Published online 13 January 2010

Nucleic Acids Research, 2010, Vol. 38, No. 7 ***e97***
doi:10.1093/nar/gkp1239

Impact of probe annotation on the integration of miRNA–mRNA expression profiles for miRNA target detection

Gabriele Sales[1], Alessandro Coppe[1], Silvio Bicciato[2], Stefania Bortoluzzi[1] and Chiara Romualdi[1,*]

[1]Department of Biology, University of Padova, via U. Bassi 58/B, 35121 Padova and [2]Department of Biomedical Sciences, University of Modena and Reggio Emilia, via Campi 287, 41100 Modena, Italy

Received June 19, 2009; Revised December 18, 2009; Accepted December 23, 2009

ABSTRACT

MicroRNAs (miRNAs) are small non-coding RNAs that mediate gene expression at the post-transcriptional and translational levels by an imperfect binding to target mRNA 3′UTR regions. While the *ab-initio* computational prediction of miRNA–mRNA interactions still poses significant challenges, it is possible to overcome some of its limitations by carefully integrating into the analysis the paired expression profiles of miRNAs and mRNAs. In this work, we show how the choice of a proper probe annotation for microarray platforms is an essential requirement to achieve good sensitivity in the identification of miRNA–mRNA interactions. We compare the results obtained from the analysis of the same expression profiles using both gene and transcript based custom CDFs that we have developed for a number of different annotations (ENSEMBL, RefSeq, AceView). In all cases, transcript-based annotations clearly improve the effectiveness of data integration and thus provide a more reliable confirmation of computationally predicted miRNA–mRNA interactions.

INTRODUCTION

MicroRNAs (miRNAs) are a family of small non-coding RNAs, derived from hairpin precursors, abundant in animals, plants and viruses (1–8). miRNAs play central roles in cell differentiation, in the development of tissues and organs, in the pathogenesis of human diseases (9,10) and tumors (11–13). At the molecular level miRNAs influence the stability and translational efficiency of target RNA messengers (mRNAs), mainly by an imperfect binding to their 3′UTR regions (14). More than 800 miRNAs have been identified in human and mouse (15); computational predictions provide even higher figures (16). Recent works estimate that, on average, each miRNA can regulate ~200 target genes (17–19), suggesting that a wide proportion of mammalian genes and biological processes respond to miRNA control mechanisms.

The computational prediction of miRNA targets is extremely challenging due to the lack of a sufficiently large group of experimentally validated targets to be used as a robust training set, and of high-throughput experimental methods for validating results (16). Tools like miRanda, TargetScan, PicTar, PITA and RNAhybrid (19–25), though based on different algorithms and philosophies, all suffer from the limited understanding of the molecular basis involving miRNA-target pairing, that probably, in turn, leads to a reduction of their predictions specificity (26,27). The integration of *in-silico* predictions with other genomic data may overcome the limits of computational predictors and facilitate the identification of functional interactions. In particular, the combination of target predictions with paired miRNA–mRNA expression profiles has been proposed as an efficient way to refine results obtained from methods based on sequences alone.

Although miRNAs may stabilize transcriptional regulation through complex feed-forward and feed-back loops (28), integrative approaches postulate that miRNAs down-regulate mRNAs and that the expression profiles of genuinely interacting miRNA–mRNA pairs are anti-correlated. The standard integrative approaches comprise three steps: (i) prediction of miRNA targets through sequence-based algorithms, (ii) quantification of target expression levels and (iii) assessment of the anti-correlation among miRNAs and their predicted targets. The anti-correlation can be quantified through a variational Bayesian model (29,30) or by computing a correlation coefficient among miRNA and mRNA expression signals (31–33).

*To whom correspondence should be addressed. Tel: +39 049 827 7401; Fax: +39 049 827 6159; Email: chiara.romualdi@unipd.it

Given that miRNA interactions depend on specific sequences in the 3′UTR regions of their targets and that alternative transcripts of a same gene may differ in such UTRs, integrative analyses of expression profiles must take into account the entire length of a transcript. This has been clearly shown by Legendre and colleagues (34) who studied 3′UTRs containing multiple EST-supported poly(A) sites, and looking for known miRNA targets and other phylogenetically conserved motifs, highlighted that motif-containing and motif-free isoforms were differentially represented in specific tissues. In addition, other studies demonstrated that the same miRNA target prediction algorithm produces significantly different results when applied to genes/transcripts defined by distinct annotations: for example, Rajewsky *et al.* (25) reported a 20% variability in the predicted regulatory relationships moving from RefSeq transcripts to UCSC 'known genes'. Target identification, moreover, was affected by alternative adenylation and multiple polyA sites in terminal exons (25,35).

The choice of a transcript-based (TB) approach influences the analysis right from the quantification of target expression. It is well known that a considerable fraction of microarray probes can be (i) entirely mis-assigned (not associated to any gene/transcript in a recent genome annotation), (ii) non-gene-specific (i.e. matching multiple genes) or (iii) non-transcript-specific (matching multiple alternative transcripts of a gene). Several groups have explored the effects of using alternative microarray annotations to quantify gene expression (36,37) and proposed the adoption of custom Chip Definition Files (CDFs) (38–41). The importance of the annotation increases when we consider the integration of miRNA and mRNA expression data because of the role played by alternative 3′UTRs. Unfortunately, the computational procedures developed so far seem to overlook this aspect and adopt gene based CDFs to correlate miRNA-target profiles. The matter is further complicated by the ambiguous definition of a gene 3′UTR region, which may be taken as the union of all 3′UTRs of its transcripts or as the longest one (42).

In this work we investigate how different microarray probe annotations affect the integrative analysis of miRNA-mRNA expression. The analysis has been performed through a computational pipeline that (i) re-annotates microarray probes into GB (gene based) and TB custom CDFs; (ii) predicts miRNA targets starting from transcripts and miRNA sequences; (iii) integrates miRNA target predictions with paired miRNA–mRNA expression signals. In particular, we explore the degree of specificity of miRNA seed pairing to alternative 3′UTR splicing variants and then compare the miRNA–mRNA expression correlation obtained from GB and TB probe annotations. The entire procedure has been tested on paired expression data originally collected to investigate the role of perineural invasion pathway (PNI) in prostate cancer (43). Results clearly show that microarray probe annotations have a substantial impact on the integrative analysis and that TB annotations outperform their GB counterparts.

MATERIALS AND METHODS

We have developed a computational pipeline (Figure 1) to compare the efficiency of GB and TB annotations in the integrative analysis of miRNA–mRNA data. Such pipeline is composed of three major steps: (i) re-annotation of microarray probes to design GB and TB custom CDFs; (ii) prediction of miRNA targets using the sequences of transcripts and miRNAs; (iii) integration of miRNA target predictions with paired miRNA–mRNA expression signals.

In the first step we used the sequences of Affymetrix microarray probes and those of transcripts and genes derived from several annotations (ENSEMBL, RefSeq, AceView) to build custom CDFs. We then obtained miRNA sequences from the mirBase database and used the miRanda, PITA and PicTar algorithms to predict their targets (both at the gene and at the transcript level). In the last step we evaluated gene and transcript expression profiles, and we integrated each with the corresponding miRNA expression signals to refine the predicted miRNA–mRNA interactions.

Transcript sequences and annotations

Transcript sequences and annotations were obtained from three databases, i.e. ENSEMBL (version 52), RefSeq (version 33) and AceView (UCSC hg18). Some RefSeq transcripts were associated to multiple UTRs of different extension; to remove redundancy, we defined a single 3′UTR as the region going from the first base after the end of the coding sequence to the first annotated polyA site.

