the interactions selected by the various algorithms even when only good scoring interactions (top-100 or less) are considered. Spearman and Kendall overlap ca. 100% because of high similarity in methodology.
Influence of dataset size and % of validated interactions	**Motivation:** The size of the dataset and the fraction of validated interactions might influence performance of the algorithms.
	Results: For all algorithms, performance does not consistently improve when using smaller random subsets with a higher fraction of validated interactions.
Influence of pre-selecting of interactions	**Motivation:** Pre-selection of relevant interactions based on (gene) expression/validated interactions might influence performance of the algorithms.
	Results: For all algorithms, performance does not consistently improve when using pre-selected/validated only sets of interactions.

Supplementary data L describes the benchmarking in detail.

5 DISCUSSION

5.1 miMsg: considerations (input and application)

Predicted interactions. A sensible trusted set of predicted interactions is a requirement to generate biological relevant output when using miMsg. Supplementary Data H illustrates the effect of random non-sense interactions. Inverse expression patterns may occur by coincidence in expression data, which is not permuted/randomized. miMsg might still identify top-scoring significant interactions in these experiments. This effect disappears when expression data is permutated/randomized.

Biological samples. miMsg generates optimal results when using biological samples that are clearly related to the studied hypothesis and include multiple comparable as well as clinically/biologically diverse samples in roughly equally sized groups. More samples indicate a lower probability of coincidental RRs. Moreover, rankings become more sensible when the samples differ biologically/clinically (Fig. 2A). However, clustering on biological/histological subtype may occur for many genes/miRs simultaneously, independent of miR/mRNA interaction. Multiple samples per cluster are required to reduce this non-specific effect.

Expression data. Adequate measurement precision of the data is of particular importance, as miMsg relies heavily on ranking. Samples with identical expression levels for certain miRs/mRNAs due to rounding off will be excluded in the analysis of interactions involving these miRs/mRNAs (Supplementary Data F). On the other hand, precision outside of the detection method's sensitivity/thresholds will not contribute to better results in miMsg.

Scope of the study, limitations and future plans. In this article, miMsg is applied to mRNA inhibition by miRs. Although there is strong evidence that miRs influence mRNA levels, the combined effect of mRNA inhibition and translational repression without influencing mRNA levels (Huntzinger and Izaurralde, 2011) will not be detected by miMsg. Moreover, miMsg is applied to cancer-related datasets. Cancer is a deregulated state in which interactions might be detected easily. Theoretically

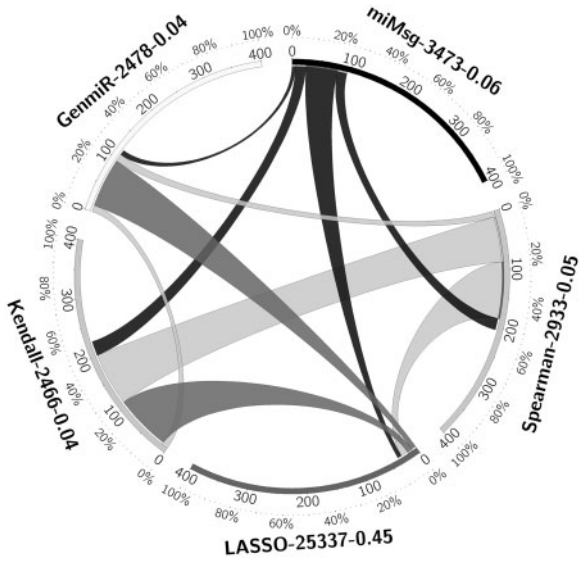


Fig. 3. Overlap between interactions selected by the various algorithms (GCT: 56 262 interactions). Maximum total overlap between five algorithms is (5–1) × 100% = 400% (inner ticks). Outer ticks identify the % of this maximal overlap that is realized for a specific algorithm/dataset. The width of the ribbon identifies the overlap between two algorithms as a % of the number of selected interactions by each algorithm separately. Algorithm labels 'name—total number of selected interactions—% of all interactions selected'. (Example of interpretation) Overall, miMsg shows ≈25% overlap with all other methods. Consider the ribbon from miMsg to LASSO: Of the 3473 interactions selected by miMsg, 54% overlaps with the selection of the LASSO algorithm. On the hand, this overlap with miMsg constitutes only 7.5% of the 25 337 interactions selected by LASSO. Figure generated using Circos (Krzywinski *et al.*, 2009)

miMsg is also applicable to other datasets/directions (non-cancer, other than miR/mRNA interactions, stimulation). miMsg might also be used to create a reference database of miR–mRNA interaction signatures in different tissues/cells. These applications need to be investigated further.

