frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 01 July 2013
doi: 10.3389/fgene.2013.00122

Abundances of microRNAs in human cells can be estimated as a function of the abundances of YRHB and RHHK tetranucleotides in these microRNAs as an ill-posed inverse problem solution

Mikhail P. Ponomarenko[1]*, Valentin V. Suslov[1], Petr M. Ponomarenko[1], Konstantin V. Gunbin[1], Irina L. Stepanenko[1], Oleg V. Vishnevsky[1,2] and Nikolay A. Kolchanov[1,2,3]

[1] *Department of Systems Biology, Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
[2] *Department of Natural Sciences, Novosibirsk State University, Novosibirsk, Russia*
[3] *Academic Council, National Research Centre "Kurchatov Institute", Moscow, Russia*

Edited by:
Peng Jin, Emory University School of Medicine, USA

Reviewed by:
Peng Jin, Emory University School of Medicine, USA
Tohru Yoshihisa, University of Hyogo, Japan

****Correspondence:***
Mikhail P. Ponomarenko, Department of Systems Biology, Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, 10 Lavrentyev Avenue, Novosibirsk 630090, Russia
e-mail: pon@bionet.nsc.ru

Mature microRNAs (miRNAs) are small endogenous non-coding RNAs 18–25 nt in length. They program the RNA Induced Silencing Complex (RISC) to make it inhibit either messenger RNAs or promoter DNAs. We have found that the mean abundance of miRNAs in Arabidopsis is correlated with the abundance of DRYD tetranucleotides near the 3′-end and the abundance of WRHB tetranucleotides in the center of the miRNA sequence. Based on this correlation, we have estimated miRNA abundances in seven organs of this plant, namely: inflorescences, stems, siliques, seedlings, roots, cauline, and rosette leaves. We have also found that the mean affinity of miRNAs for two proteins in the *Argonaute* family (Ago2 and Ago3) in man is correlated with the abundance of YRHB tetranucleotides near the 3′-end and that the preference of miRNAs for Ago2 is correlated with the abundance of RHHK tetranucleotides in the center of the miRNA sequence. This allowed us to obtain statistically significant estimates of miRNA abundances in human embryonic kidney cells, HEK293T. These findings in relation to two taxonomically distant entities (man and Arabidopsis) fit one another like pieces of a jigsaw puzzle, which allowed us to heuristically generalize them and state that the miRNA abundance in the human brain may be determined by the abundance of YRHB and RHHK tetranucleotides in these miRNAs.

Keywords: microRNA, *Argonote*, miRNA/Ago-affinity, miRNA abundance, quantitative sequence-activity relationship (QSAR), ill-posed inverse problem, linear-additive approximation, "limiting stage" approximation

INTRODUCTION

MicroRNAs (miRNAs) are small endogenous non-coding RNAs (Kozomara and Griffiths-Jones, 2010). Within the canonical biogenesis of miRNAs, their genes are transcribed by RNA polymerase II into the primary transcripts (pri-miRNAs). Special Microprocessor proteins cut away the first precursor of miRNA (pre-miRNA) and then the mature miRNA 18–25 nt in length (Kozomara and Griffiths-Jones, 2010). The miRNAs that maturate from other sources, including spliced-out introns (they are the source of mirtrons) and transfer RNAs (tRNAs), are called "non-canonical." They are less abundant in cells and their maturation time is deviant (Havens et al., 2012).

Mature miRNAs program the RISC (RNA-Induced Silencing Complex) to make it inhibit either messenger RNAs (mRNAs) or promoter DNAs (pDNAs) through the formation of mRNA(pDNA):miRNA-RISC complexes (Song et al., 2004). The function of the RISC depends upon what of the proteins in the *Argonaute* family is incorporated in the RISC (Gagnon and Corey, 2012). In the 3D structure of the mRNA:miRNA-Ago-RISC in Archaea (Song et al., 2004), the Ago protein interacts with the 3′-end of the miRNA.