Construction of custom CDFs

GB and TB custom CDFs have been built for a number of human Affymetrix arrays (i.e. HG95v2, HG133A 2.0, HG133plus2, and Human Exon 1.0 ST) using the ENSEMBL, RefSeq and AceView annotations. Specifically, the custom CDFs were generated (i) matching gene/transcript sequences with all the probes of the microarray, (ii) filtering out all non-specific probes, i.e. those matching more than one gene/transcript, (iii) grouping probes into meta-probe sets with at least four entries, and finally, (iv) discarding all those probes not belonging to any meta-probe set. Details on the number of specific probes and of recognized genes and transcripts are reported in Supplementary Table S1.

Prediction of miRNA targets

Human miRNA sequences were obtained from the miRBase::Sequences repository (version 12). We updated target predictions using three different algorithms, characterized by different target identification strategies (miRanda, PITA and PicTar). We ran miRanda and PITA over the ENSEMBL, RefSeq and AceView annotations; PicTar target predictions, based on RefSeq sequences, were downloaded as provided by PicTar developers since the software is not freely available.

The thresholds for miRanda and PITA were defined applying the two algorithms to artificial sequences

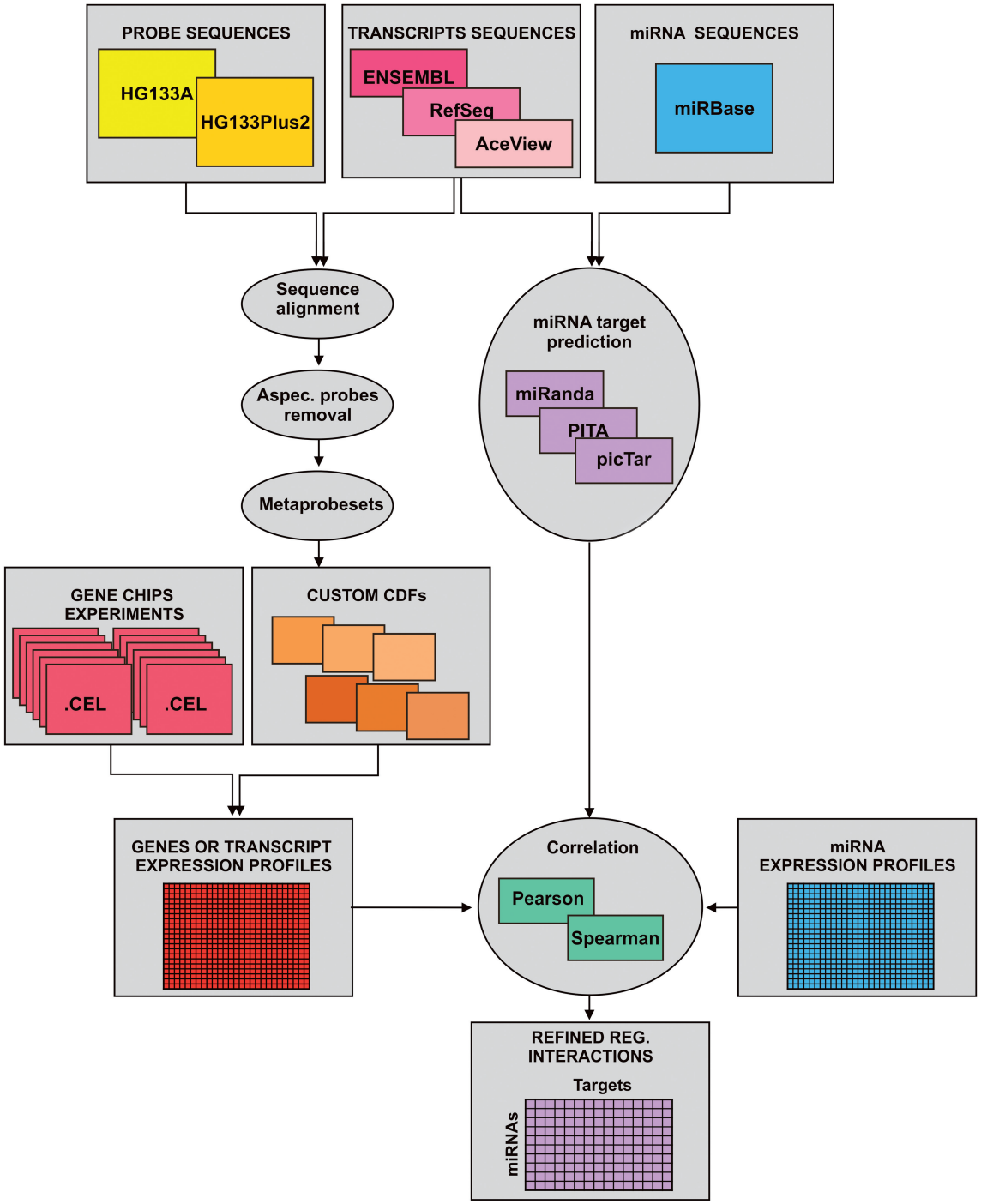

Figure 1. Computational pipeline. Detailed scheme of the pipeline implemented for the construction of our custom CDFs and for the miRNA–mRNA expression profiles integration.

generated through a permutational approach, i.e. shuffling 3′UTR sequences. miRanda scores were all smaller than 200 when applied to shuffled data, while PITA did not recognize any target at all. As a result, the threshold score of miRanda was set to 200 while targets predicted by PITA were further limited to those with the top 10% scores.

Integrative analysis of mRNA and miRNA expression data

We obtained from the GEO database matched mRNA and miRNA expression data of 57 prostate cancer samples [GSE7055 (43)] generated using Affymetrix HGU133A 2.0 microarrays and OSU-CCC MicroRNA custom arrays, respectively.

The Robust Multichip Average model with quantile normalization and HG133A 2.0 custom-CDFs were used to generate and normalize mRNA expression signals. miRNA expression levels were pre-processed using the approach adopted in the original publication (43). Briefly, spots having signal/background ratio below a specific threshold (calculated as the average of blank spots) were filtered out; each experiment was normalized dividing the expression values by their corresponding median level; replicate signals within the array were averaged. This procedure resulted in a matrix containing the expression levels of 426 miRNAs, 236 of which were human-specific.

To evaluate the impact of GB and TB annotations on the integration of miRNA–mRNA expression data, we used only those probe sets of the HG133A 2.0 GB-CDF that measured genes having at least one transcript represented by a probe set in the HG133A 2.0 TB-CDF. In addition, we filtered out genes having a single transcript (the choice of the type of annotation does not affect the quantification of their expression signal). The filtering procedure resulted in 1715 genes and 1818 transcripts using ENSEMBL, 621 genes and 746 transcripts with RefSeq and 12 184 genes and 4599 transcripts considering the AceView annotation. These genes and transcripts were then used as targets to predict miRNA-target interactions with miRanda, PITA and PicTar (the latter is limited to RefSeq sequences only; see Supplementary Table S2).

We calculated the correlation among all miRNA and mRNA expression profiles using both parametric (Pearson) and non-parametric (Spearman) coefficients. To quantify the impact of different annotations we defined the delta correlations (Δc) as the differences between the correlation levels of a miRNA-gene pair and all of the corresponding miRNA-transcript pairs. The significance of the Δc was assessed comparing the observed Δc with Δc^*, the distribution of Δc calculated by randomly permuting 100 times the mRNA expression levels. Specifically, since the maximum absolute value of Δc^* resulted equal to 0.2, a $|\Delta c| > 0.2$ was considered as an indication of a significant impact of the type of probe annotation (GB or TB) on the correlation of miRNA–mRNA expression data.