5.2 Preliminary assessment of GCT results

miMsg was applied to expression data from a defined set of human GCTs. They represent the so-called type II (SE and EC) and type III (SS) subtypes. Despite a good response to systemic treatment, metastatic type II GCTs are a major cause of death in young adults. Both SE and EC originate from primordial germ cells/gonocytes. These cells have an intrinsic totipotent capacity and mimic totipotent embryonic stem (ES) cells. Type III GCTs originate from more differentiated germ cells and show no embryonic features (Looijenga, 2009; Oosterhuis and Looijenga, 2005). GCTs and their precursor lesions (either carcinoma *in situ* or gonadoblastoma) can be detected and classified by specific immunohistochemical markers and show distinct patterns/aberrations in DNA copy number and mRNA/miR expression (Gillis et al., 2007; Looijenga et al., 2006; Veltman et al., 2005).

Based on a systematic literature review top-scoring (top-100) and validated (FDR ≤ 5%) miRs/mRNAs associated with GCs/GCTs were shown to be involved in (i) pluripotency/germ line development, (ii) TP53- and TFGβ-mediated apoptosis + therapy sensitivity, (iii) fetal development (iv) germ cell micro-environment, (v) proliferation and (vi) carcinogenesis in general (extended review and references presented in Supplementary Data K).

Moreover, miR-23ab and miR-27ab were highly enriched in the top-100 top-scoring subset (n = 19/100, Fig. 4). These miRs are part of the miR-23 paralog clusters: miR23a/27a/24-2 (cluster A, chromosome 19) and miR23b/27b/24-1 (cluster B, chromosome 9).

The miRs in these clusters are not necessarily regulated identically, but are thought to be closely related based on their evolutionary conservation, sequence homology, genomic clustering and simultaneous expression level changes in human diseases including cancer (Chhabra et al., 2010; Tanzer and Stadler, 2004).

There was significant differential expression of these miR families, clusters and individual miRs between EC, SE and SS ($P < 0.05$). The low expression in SE as compared with EC (both type II GCTs) is interesting because these tumor types differ in prognosis and biological behavior. Moreover, around half of the 19 top-scoring miR-23 paralog targets were shown to be targeted by miRs in human ES cells (Lipchina *et al.*, 2011) (Supplementary Data K).

This overview does not imply proven functional relations, but shows that miMsg can be successfully applied to identify relevant targets for further study and validation in GCT-specific models.

6 CONCLUSION

To conclude, the miMsg algorithm is an unbiased tool to enrich predicted miR–mRNA interactions for a specific biological context using matched high-throughput miR/mRNA expression data (Fig. 1). As interactions are graded on scores directly related to combined ranking profiles of miR/mRNA expression, miMsg is suitable to detect subtle miR–mRNA interactions (Fig. 2). miMsg selected 0.18/1.64/6.17% (FDR 0.1/1/5%) highly relevant and significant interactions in the GCT data (Table 1, Fig. 4). In the largest datasets (GCT/MM: 56 262/27 129 interactions), miMsg outperforms the other algorithms by achieving 37/33% enrichment for validated interactions. Overlap between interactions selected by different algorithms is limited. Overall, the benchmarking results illustrate that there is no obvious best choice between miMsg, LASSO and GenmiR++. Spearman's and Kendall's correlation methods perform worst. Results depend strongly on the dataset used. Although we favor miMsg because of its limited fraction of selected interactions [≈5 versus ≈50% (LASSO)] and unbiased distribution-free approach, application of different algorithms can aid selection of relevant interactions.

A systematic literature review of the GCT results showed association of top-scoring miRs/mRNAs with processes regulating germ cells (proliferation/differentiation) and GCT development. Moreover, interactions involving miR-23 paralogs were highly enriched in the top-100. Fifty percent of the mRNAs involved in these interactions were shown to be miR-binding in ES cells. Based on (i) top-100 ranking by miMsg/FDR ≤ 1%, (ii) involvement of the enriched miR-23 paralog cluster, (iii) known miR–mRNA interaction and/or (iv) experimental validation, 28 high