Changes in mature miRNA abundance and sequence affecting interactions between the miRNAs and their targets were associated with various abnormalities, including neurodegeneration (Barbato et al., 2009) and cancer (Winter and Diederichs, 2011a). Winter and Diederichs (2011b) showed experimentally that the miRNA abundance in Ago2-deficient cells treated by the transcription inhibitor actinomycin D increases when the Ago2 protein is introduced to them ectopically, because the affinity of a particular miRNA for Ago2 protein influences the half-life of this miRNA in cells. Furthermore, Martinez and Gregory (2013) showed that Ago2 expression in mouse embryonic stem cells, originally low in Ago2 and then transfected by a vector containing Ago2, is dependent on miRNA abundance post-transcriptionally. Therefore, that miRNAs and human Ago2 stabilize each other is an experimentally established fact. How can we benefit from this fact?

Although nobody has measured *in vivo* the affinity of mature miRNAs for the different *Argonaute* proteins (Azuma-Mukai et al., 2008) or the abundance of these miRNAs in cells (Axtell and Bartel, 2005) under identical experimental conditions simultaneously, we have earlier demonstrated (Ponomarenko et al.,

2001) that the patterns and features found *in silico* in one experiment readily apply to the next, at least within the limits of applicability of the theory that underlies these experiments. Consequently, we have had to work with disembodied experimental data on two taxonomically distant entities (man and Arabidopsis) using original ACTIVITY tools (Ponomarenko et al., 1997). As a result, we have successfully found correlations which fit each other like pieces of a puzzle created in two experiments, one by Winter and Diederichs (2011b) and another by Martinez and Gregory (2013), the mutual complementarity of which was, in fact, the starting point of our work. These correlations allowed us to generalize them into a heuristic hypothesis stating that miRNA abundance in the human brain depends on the abundance of YRHB and RHHK tetranucleotides in these miRNAs. This hypothesis was further confirmed using independent experimental data taken from the Sestan Brain Atlases (Kang et al., 2011).

The results obtained are discussed in terms of the "limiting stage" approximation, the linear-additive approximation, and an ill-posed inverse problem. This allowed us to conclude that *in silico* estimates like these can reach an acceptable accuracy level for their practical consideration by cancer and neurodegeneration researches once the preference of these miRNAs for the proteins in the *Argonaute* family has become known, and so have yet unknown values of the affinity of any miRNA for two of the four proteins (50%), Ago1 and Ago4, which is absolutely required for a more accurate approximation.

MATERIALS AND METHODS

NUCLEOTIDE SEQUENCES

The nucleotide sequences of the mature canonical Arabidopsis miRNAs $\{\xi_i\}$ were taken from a work by Axtell and Bartel (2005), $\xi \in \{a, u, g, c\}$. Seventeen out of 27 miRNAs were used as the training dataset (**Table 1**). Because miRNA lengths varied from 20 to 22 nt, our *in silico* processing was only confined to miRNAs of a given length (in **Tables 1**, **2**, these sequences are typed in CAPITALS).

The other 10 miRNAs (**Figures 6**, **7**) were used as an independent experimental dataset (sequences not shown). Twenty-two Arabidopsis miRNAs taken from a work by Lu et al. (2005) were used as independent experimental control datasets (**Figure 4C**; sequences not shown).

The nucleotide sequences of human mature miRNA were taken from a work by Azuma-Mukai et al. (2008). Twelve out of 28 mature canonical miRNAs were used as the training dataset (**Table 2**).

The other 16 canonical miRNAs were used as independent experimental control datasets (**Figure 3**), and 48 miRNAs named the "individual variants" by Azuma-Mukai et al. (2008) because of their 5′- and/or 3′-terminal differences from canonical mature miRNAs, which were associated by Azuma-Mukai et al. (2008) with (i) alternative maturation (Azuma-Mukai et al., 2008) or (ii) post-maturation processing (Azuma-Mukai et al., 2008), 96 human miRNAs taken from a work by Bail et al. (2010), and 318 human miRNAs taken from miRBase (Kozomara and Griffiths-Jones, 2010) according to their identifiers in the Sestan Brain

Table 1 | The training dataset (this work) of miRNA abundances in Arabidopsis (Axtell and Bartel, 2005) contains 17 miRNA fragments each 20 nt in length (they are in CAPITALS).