Functional enrichment

We calculated the functional enrichment of target genes obtained through TB and GB approaches using the hypergeometric distribution (Fisher exact test) and the GSEA (44,45). The hypergeometric test was performed in DAVID (46) with KEGG (47) and Biocarta pathways (EASE score less than 0.1), while we used the Java application of the Broad Institute (http://www.broadinstitute.org/gsea/) for the GSEA.

RESULTS AND DISCUSSIONS

Analysis of 3′UTR alternative transcripts

The transcript annotation databases, ENSEMBL (v.52), AceView (UCSC hg18) and RefSeq (v.33), provided a total of 54 617, 260 113 and 46 417 transcripts (respectively), resulting in 39 680, 210 003 and 33 518 sequences with annotated 3′UTR regions. The distribution of the number of transcripts *per* gene according to ENSEMBL, RefSeq and AceView showed that a significant fraction of genes (30% for ENSEMBL, 20% for RefSeq and 29% for AceView) have at least two alternative transcripts (Figure 2A and Table 1). Moreover, as different transcripts may share the same 3′UTR, transcripts with the same 3′UTR have been considered as putative targets of the same set of miRNAs. We define a *3′UTR equivalence class* as a set of transcripts of a gene sharing exactly the same 3′UTR sequence. A significant fraction of genes (71% for ENSEMBL, 36% for RefSeq and 94% for AceView) has two or more equivalence classes (Figure 2B and Table 2); such variability in the proportions may be ascribed to differences in terms of annotations of the human genome (48). For instance, predicted alternative transcripts with different 3′UTRs are more numerous and longer in ENSEMBL than in RefSeq. On the contrary AceView, that has been developed to provide a strictly cDNA-supported view of the human transcriptome and to summarize all quality-filtered cDNA data from GenBank, dbEST and RefSeq, is characterized by a larger number of alternative transcripts within the same gene, most of which have different 3′UTR sequences (Figure 2B).

We used the miRanda, PITA and PicTar algorithms to evaluate the specificity of miRNA target predictions with respect to 3′UTR equivalence classes. We computed for each putative miRNA-gene pair the percentage of equivalence classes recognized by the miRNA. Figure 2C and D and Supplementary Figure S2A shows the distribution of the average percentage of 3′UTR equivalence classes *per* miRNA over all its putative target genes using miRanda, PITA and PicTar, respectively. These findings indicate that the heterogeneity of alternative 3′UTRs results in miRNAs highly specific in their targeting 3′UTR equivalence classes. Indeed, while using ENSEMBL and RefSeq approximately half of all 3′UTR equivalence classes of a protein-coding gene are recognized by a specific miRNA; with AceView this quantity drops to <20%, indicating a greater miRNA specificity.

Considering that 26% of genes have more than one transcript (taking the average over the three annotation databases), GB data integration could be deceptive for a significant proportion of protein-coding genes. Indeed, 71% of ENSEMBL genes with more than one transcript exhibit more than one 3′UTR equivalence class; a GB data integration would have been ambiguous for at least 18% of them (9 and 23% for RefSeq and AceView, whose proportions of genes having more than one 3′UTR equivalence class are 36 and 94%, respectively). As an example, the BAIAP2 gene (brain-specific angiogenesis inhibitor-1, ENSG00000175866, a secretin receptor family member whose expression is induced by p53) is associated to three different transcripts which differ in their 3′ region (3′CDS and 3′UTR) and encode different isoforms of an insulin receptor tyrosine kinase substrate of the secretin receptor family (49). Figure 3 shows the alternative transcripts of BAIAP2 that are characterized by the same 5′ region and their regulating miRNAs. Among the 95 miRNAs regulating BAIAP2, only 7 (7%) are shared by

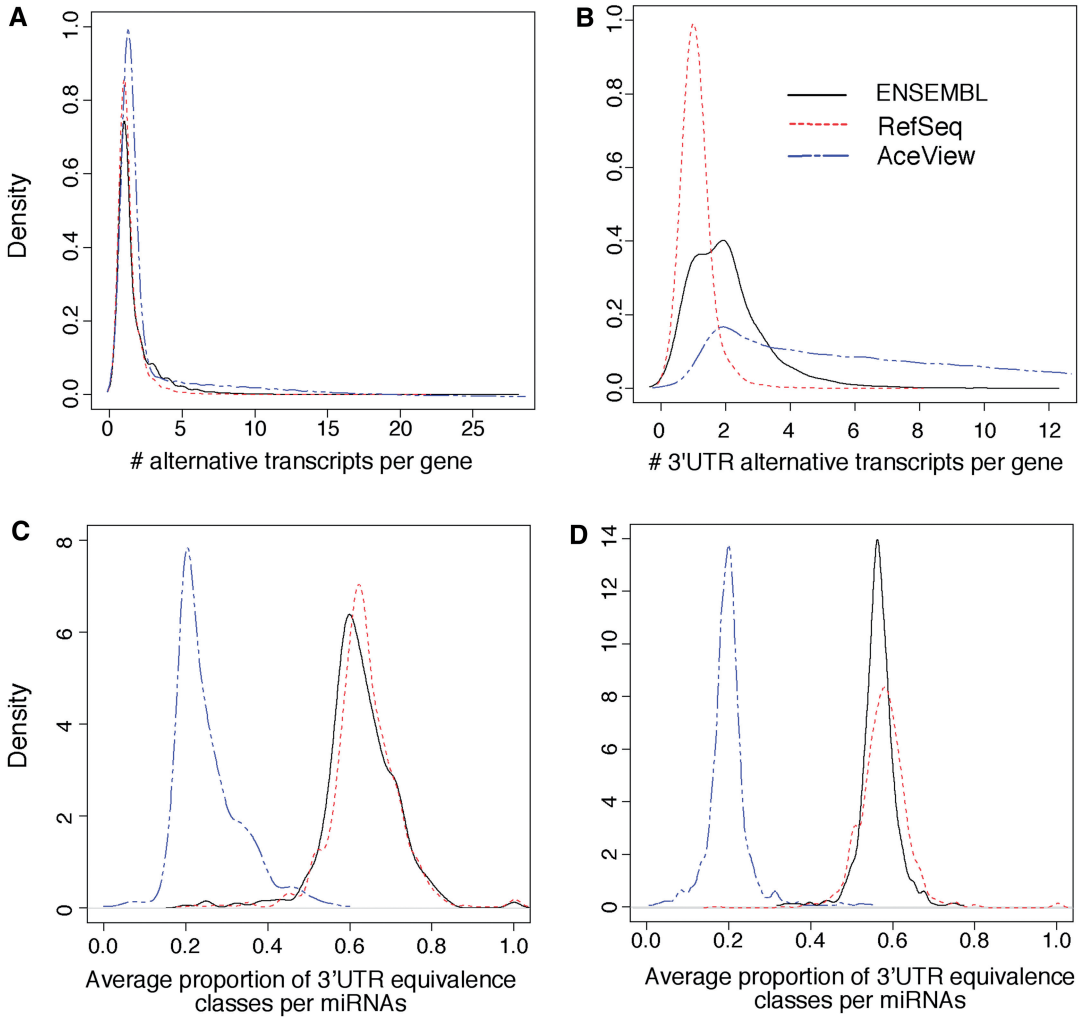

Figure 2. The 3′UTRs transcript variability. (**A**) Distribution of the number of transcripts *per* gene. (**B**) Distribution of the number of different 3′UTRs *per* gene (hereafter called 3′UTR equivalence classes) with at least two alternative transcripts. (**C**) Distribution of the average number of 3′UTR equivalence classes *per* gene targeted by miRNAs using ENSEMBL, RefSeq and AceView databases with miRanda target prediction algorithm. (**D**) Distribution of the average number of 3′UTR equivalence classes *per* gene targeted by miRNAs using ENSEMBL, RefSeq and AceView databases with the PITA target prediction algorithm.