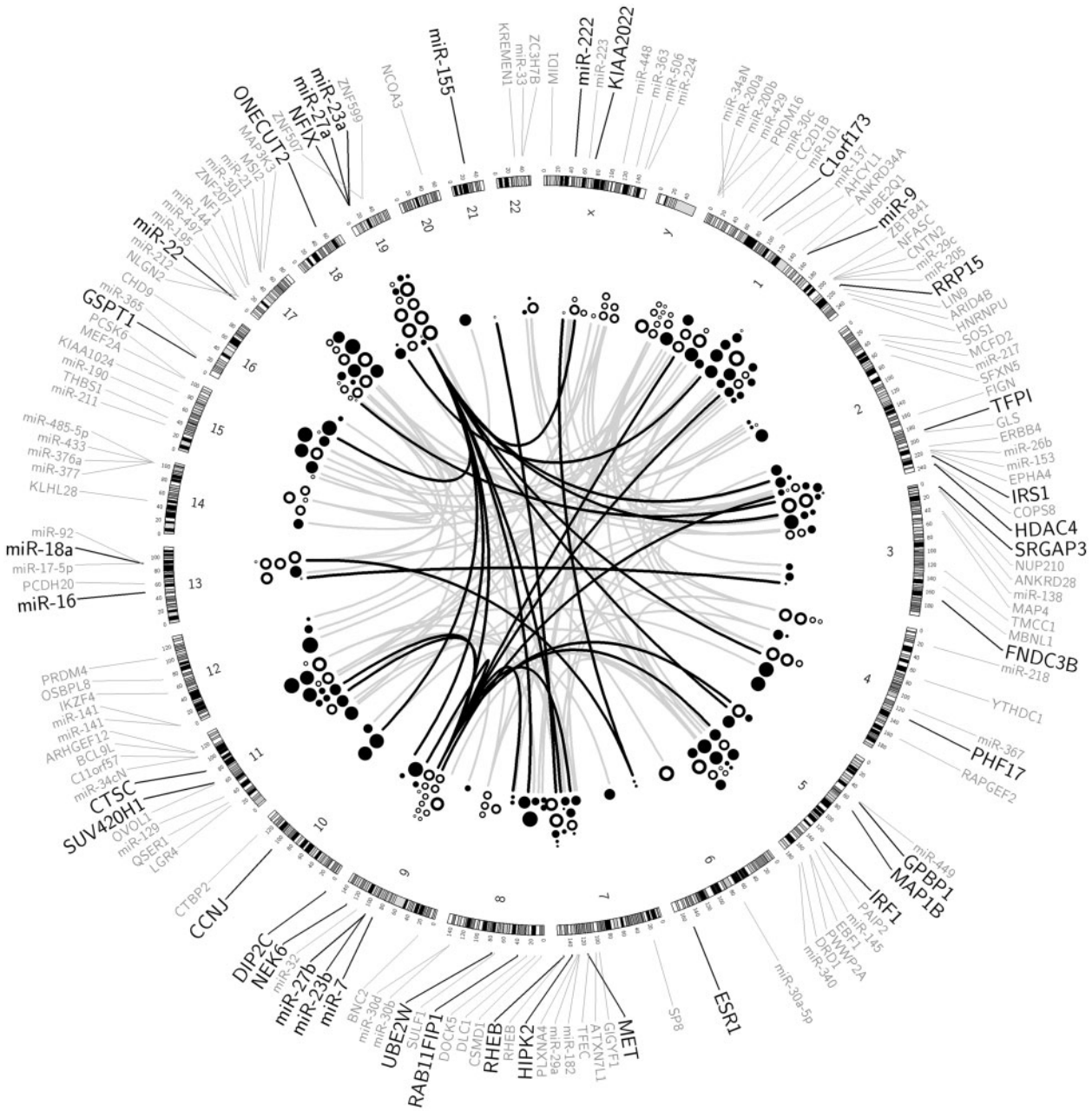


Fig. 4. Visualization of top-scoring significant interactions in the GCT data. Circular representation of the human karyotype, generated using Circos (Krzywinski *et al.*, 2009). Lines represent interactions between miRs and mRNAs. Black text/lines = miR-23ab/27ab cluster (top-100) or validated interactions in top-scoring significant interactions (FDR ≤ 1%). Gray text/lines = other miRs/mRNAs (top-100). Circles represent miRs (open circles) and targets (mRNAs, closed circles). Larger circles indicate better scoring interactions. Depending on the local density of the plot and the available space, the location of the circles cannot always exactly correspond to the genomic position of the miR/mRNA. Based on the known functions of the miRs/mRNAs, four top-scoring interactions were identified as most important (*PHF17*/miR-23a) *PHF17* promotes apoptosis, is a known renal tumor suppressor and associates with OCT3/4 (Pardo *et al.*, 2010), a key protein in germ cell cancer (Looijenga, 2009). Moreover, miR23a has been shown to function as a tumor suppressor-miR via repression by c-Myc. (*IRF1*/miR-23b) *IRF1* is involved in the pathogenesis of infertility in men with germ cell maturation arrest (validated in GCT cell line NT2) (Lian *et al.*, 2010). Moreover, miR-23b is associated with the hormone regulated process of spermiation in Sertoli cells (Nicholls *et al.*, 2011). (*IRS1*/miR-7) *IRS1* is present in peritubular myoid and interstitial cells (Kokk *et al.*, 2005), which might suggest a role in the micro-environment of germ cells. miR-7 interferes with germ line differentiation in Drosophila (Pek *et al.*, 2009). (*HIPK2*/miR-27a) This interaction is validated in ovarian carcinoma cell lines when studying multidrug resistance (Li *et al.*, 2010). miR-27a has also been shown to influence oncogenesis, proliferation and differentiation in various forms of cancer Chhabra *et al.*, 2010). This interaction was present twice in the FDR ≤ 1% top set

confidence interactions were selected. Literature review identified four of these as most relevant for further study (Fig. 4). To conclude, miMsg was highly effective in reducing a high number of predicted miR–mRNA interactions in GCT data, generating a small high-confidence set directly applicable to further research.

ACKNOWLEDGEMENTS