Experimental data *in vivo* (Axtell and Bartel, 2005)			Analysis *in silico* (this work)	
miRNA	**Canonical miRNA sequence**	**ln[miRNA], ln-unit**	**$[WRHW]_{F1}$**	**$[DRYD]_{F2}$**
ath-mir-156	UGACAGAAGAGAGUGAGCAC	3.09	0.97	2.08
ath-mir-157	UUGACAGAAGAUAGAGAGCAc	1.72	1.89	1.02
ath-mir-158	UCCCAAAUGUAGACAAAGCA	4.85	1.79	2.01
ath-mir-159	UUUGGAUUGAAGGGAGCUCUa	5.21	1.30	1.40
ath-mir-160	UGCCUGGCUCCCUGUAUGCCa	3.81	0.69	2.43
ath-mir-161.1	UGAAAGUGACUACAUCGGGGt	4.68	1.22	1.87
ath-mir-161.2	UCAAUGCAUUGAAAGUGACUa	3.90	1.78	1.63
ath-mir-163	UUGAAGAGGACUUGGAACUUcgau	1.96	0.61	1.70
ath-mir-164	UGGAGAAGCAGGGCACGUGCa	4.24	1.64	1.46
ath-mir-165	UCGGACCAGGCUUCAUCCCCc	0.90	0.00	0.70
ath-mir-166	UCGGACCAGGCUUCAUUCCCc	1.48	0.00	0.70
ath-mir-168	UCGCUUGGUGCAGGUCGGGAa	4.02	1.00	1.25
ath-mir-169	CAGCCAAGGAUGACUUGCCGa	2.11	0.00	1.61
ath-mir-171	UGAUUGAGCCGCGCCAAUAUc	2.02	0.48	1.17
ath-mir-390	AAGCUCAGGAGGGAUAGCGCc	2.65	0.43	2.13
ath-mir-394	UUGGCAUUCUGUCCACCUCC	2.00	0.00	0.25
ath-mir-398	UGUGUUCUCAGGUCACCCCUg	1.74	0.57	0.35
	Coefficient of linear correlation		$r = 0.67$	$r = 0.58$
	Statistical significance		$\alpha < 0.005$	$\alpha < 0.025$

Table 2 | The training dataset (this work) of miRNA affinities for the human Ago2 and Ago3 proteins (Azuma-Mukai et al., 2008); 12 miRNA fragments each 22 nt in length are in CAPITALS.

Experimental data *in vivo* (Azuma-Mukai et al., 2008)				Analysis *in silico* (this work)			
miRNA	**Canonical miRNA sequence**	**[miR/Ago2] ln-un., X_2**	**[miR/Ago3] ln-un., X_3**	**Δ, $(X_2 - X_3)/2$**	**$[RHHK]_{F3}$**	**Σ, $(X_2 + X_3)/2$**	**$[YRHB]_{F4}$**
hsa-mir-342	*u*CUCACACAGAAAUCGCACCCGU	7.81	4.85	1.48	1.95	6.33	0.77
hsa-mir-21	UAGCUUAUCAGACUGAUGUUGA	8.34	7.24	0.55	1.59	7.79	1.45
hsa-mir-378	ACUGGACUUGGAGUCAGAAGGC	4.97	5.89	−0.46	1.26	5.43	0.00
hsa-mir-629	GUUCUCCCAACGUAAGCCCAGC	5.28	7.98	−1.35	0.91	6.63	0.37
hsa-mir-92b	UAUUGCACUCGUCCCGGCCUCC	6.48	8.98	−1.25	0.22	7.73	1.06
hsa-mir-221	AGCUACAUUGUCUGCUGGGUUU	6.08	5.26	0.41	2.37	5.67	0.50
hsa-mir-29c	UAGCACCAUUUGAAAUCGGUUA	6.08	6.05	0.02	1.36	6.06	0.42
hsa-mir-210	CUGUGCGUGUGACAGCGGCUGA	6.81	7.96	−0.57	0.59	7.38	1.32
hsa-mir-let7d	CUAUACGACCUGCUGCCUUUCU	6.49	6.46	0.01	1.53	6.47	1.00
hsa-mir-99b	CACCCGUAGAACCGACCUUGCG	8.55	6.70	0.92	1.43	7.62	2.09
hsa-mir-191	CAACGGAAUCCCAAAAGCAGCU	6.40	7.47	−0.53	1.05	6.94	0.82
hsa-mir-425	*a*AUGACACGAUCACUCCCGUUGA	7.33	8.50	−0.59	0.10	7.92	1.99
		Coefficient of linear correlation		$r = 0.75$		$r = 0.86$	
		Statistical significance		$\alpha < 0.005$		$\alpha < 0.001$	

Atlases (Kang et al., 2011) (**Figures 7**, **8**, **9**, respectively; sequences not shown).