Table 1. Distribution and cumulative distribution of the number of genes with transcript variants according to ENSEMBL (v 52), RefSeq (v.33) and AceView (UCSC hg18)

Number of transcript variants within a gene	ENSEMBL		RefSeq		AceView	
	Frequency (%)	Cumulative frequency (%)	Frequency (%)	Cumulative frequency (%)	Frequency (%)	Cumulative frequency (%)
1	69.9	69.9	80.4	80.4	70.8	70.8
2	13.7	83.6	13.3	93.7	5.0	75.8
3	7.3	90.9	3.5	97.2	3.1	78.9
4	3.9	94.8	1.4	98.6	2.7	81.6
5	2.0	96.8	0.5	99.1	2.3	83.9
>5	3.2	100	0.9	100	16.1	100

all transcripts, while 45% of them (9% for ENST00000321300, 36% for ENST00000321280 and none for ENST00000321238) are transcript-specific. This evidence supports the hypothesis that using GB instead of TB annotation for miRNA–mRNA data integration could lead to misleading results.

Construction of custom CDFs

Several groups have explored the effect of using alternative microarray annotations to improve the estimation of expression values. For instance, Dai and colleagues periodically update several custom CDFs for various

Table 2. Distribution and cumulative distribution of the number of 3′UTR equivalence classes per gene with at least two alternative 3′UTRs according to ENSEMBL (v 52), RefSeq (v.33) and AceView (UCSC hg18)

Number of 3′UTRs within a gene	ENSEMBL		RefSeq		AceView	
	Frequency (%)	Cumulative frequency (%)	Frequency (%)	Cumulative frequency (%)	Frequency (%)	Cumulative frequency (%)
1	29.0	29.0	64.0	64.0	6.0	6.0
2	45.7	74.7	30.8	94.8	17.7	23.7
3	15.7	90.4	4.2	99.0	10.8	34.5
4	5.7	96.1	0.9	99.9	9.5	44.0
5	2.5	98.6	0.02	99.92	8.2	52.2
>5	1.4	100	0.08	100	47.8	100

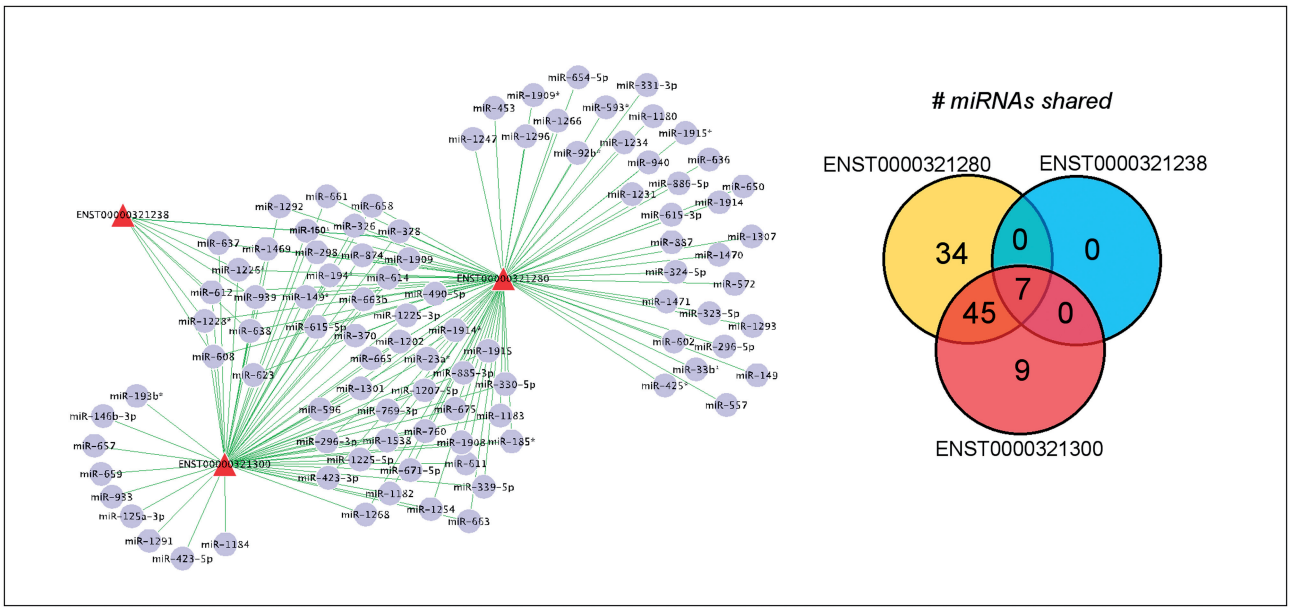

Figure 3. Differences between gene and transcript-based annotation approaches. miRNAs regulating the three alternative transcripts of the gene BAIAP2. Triangles represent transcripts and circles miRNAs. The Venn diagram highlights the fact that only 7% of the 95 miRNAs regulating BAIAP2 are shared by all transcripts, while 45% of them (9% for ENST00000321300, 36% for ENST00000321280 and none for ENST00000321238) are transcript-specific.

Affymetrix platforms (38). In their annotation pipeline, however, probes of a given meta-probe set may match different transcripts. As such transcripts may have different expression profiles, the use of non-specific probes in the process of signal quantification could bias the expression value estimates by increasing expression variability. To overcome this limitation, we have developed an alternative annotation scheme and, as suggested by Moll *et al.* (41), eliminated all non-specific probes. In particular, we reconstructed TB custom CDFs for the most commonly used Affymetrix arrays using ENSEMBL, RefSeq and AceView sequences. Table 3 and Supplementary Table S1 shows the details, in terms of number of genes, probes and transcripts contained in the custom CDF for the HGU133A 2.0 platform based on the three different annotation databases.

Integrative analysis of paired miRNA–mRNA expression profiles

We have used the computational pipeline of Figure 1 for the analysis of paired miRNA–mRNA expression data from 57 prostate cancer samples (43) in order to evaluate the impact of TB annotations on the identification of miRNA targets. The comparative evaluation of GB and TB approaches focused only on those genes having at least two alternative transcripts, i.e. on those cases where the TB annotation should improve data integration. In particular, we evaluated the distributions of differences between correlation estimates (Δc), i.e. the impact of the annotation adopted for expression signal quantification (GB or TB), as a function of the algorithm used for the prediction of miRNA targets (miRanda, PITA or PicTar) and of the type of correlation (parametric or non-parametric coefficients). Figure 4A–C shows the distribution of Δc using the miRanda algorithm for the prediction of miRNA targets and similar results are reported in Supplementary Figures S1A and S2B when using PITA and PicTar, respectively. Δc distributions are centered on zero for all annotation databases and for all type of correlations, but are interestingly characterized by strong kurtosis levels (fat distribution tails), suggesting the

Table 3. Number of unique probes covering genes and transcripts for the construction of our custom CDFs for the Affymetrix platform HG133A 2.0

Platform ID	ENSEMBL	RefSeq	AceView
Number of genes	12 136	12 011	12 184
Number of unique probes covering genes	186 624	185 315	193 901
Number of transcripts	6583	8842	4599
Number of unique probes covering transcripts	86 756	124 420	51 227