The authors thank the Department of Bioinformatics, Erasmus MC, Rotterdam, for their support. They especially thank Dr Mirjam van den Hout - van Vroonhoven, Ms Sylvia de Does, Mr Ivo Palli and Dr Andreas Kremer. M.R. financially supported by a Translational Grant, Erasmus MC but no specific funding has been received for this project.

Conflict of Interest: none declared.

REFERENCES

Baek,D. *et al.* (2008) The impact of microRNAs on protein output. *Nature*, **455**, 64–71.

Barbato,C. *et al.* (2009) Computational challenges in miRNA target predictions: to be or not to be a true target? *J. Biomed. Biotechnol.*, **2009**, 803069.

Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.

Carthew,R.W. and Sontheimer,E.J. (2009) Origins and Mechanisms of miRNAs and siRNAs. *Cell*, **136**, 642–655.

Chhabra,R. *et al.* (2010) Cooperative and individualistic functions of the microRNAs in the miR-23a~27a~24-2 cluster and its implication in human diseases. *Mol. Cancer*, **9**, 232–248.

Esquela-Kerscher,A. and Slack,F.J. (2006) Oncomirs—microRNAs with a role in cancer. *Nat. Rev. Cancer*, **6**, 259–269.

Eulalio,A. *et al.* (2008) Getting to the root of miRNA-mediated gene silencing. *Cell*, **132**, 9–14.

Filipowicz,W. *et al.* (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev.*, **9**, 102–114.

Gillis,A.J. *et al.* (2007) High-throughput microRNAome analysis in human germ cell tumours. *J. Pathol.*, **213**, 319–328.

Griffiths-Jones,S. *et al.* (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.

Guo,H. *et al.* (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, **466**, 835–840.

Gutierrez,N.C. *et al.* (2010) Deregulation of microRNA expression in the different genetic subtypes of multiple myeloma and correlation with gene expression profiling. *Leukemia*, **24**, 629–637.

He,L. and Hannon,G.J. (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev.*, **5**, 522–531.

Hendrickson,D.G. *et al.* (2009) Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. *PLoS Biol.*, **7**, e1000238.

Hsu,S.D. *et al.* (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, **39**, D163–D169.

Huang,J.C. *et al.* (2007a) Using expression profiling data to identify human microRNA targets. *Nat. Methods*, **4**, 1045–1049.

Huang,J.C. *et al.* (2007b) Bayesian inference of MicroRNA targets from sequence and expression data. *J. Comput. Biol.*, **14**, 550–563.

Huntzinger,E. and Izaurralde,E. (2011) Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat. Rev.*, **12**, 99–110.

John,B. *et al.* (2004) Human MicroRNA targets. *PLoS Biol.*, **2**, e363.