BIOLOGICAL ACTIVITY

The relative values ranging from −0.5 to 7.8 ln for the miRNA abundance in Arabidopsis taken from a work by Axtell and Bartel (2005) are partly presented in **Table 1** and fully in **Figures 3**, **4** (the *y*-axis).

The relative values ranging from 0 to 4 ln for the miRNA abundance in Arabidopsis obtained using Massively Parallel Signature Sequencing (MPSS) were taken from a work by Lu et al. (2005) and used as an independent experimental control dataset (**Figure 4**, the *y*-axis).

The values ranging from 4.85 to 9.43 ln for the *in vivo* measured affinity of canonical miRNAs for the human Ago2 and Ago3 proteins were taken from a work by Azuma-Mukai et al. (2008), are partly presented in **Table 2** and fully in **Figures 5**, **6** (the *y*-axis), while those for the affinity of 48 miRNAs named the "individual variants" by Azuma-Mukai et al. (2008) because of their 5′- and/or 3′-terminal differences from canonical mature miRNAs, which were associated by Azuma-Mukai et al. (2008) with (i) alternative maturation (Azuma-Mukai et al., 2008) or (ii) post-maturation processing (Azuma-Mukai et al., 2008), are shown in **Figure 7**.

The relative values ranging from −9.0 to 0.0 ln for the abundance of 96 human miRNAs in the human embryonic kidney cells HEK293T, some preincubated for 8 h with the transcription inhibitor actinomycin D and others not preincubated, were taken from a work by Bail et al. (2010) and used as an independent experimental control dataset (**Figure 8**, the *y*-axis).

The relative values ranging from 0 to 16 rel. un. for the miRNA abundance measured within 95 human brain regions or neocortical areas were taken from the Sestan Brain Atlases (Kang et al., 2011) and used as an independent experimental control dataset (**Figure 9**, the *y*-axis).

CORRELATIONS BETWEEN BIOLOGICAL ACTIVITY AND miRNA NUCLEOTIDE SEQUENCES

We have used our original development called ACTIVITY (Ponomarenko et al., 1997), which is a tool intended for the processing of input data on a pre-set biological activity, $X(\{\xi_i\})$ in known miRNA sequences, $\{\xi_i\}$ and searching for correlations in them.

Although ACTIVITY has been described in detail elsewhere (Ponomarenko et al., 1997), we will additionally provide a brief descriptions of its features that were critical to our current study.

First of all, ACTIVITY (Ponomarenko et al., 1997) searches for correlations between the biological activity of a miRNA, $X(\{\xi_i\})$, (expressed as its expression level; herein, as the experimentally measured abundance and miRNA/Ago affinity) and the weighted abundance of the tetranucleotides, $[z_1z_2z_3z_4]_F$, in the sequence $\{\xi_i\}$ of this miRNA:

$$[z_1z_2z_3z_4]_F\left\{\xi_{1\le i\le L}\right\} = \sum_{\substack{1 \le i \le L-3 \\ \xi_i\xi_{i+1}\xi_{i+2}\xi_{i+3} \in z_1z_2z_3z_4}} F(i); \quad (1)$$

where $z \in \{a, u, g, c, w, r, m, k, y, s, b, v, h, d, n\}$ (IUPAC-IUB, 1971); $0 \le F(i) \le 1$ is the weight of the tetranucleotide $z_1z_2z_3z_4$ at the i-th position, with which we heuristically assessed its linear additive contribution to the $X(\{\xi_i\})$ value using the rule "the higher F(i), the greater the contribution" (**Figure 1**).