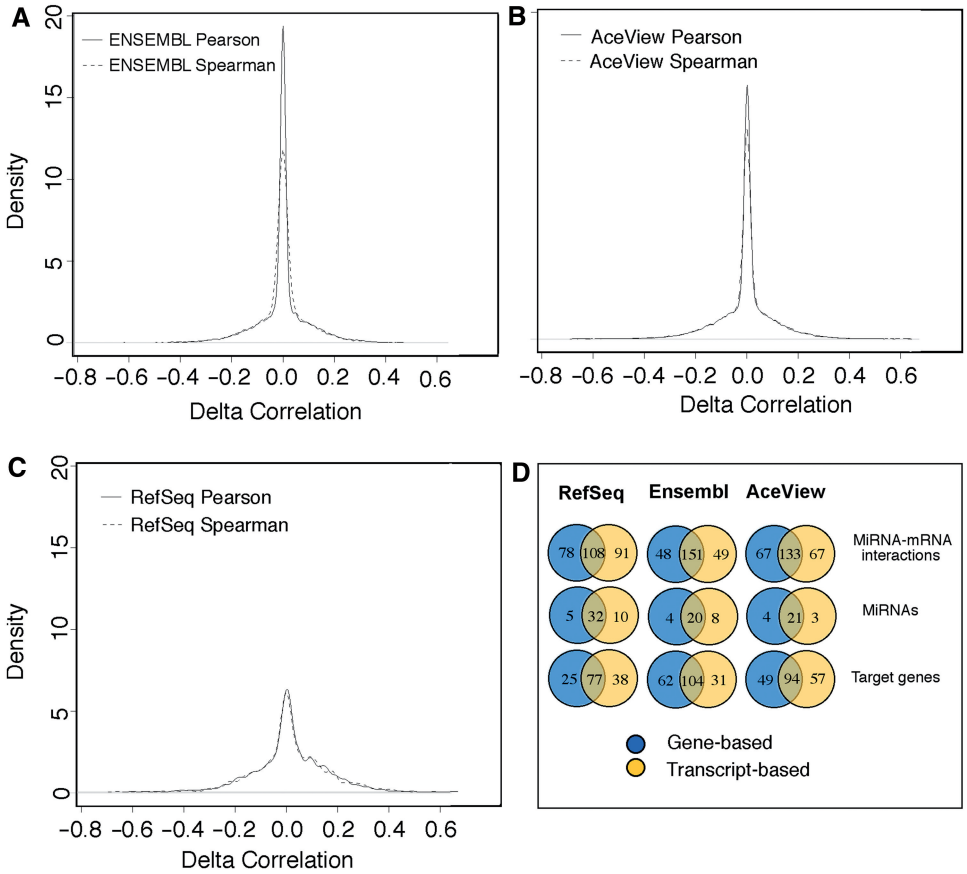

Figure 4. miRNA–mRNA correlations. Differences in parametric and non-parametric correlation estimates obtained using transcript and gene-based annotations, respectively, using ENSEMBL (**A**), AceView (**B**) RefSeq (**C**) sequences and miRanda algorithm. (**D**) Number of miRNA–mRNA interactions, miRNAs and target genes involved in the putative 1% most relevant interactions detected using gene- and transcript-based annotations.

presence of feed-forward and -back transcriptional regulation (28). The significance threshold for Δc has been assessed through a permutational approach and set equal to 0.2 (see 'Materials and methods' section for details and Supplementary Figure S3 for the distributions of delta correlations of real and randomly permuted data). This threshold allowed us to select those genes/transcripts whose correlation coefficient with at least one miRNA is affected by the choice of the annotation (GB or TB). Specifically, 7% of ENSEMBL and AceView gene/transcripts and 14% of RefSeq ones resulted in delta correlations exceeding the threshold of 0.20 ($|\Delta c| > 0.2$ at a FDR < 0.1, Supplementary Figure S4). Among this remaining fraction of interactions, we further considered only those miRNA–mRNA pairs having the top 1% anti-correlation coefficients. As expected, the adoption of GB or TB annotations severely affects the number of miRNA–mRNA interactions, as well as the number of relevant miRNAs and target genes involved in putative interactions (Figure 4D), irrespective of the considered database. Although the aim of Legendre *et al.* study (34) was different from our goal (they do not perform integrative analysis with microarray data and do not evaluate the impact that chip annotation has on correlation calculation), their findings on few specific miRNAs were concordant with our results. For instance, among the 248 most significant anti-correlated mRNA–miRNA pairs identified using miRanda and ENSEMBL, only 151 (60%) are shared between GB and TB lists. Forty-eight miRNA–mRNA pairs, on the other hand, are specific to the GB and 49 to the TB annotations, respectively. Differences in the top 200 anti-correlated interactions identified through