Kokk,K. *et al.* (2005) Expression of insulin receptor substrates 1-3, glucose transporters GLUT-1-4, signal regulatory protein 1alpha, phosphatidylinositol 3-kinase and protein kinase B at the protein level in the human testis. *Anat. Sci. Int.*, **80**, 91–96.

Krek,A. *et al.* (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.

Krol,J. *et al.* (2010) The widespread regulation of microRNA biogenesis, function and decay. *Nat. Rev.*, **11**, 597–610.

Krzywinski,M. *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.

Lewis,B.P. *et al.* (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.

Li,Z. *et al.* (2010) MiR-27a modulates MDR1/P-glycoprotein expression by targeting HIPK2 in human ovarian cancer cells. *Gynecol. Oncol.*, **119**, 125–130.

Lian,J. *et al.* (2010) Downregulation of microRNA-383 is associated with male infertility and promotes testicular embryonal carcinoma cell proliferation by targeting IRF1. *Cell Death Dis.*, **1**, e94.

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

Lipchina,I. *et al.* (2011) Genome-wide identification of microRNA targets in human ES cells reveals a role for miR-302 in modulating BMP response. *Genes Dev.*, **25**, 2173–2186.

Looijenga,L. (2009) Human testicular (non)seminomatous germ cell tumours: the clinical implications of recent pathobiological insights. *J. Pathol.*, **217**, 146–162.

Looijenga,L.H. *et al.* (2006) Genomic and expression profiling of human spermatocytic seminomas: primary spermatocyte as tumorigenic precursor and DMRT1 as candidate chromosome 9 gene. *Cancer Res.*, **66**, 290–302.

Lu,Y. *et al.* (2011) A Lasso regression model for the construction of microRNA-target regulatory networks. *Bioinformatics*, **27**, 2406–2413.

Mallory,A.C. and Vaucheret,H. (2006) Functions of microRNAs and related small RNAs in plants. *Nat. Genet.*, **38** (**Suppl**), S31–S36.

Mestdagh,P. *et al.* (2009) A novel and universal method for microRNA RT-qPCR data normalization. *Genome Biol.*, **10**, R64.

Nicholls,P.K. *et al.* (2011) Hormonal regulation of sertoli cell micro-RNAs at spermiation. *Endocrinology*, **152**, 1670–1683.

Oosterhuis,J.W. and Looijenga,L.H. (2005) Testicular germ-cell tumours in a broader perspective. *Nat. Rev. Cancer*, **5**, 210–222.

Pardo,M. *et al.* (2010) An expanded Oct4 interaction network: implications for stem cell biology, development, and disease. *Cell Stem Cell*, **6**, 382–395.

Pauli,A. *et al.* (2011) Non-coding RNAs as regulators of embryogenesis. *Nat. Rev.*, **12**, 136–149.

Pek,J.W. *et al.* (2009) Drosophila maelstrom ensures proper germline stem cell lineage differentiation by repressing microRNA-7. *Dev. Cell*, **17**, 417–424.

Rajewsky,N. (2006) microRNA target predictions in animals. *Nat. Genet.*, **38** (**Suppl**), S8–S13.

Schmitter,D. *et al.* (2006) Effects of Dicer and Argonaute down-regulation on mRNA levels in human HEK293 cells. *Nucleic Acids Res.*, **34**, 4801–4815.

Selbach,M. *et al.* (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, **455**, 58–63.

Sobkowiak,L., Szarzynska,B. and Szweykowska-Kulinska,Z. (2008) Plant micro RNA biogenesis. *Postepy Biochemii.*, **54**, 308–316.

Taft,R.J. *et al.* (2010) Non-coding RNAs: regulators of disease. *J. Pathol.*, **220**, 126–139.

Tanzer,A. and Stadler,P.F. (2004) Molecular evolution of a microRNA cluster. *J. Mol. Biol.*, **339**, 327–335.

Thomas,M. *et al.* (2010) Desperately seeking microRNA targets. *Nat. Struct. Mol. Biol.*, **17**, 1169–1174.

Veltman,I. *et al.* (2005) Identification of recurrent chromosomal aberrations in germ cell tumors of neonates and infants using genomewide array-based comparative genomic hybridization. *Genes Chromosomes Cancer*, **43**, 367–376.

Voinnet,O. (2009) Origin, biogenesis, and activity of plant microRNAs. *Cell*, **136**, 669–687.

Wu,L. and Belasco,J.G. (2008) Let me count the ways: mechanisms of gene regulation by miRNAs and siRNAs. *Mol. Cell*, **29**, 1–7.