ACTIVITY (Ponomarenko et al., 1997) has built-in F(i) profiles: 180 U-shaped and 180 S-shaped curves for F(i) values, in which low and high weights have different locations and interval lengths. ACTIVITY works uniformly on each of all the possible

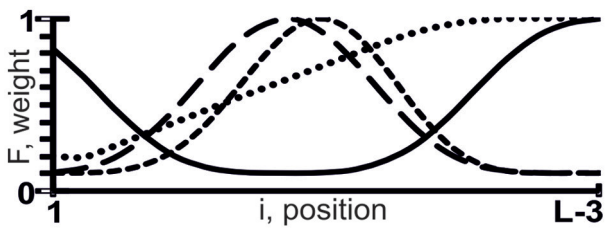

FIGURE 1 | Sample weights (the *y*-axis), F(i), of tetranucleotide $z_1z_2z_3z_4$ at the *i*-th position (the *x*-axis) of the sequence $\{\xi_i\}$ composed of L nt, with which we assessed the linear additive contribution to $X(\{\xi_i\})$ using the rule "the higher F(i), the greater the contribution."

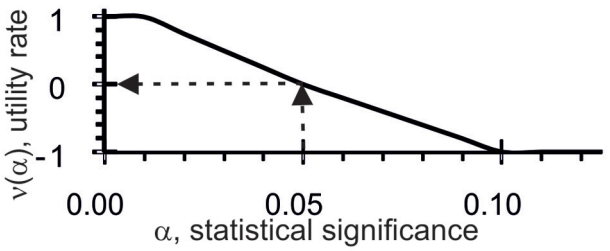

FIGURE 2 | Utility estimates (the *y*-axis), $\upsilon(\alpha)$, for correlations between $X(\{\xi_i\})$ and $[z_1z_2z_3z_4]_F(\{\xi_i\})$ based on statistical significance (α, the *x*-axis) for the given dataset.

variants of weighted tetranucleotide abundance ($[z_1z_2z_3z_4]_F$), their number being $360 \times 15^4 = 18225000 \approx 10^7$.

Furthermore, each $[z_1z_2z_3z_4]_F\{\xi_i\}$ value was compared with $X(\{\xi_i\})$ using bootstrapping (Hayes et al., 1989) in 7 subsets: (i) the entire dataset; (ii) 50% of the entries that have the lowest $[z_1z_2z_3z_4]_F\{\xi_i\}$ values; (iii) 50% of entries that have the highest $[z_1z_2z_3z_4]_F\{\xi_i\}$ values; (iv) 50% of the entries that are closest to the mean of all the $[z_1z_2z_3z_4]_F\{\xi_i\}$ values; (v) 50% of the entries that have the lowest $X(\{\xi_i\})$ values; (vi) 50% of the entries that have the highest $X(\{\xi_i\})$ values; (vii) 50% of the entries that are closest to the mean of all the $X(\{\xi_i\})$ values. By varying statistical test data [bootstrapping (Hayes et al., 1989)], we seek to minimize the dependence of search results on the input dataset.

In each of these seven subsets, ACTIVITY (Ponomarenko et al., 1997) checks five types of correlation between $X(\{\xi_i\})$ and $[z_1z_2z_3z_4]_F(\{\xi_i\})$: i) linear correlation; ii) Spearman's rank correlation; iii) Kendall's rank correlation; iv) dichotomous correlations tested by χ^2; and v) dichotomous correlations tested by the Fisher-Irwin test. Because it is possible to obtain quantitative estimates using linear correlations, such correlations could be useful if it were not for their sensitivity to data heterogeneity. By contrast, dichotomous correlations do not depend on data heterogeneity; however, they provide the least informative estimates above/below any pre-set threshold. Based on the usefulness-to-robustness ratio, rank correlations are between linear and dichotomous correlations. We search the input training dataset for different types of correlation and identify the best trade-offs.