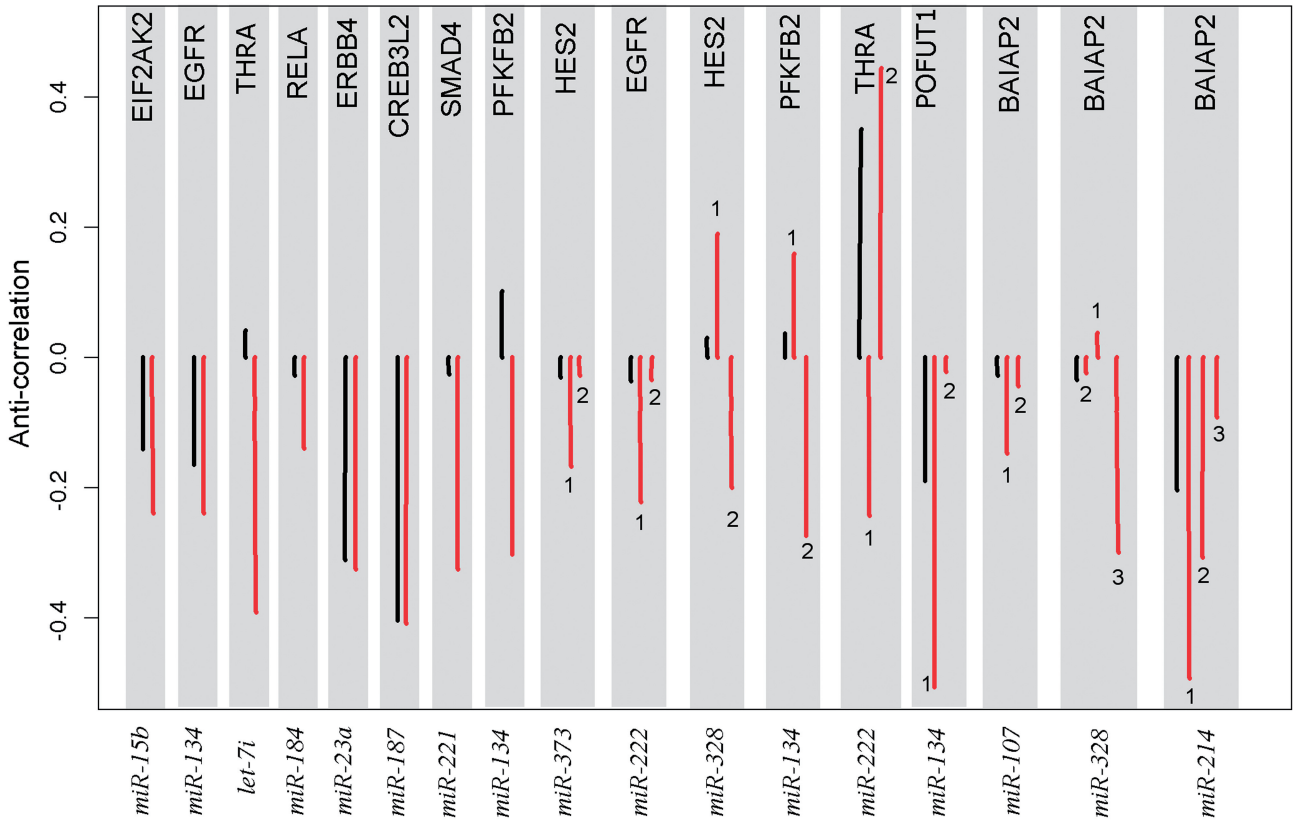

Figure 5. Differences between gene and transcript-based annotation. Selected cases of correlations obtained between mRNA and miRNA with a gene-based (black bars) and a transcript-based (red bars) annotation (only for genes with more than one 3′UTR alternative transcript). The symbol of the gene is reported at the top of the panel, the name of the miRNA targeting at least one transcript is reported in the *x*-axis and the numbers 1, 2 and 3 on the red bars represent the alternative transcripts of the same gene.

the PITA and miRanda are reported in Supplementary Figure S1 (panel B).

Figure 5 shows some examples of miRNA–mRNA interactions together with their GB and TB delta correlations; this highlights the bias introduced by the GB annotation. As an example, all three transcripts of the BAIAP2 are putative targets of *miR-328*, but the anti-correlation of expression signals is not significant (e.g. −0.06) using the GB approach. Using TB annotations, on the other hand, ENST00000321300 and *miR-328* show a significant negative correlation (e.g. −0.3), whereas expression data indicate no correlation between ENST00000321280 and ENST00000321280 and *miR-328* (−0.02 and 0.03, respectively).

Enrichment analyses

An optimal approach should identify, among the supported miRNA–mRNA interactions, a significant proportion of targets and miRNAs with a known role in the pathological processes under examination. As such, we verified the functional enrichment of the most significant anti-correlated mRNA-miRNA pairs among those predicted by miRanda on the ENSEMBL database and confirmed using GB and TB annotations (Figure 4). Specifically, we performed gene set enrichment analysis on GB and TB lists of targets and a literature search on the identified miRNAs. The list of targets obtained through the TB annotation led to highly enriched metabolic pathways using both the hypergeometric and the GSEA approaches (Table 4 and Supplementary Table S3). Both enrichment statistics identified several oncogenes involved in pancreatic and prostate cancer pathways, like CDC42 (associated to human *miR-214*), EGFR (associated to human *miR-134*), RELA (associated to *miR-205*), SMAD4 (associated to *miR-200c*), CREB3L2 (associated to *miR-187*), ERBB4 (associated to *miR-31*) and several other genes involved in the ErbB and GnRH signaling pathways. ERBB2, ERBB4 and EGFR have been implicated in the development of many types of human cancer (50), while the gonadotropin-releasing hormone (GnRH) receptor activation has been demonstrated to inhibit cell proliferation *in vitro* and *in vivo* in prostate cancer (51–53) and GnRH agonists have been used as therapeutical treatment in prostate cancer clinical trials since the early 1980s (54). Interestingly, only some of these interesting pathways were found enriched using the list of targets obtained by the GB annotation.

Among the miRNAs shared by both approaches, 20 out of 32 are highly involved in prostate carcinogenesis [such as *miR-221, miR-222* (55,56) and *miR-145* (57,58)], in bladder cancer, [*miR-23a*, *miR-23b* and *miR-205* (59)] and in testis cancer [*miR-373* (60)]. Among the miRNAs identified only through the TB annotation,

Table 4. Enriched metabolic pathways of the GB and TB significantly anti-correlated miRNA–mRNA interactions (using ENSEMBL and miRanda)

TB		GB	
Pathway	*P*-value	Pathway	*P*-value
Pancreatic cancer	0.016	Wnt signaling pathway	0.05
Adherens junction	0.016	GnRH signaling pathway	0.1
Dorso-ventral axis formation	0.018	beta-Alanine metabolism	0.15
ErbB signaling pathway	0.024	MAPK signaling pathway	0.18
Neuroregulin receptor degredation protein-1 Controls ErbB3 receptor recycling	0.06		
Regulation of actin cytoskeleton	0.07		
Wnt signaling pathway	0.09		
Adipocytokine signaling pathway	0.1		
Prostate cancer	0.1		
GnRH signaling pathway	0.1		
Focal adhesion	0.1		

29% (8 out of 28) are still cancer related, e.g. *miR-106a* and *miR-106b* are known to be involved in prostate cancer (58,61), *miR-223* is involved in bladder cancer (59), *miR-200* in hepatocellular carcinoma (62), *miR-15b* in chronic lymphocytic leukemia (63,64) and *miR-17* in lung cancer and lymphomas (65). Finally, among the miRNAs identified using the GB annotation, only miRNA (*miR-184*) has been reported to be involved in prostate cancer development (58,66). These results become even more intriguing if considering that correlation and enrichment analyses have been performed on a subset of all possible transcripts, i.e. those belonging to genes with more than one alternative 3′UTRs. This suggests that the use of a GB annotation results in a significant loss of information about post-transcriptional regulation, thus impairing the effectiveness of integrative analyses in the identification of real miRNA targets.

CONCLUSIONS

The ENCODE consortium recently completed the characterization of 1% of the human genome showing a striking picture of its complex molecular activity. While the human genome sequencing revealed a number of protein-coding genes lower than previously estimated (<21 000, according to ENSEMBL), ENCODE identified an extensive transcriptional activity of the genome and highlighted the complexity of the RNA transcriptome (67). At the same time, the miRNA revolution in cell-biology and functional genetics has deeply changed the scenario of gene expression regulation, assigning an increasing importance to post-transcriptional mechanisms in development, physiology and disease. Thus, in the light of these new insights, the definition of gene should be somehow revised (67). Recently Gerstein *et al.* (67) proposed an alternative definition of gene as the 'union of genomic sequences encoding a coherent set of potentially overlapping functional products'. This definition pays particular attention to 5′ and 3′ UTRs whose key roles in translation, regulation, stability and localization of mRNAs is widely accepted. Neglecting UTRs from the definition of a gene, one can avoid the problem of multiple 5′ and 3′ ends. Most of the longer protein-coding transcripts identified by ENCODE, differ only in their UTRs (67), thus reinforcing the Gerstein's new suggested definition of gene. This is particularly important when studying post-transcriptional regulation, where the 3′UTRs is the key region for a miRNA–mRNA seed pairing.

Integrative approaches that aim at improving miRNA target identification through the integration of miRNA and mRNA expression profiles seem to underestimate this problem. GB annotation (which ignores the issue of alternative transcripts) is usually adopted to quantify mRNA expression and to calculate miRNA–mRNA expression correlation. Here, we evaluated the impact of using a GB annotation approach rather than a more appropriate TB one. Using prostate cancer as a case study, we demonstrated how TB array annotation shows more consistent results with the pathological state investigated, even when limiting the analysis to genes with multiple alternative transcripts. We identified a considerable number of miRNA–mRNA interactions whose GB anti-correlations show strong biases due to the presence of alternative 3′UTR transcripts with highly different expression profiles. Furthermore, the TB approach was able to predict new putative miRNA–mRNA interactions involving known oncogenes such as EGFR, RELA and ERBB4 whose regulators, i.e. *miR-134*, *miR-205* and *miR-31*, respectively, could represent valid candidates for further experimental validations. Unfortunately, the use of a TB annotation lead to loss of information in terms of filtered non-specific probes, thus reducing the possibility of an exhaustive exploration of the transcriptome regulation. In this perspective, alternative technologies such as new generation deep-sequencers, Affymetrix Exon Arrays or custom arrays, such as Combimatrix, Agilent, Nimblegen (68), would provide a wider coverage.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charges: Fondazione Cassa di Risparmio di Padova e Rovigo (Progetti Eccellenza 2006, 'A computational approach to the study of skeletal muscle genomic expression in health and disease'), University of Padova (CPDR070805 and CPDA07591), MIUR (PRIN 2007Y84HTJ), University of Modena (Finanziamento Linee Strategiche di Sviluppo dell'Ateneo, Medicina Molecolare e Rigenerativa, 2008) and Fondazione Cassa di Risparmio di Modena (Bando ricerca, 2007).

Conflict of interest statement. None declared.

REFERENCES

1. Ambros,V. (2003) MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. *Cell*, **113**, 673–676.
2. Brennecke,J., Hipfner,D.R., Stark,A., Russell,R.B. and Cohen,S.M. (2003) bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell*, **113**, 25–36.
3. Lagos-Quintana,M., Rauhut,R., Lendeckel,W. and Tuschl,T. (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.
4. Lee,R.C., Feinbaum,R.L. and Ambros,V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.
5. Lim,L.P., Glasner,M.E., Yekta,S., Burge,C.B. and Bartel,D.P. (2003) Vertebrate microRNA genes. *Science*, **299**, 1540.
6. Palatnik,J.F., Allen,E., Wu,X., Schommer,C., Schwab,R., Carrington,J.C. and Weigel,D. (2003) Control of leaf morphogenesis by microRNAs. *Nature*, **425**, 257–263.
7. Pfeffer,S., Zavolan,M., Grasser,F.A., Chien,M., Russo,J.J., Ju,J., John,B., Enright,A.J., Marks,D., Sander,C. *et al.* (2004) Identification of virus-encoded microRNAs. *Science*, **304**, 734–736.
8. Reinhart,B.J., Slack,F.J., Basson,M., Pasquinelli,A.E., Bettinger,J.C., Rougvie,A.E., Horvitz,H.R. and Ruvkun,G. (2000) The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*, **403**, 901–906.
9. Stefani,G. and Slack,F.J. (2008) Small non-coding RNAs in animal development. *Nat. Rev. Mol. Cell Biol.*, **9**, 219–230.
10. Zhang,C. (2008) MicroRNomics: a newly emerging approach for disease biology. *Physiol. Genomics*, **33**, 139–147.
11. Blenkiron,C., Goldstein,L.D., Thorne,N.P., Spiteri,I., Chin,S.F., Dunning,M.J., Barbosa-Morais,N.L., Teschendorff,A.E., Green,A.R., Ellis,I.O. *et al.* (2007) MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biol.*, **8**, R214.
12. Hobert,O. (2007) miRNAs play a tune. *Cell*, **131**, 22–24.
13. Lu,J., Getz,G., Miska,E.A., Alvarez-Saavedra,E., Lamb,J., Peck,D., Sweet-Cordero,A., Ebert,B.L., Mak,R.H., Ferrando,A.A. *et al.* (2005) MicroRNA expression profiles classify human cancers. *Nature*, **435**, 834–838.
14. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
15. Griffiths-Jones,S. (2004) The microRNA Registry. *Nucleic Acids Res*, **32**, D109–D111.
16. Bentwich,I. (2005) Prediction and validation of microRNAs and their targets. *FEBS Lett.*, **579**, 5904–5910.
17. Gennarino,V.A., Sardiello,M., Avellino,R., Meola,N., Maselli,V., Anand,S., Cutillo,L., Ballabio,A. and Banfi,S. (2009) MicroRNA target prediction by expression analysis of host genes. *Genome Res.*, **19**, 481–90.
18. Griffiths-Jones,S., Saini,H.K., van Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
19. Krek,A., Grun,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C., Stoffel,M. *et al.* (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.
20. John,B., Enright,A.J., Aravin,A., Tuschl,T., Sander,C. and Marks,D.S. (2004) Human MicroRNA targets. *PLoS Biol.*, **2**, e363.
21. Kertesz,M., Iovino,N., Unnerstall,U., Gaul,U. and Segal,E. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.
22. Kruger,J. and Rehmsmeier,M. (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.*, **34**, W451–W454.
23. Kuhn,D.E., Martin,M.M., Feldman,D.S., Terry,A.V. Jr, Nuovo,G.J. and Elton,T.S. (2008) Experimental validation of miRNA targets. *Methods*, **44**, 47–54.
24. Lewis,B.P., Shih,I.H., Jones-Rhoades,M.W., Bartel,D.P. and Burge,C.B. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.
25. Rajewsky,N. (2006) microRNA target predictions in animals. *Nat. Genet.*, **38(Suppl.)**, S8–S13.
26. Didiano,D. and Hobert,O. (2008) Molecular architecture of a miRNA-regulated 3′ UTR. *Rna*, **14**, 1297–1317.
27. Grimson,A., Farh,K.K., Johnston,W.K., Garrett-Engele,P., Lim,L.P. and Bartel,D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.
28. Hobert,O. (2008) Gene regulation by transcription factors and microRNAs. *Science*, **319**, 1785–1786.
29. Huang,J.C., Babak,T., Corson,T.W., Chua,G., Khan,S., Gallie,B.L., Hughes,T.R., Blencowe,B.J., Frey,B.J. and Morris,Q.D. (2007) Using expression profiling data to identify human microRNA targets. *Nat. Methods*, **4**, 1045–1049.
30. Huang,J.C., Morris,Q.D. and Frey,B.J. (2007) Bayesian inference of MicroRNA targets from sequence and expression data. *J. Comput. Biol.*, **14**, 550–563.
31. Xin,F., Li,M., Balch,C., Thomson,M., Fan,M., Liu,Y., Hammond,S.M., Kim,S. and Nephew,K.P. (2009) Computational analysis of microRNA profiles and their target genes suggests significant involvement in breast cancer antiestrogen resistance. *Bioinformatics*, **25**, 430–434.
32. Wang,Y.P. and Li,K.B. (2009) Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. *BMC Genomics*, **10**, 218.
33. Ruike,Y., Ichimura,A., Tsuchiya,S., Shimizu,K., Kunimoto,R., Okuno,Y. and Tsujimoto,G. (2008) Global correlation analysis for micro-RNA and mRNA expression profiles in human cell lines. *J. Hum. Genet.*, **53**, 515–523.
34. Legendre,M., Ritchie,W., Lopez,F. and Gautheret,D. (2006) Differential repression of alternative transcripts: a screen for miRNA targets. *PLoS Comput. Biol.*, **2**, e43.
35. Thierry-Mieg,D. and Thierry-Mieg,J. (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.*, **7(Suppl. 1)**, S12 11–14.
36. Draghici,S., Khatri,P., Eklund,A.C. and Szallasi,Z. (2006) Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet.*, **22**, 101–109.
37. Mecham,B.H., Klus,G.T., Strovel,J., Augustus,M., Byrne,D., Bozso,P., Wetmore,D.Z., Mariani,T.J., Kohane,I.S. and Szallasi,Z. (2004) Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res.*, **32**, e74.
38. Dai,M., Wang,P., Boyd,A.D., Kostov,G., Athey,B., Jones,E.G., Bunney,W.E., Myers,R.M., Speed,T.P., Akil,H. *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.*, **33**, e175.
39. Ferrari,F., Bortoluzzi,S., Coppe,A., Sirota,A., Safran,M., Shmoish,M., Ferrari,S., Lancet,D., Danieli,G.A. and Bicciato,S. (2007) Novel definition files for human GeneChips based on GeneAnnot. *BMC Bioinformatics*, **8**, 446.
40. Lu,J., Lee,J.C., Salit,M.L. and Cam,M.C. (2007) Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: high-resolution annotation for microarrays. *BMC Bioinformatics*, **8**, 108.

41. Moll,A.G., Lindenmeyer,M.T., Kretzler,M., Nelson,P.J., Zimmer,R. and Cohen,C.D. (2009) Transcript-specific expression profiles derived from sequence-based analysis of standard microarrays. *PLoS ONE*, **4**, e4702.
42. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
43. Prueitt,R.L., Yi,M., Hudson,R.S., Wallace,T.A., Howe,T.M., Yfantis,H.G., Lee,D.H., Stephens,R.M., Liu,C.G., Calin,G.A. *et al.* (2008) Expression of microRNAs and protein-coding genes associated with perineural invasion in prostate cancer. *Prostate*, **68**, 1152–1164.
44. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
45. Mootha,V.K., Lindgren,C.M., Eriksson,K.F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstrale,M., Laurila,E. *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.
46. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
47. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
48. Larsson,T.P., Murray,C.G., Hill,T., Fredriksson,R. and Schioth,H.B. (2005) Comparison of the current RefSeq, Ensembl and EST databases for counting genes and gene discovery. *FEBS Lett.*, **579**, 690–698.
49. Okamura-Oho,Y., Miyashita,T. and Yamada,M. (2001) Distinctive tissue distribution and phosphorylation of IRSp53 isoforms. *Biochem. Biophys. Res. Commun.*, **289**, 957–960.
50. Holbro,T. and Hynes,N.E. (2004) ErbB receptors: directing key signaling networks throughout life. *Annu. Rev. Pharmacol. Toxicol.*, **44**, 195–217.
51. Bahk,J.Y., Hyun,J.S., Lee,H., Kim,M.O., Cho,G.J., Lee,B.H. and Choi,W.S. (1998) Expression of gonadotropin-releasing hormone (GnRH) and GnRH receptor mRNA in prostate cancer cells and effect of GnRH on the proliferation of prostate cancer cells. *Urol. Res.*, **26**, 259–264.
52. Dondi,D., Limonta,P., Moretti,R.M., Marelli,M.M., Garattini,E. and Motta,M. (1994) Antiproliferative effects of luteinizing hormone-releasing hormone (LHRH) agonists on human androgen-independent prostate cancer cell line DU 145: evidence for an autocrine-inhibitory LHRH loop. *Cancer Res.*, **54**, 4091–4095.
53. Halmos,G., Arencibia,J.M., Schally,A.V., Davis,R. and Bostwick,D.G. (2000) High incidence of receptors for luteinizing hormone-releasing hormone (LHRH) and LHRH receptor gene expression in human prostate cancers. *J. Urol.*, **163**, 623–629.
54. Tolis,G., Ackman,D., Stellos,A., Mehta,A., Labrie,F., Fazekas,A.T., Comaru-Schally,A.M. and Schally,A.V. (1982) Tumor growth inhibition in patients with prostatic carcinoma treated with luteinizing hormone-releasing hormone agonists. *Proc. Natl Acad. Sci. USA*, **79**, 1658–1662.
55. Shi,X.B., Tepper,C.G. and White,R.W. (2008) MicroRNAs and prostate cancer. *J. Cell Mol. Med.*, **12**, 1456–1465.
56. Galardi,S., Mercatelli,N., Giorda,E., Massalini,S., Frajese,G.V., Ciafre,S.A. and Farace,M.G. (2007) miR-221 and miR-222 expression affects the proliferation potential of human prostate carcinoma cell lines by targeting p27Kip1. *J. Biol. Chem.*, **282**, 23716–23724.
57. Ozen,M., Creighton,C.J., Ozdemir,M. and Ittmann,M. (2008) Widespread deregulation of microRNA expression in human prostate cancer. *Oncogene*, **27**, 1788–1793.
58. Schaefer,A., Jung,M., Kristiansen,G., Lein,M., Schrader,M., Miller,K., Stephan,C. and Jung,K. (2008) MicroRNAs and cancer: Current state and future perspectives in urologic oncology. *Urol. Oncol.*, Dec 29 [Epub ahead of print].
59. Gottardo,F., Liu,C.G., Ferracin,M., Calin,G.A., Fassan,M., Bassi,P., Sevignani,C., Byrne,D., Negrini,M., Pagano,F. *et al.* (2007) Micro-RNA profiling in kidney and bladder cancers. *Urol. Oncol.*, **25**, 387–392.
60. Voorhoeve,P.M., le Sage,C., Schrier,M., Gillis,A.J., Stoop,H., Nagel,R., Liu,Y.P., van Duijse,J., Drost,J., Griekspoor,A. *et al.* (2007) A genetic screen implicates miRNA-372 and miRNA-373 as oncogenes in testicular germ cell tumors. *Adv. Exp. Med. Biol.*, **604**, 17–46.
61. Ambs,S., Prueitt,R.L., Yi,M., Hudson,R.S., Howe,T.M., Petrocca,F., Wallace,T.A., Liu,C.G., Volinia,S., Calin,G.A. *et al.* (2008) Genomic profiling of microRNA and messenger RNA reveals deregulated microRNA expression in prostate cancer. *Cancer Res.*, **68**, 6162–6170.
62. Murakami,Y., Yasuda,T., Saigo,K., Urashima,T., Toyoda,H., Okanoue,T. and Shimotohno,K. (2006) Comprehensive analysis of microRNA expression patterns in hepatocellular carcinoma and non-tumorous tissues. *Oncogene*, **25**, 2537–2545.
63. Calin,G.A., Ferracin,M., Cimmino,A., Di Leva,G., Shimizu,M., Wojcik,S.E., Iorio,M.V., Visone,R., Sever,N.I., Fabbri,M. *et al.* (2005) A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *N. Engl. J. Med.*, **353**, 1793–1801.
64. Cimmino,A., Calin,G.A., Fabbri,M., Iorio,M.V., Ferracin,M., Shimizu,M., Wojcik,S.E., Aqeilan,R.I., Zupo,S., Dono,M. *et al.* (2005) miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proc. Natl Acad. Sci. USA*, **102**, 13944–13949.
65. Zhang,B., Pan,X., Cobb,G.P. and Anderson,T.A. (2007) microRNAs as oncogenes and tumor suppressors. *Dev. Biol.*, **302**, 1–12.
66. Lin,S.L., Chiang,A., Chang,D. and Ying,S.Y. (2008) Loss of mir-146a function in hormone-refractory prostate cancer. *Rna*, **14**, 417–424.
67. Gerstein,M.B., Bruce,C., Rozowsky,J.S., Zheng,D., Du,J., Korbel,J.O., Emanuelsson,O., Zhang,Z.D., Weissman,S. and Snyder,M. (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Res.*, **17**, 669–681.
68. Ghindilis,A.L., Smith,M.W., Schwarzkopf,K.R., Roth,K.M., Peyvan,K., Munro,S.B., Lodes,M.J., Stover,A.G., Bernards,K., Dill,K. *et al.* (2007) CombiMatrix oligonucleotide arrays: genotyping and gene expression assays employing electrochemical detection. *Biosens. Bioelectron.*, **22**, 1853–1860.