Furthermore, ACTIVITY (Ponomarenko et al., 1997) checks the following six criteria of the applicability of regression analysis to the $\{[z_1z_2z_3z_4]_F(\{\xi_i\}); X(\{\xi_i\})\}$ data: i) how uniform the $X(\{\xi_i\})$ values are; ii) how uniform the $[z_1z_2z_3z_4]_F(\{\xi_i\})$ values are; iii) whether the departures of the $[z_1z_2z_3z_4]_F(\{\xi_i\})$ values from the $\{\lambda X(\{\xi_i\})+\mu\}$ regression are normal; iv) whether the departures of $[z_1z_2z_3z_4]_F(\{\xi_i\})$ from $\{\lambda X(\{\xi_i\}) + \mu\}$ independent of each other; v) whether the departures of the $X(\{\xi_i\})$ values from the $\{\varphi[z_1z_2z_3z_4]_F(\{\xi_i\})+\psi\}$ regression are normal; and vi) whether the departures of the $X(\{\xi_i\})$ values from $\{\varphi[z_1z_2z_3z_4]_F(\{\xi_i\}) + \psi\}$ independent of each other. We search the input training dataset for the correlations that satisfy additional criteria for their applicability to making *in silico* estimates, for example, for estimating unknown $X(\{\xi_i\})$ values from known $\{\xi_i\}$ values based on the weighted abundance, $[z_1z_2z_3z_4]_F(\{\xi_i\})$.

Thus, ACTIVITY (Ponomarenko et al., 1997) checks 11 criteria (five types of correlation and six criteria of the applicability of regressions) on each of 10^7 $[z_1z_2z_3z_4]_F$ variants and in each of 7 subsets yields $11 \times 7 = 77$ values for the statistical significance (α). In terms of fuzzy set theory (Zadeh, 1965) and utility theory for decision making (Fishburn, 1970), each α is transformed into a numerical utility value, $\alpha \rightarrow \upsilon(\alpha)$, for the correlation between $X(\{\xi_i\})$ and $[z_1z_2z_3z_4]_F(\{\xi_i\})$, see **Figure 2**.

As can be seen, the threshold α set at 0.05 was the reference point for $\upsilon = 0$: when the tests were statistically significant, the utility values were positive and when the tests were not statistically significant, the utility values were negative. Each $[z_1z_2z_3z_4]_F$ estimate in the input training dataset, $X(\{\xi_i\})$ and $\{\xi_i\}$, was equal to their mean:

$$\Xi\left([z_1z_2z_3z_4]_F\{\xi_i\}; X\{\xi_i\}\right) = \frac{1}{77}\sum_{k=1}^{7}\sum_{q=1}^{11}\left(\alpha_{kq}\right). \qquad (2)$$

Finally, in the given input training dataset $\{\{\xi_i\}; X(\{\xi_i\})\}$, ACTIVITY (Ponomarenko et al., 1997) finds the only $[z_1z_2z_3z_4]_F$ value with the highest $\Xi([z_1z_2z_3z_4]_F(\{\xi_i\});X(\{\xi_i\})) > 0$ or infers that the correlations found with the input training dataset are useless.

VERIFICATION OF THE CORRELATIONS FOUND

Because ACTIVITY finds the only best correlation from among 10^7 variants in any given input training dataset (Ponomarenko et al., 1997), verification is absolutely required.

First of all, the Bonferroni test yields $p(\Xi > 0) < 10^{-20}$ (Omelyianchuk et al., 2011). This implies that it is quite unlikely that half of the 77 tests run can be satisfied simultaneously for random chance at $\alpha < 0.05$.

Also, the training dataset, $\{\xi_i\}$ and $X(\{\xi_i\})$, is brought to the permutation test (Sohn et al., 2009): all the data are randomly rearranged $\{\{\xi_i^{\#}\}; X(\{\xi_i^{\$}\})\}$ and fed to ACTIVITY. The lack of the same $[z_1z_2z_3z_4]_F$ value or its correlated $[z_1^{\#}z_2^{\$}z_3^{\&}z_4^{*}]_F$ value in 100 independent cycles is statistically significant ($\alpha < 0.05$; binomial law).

In turn, the statistical significance of the correlation found by ACTIVITY (Ponomarenko et al., 1997) with the training dataset

is tested for on the control dataset with unprocessed experimental data (**Figures 3**, **5**).

Finally, the statistical significance of the correlations found by ACTIVITY (Ponomarenko et al., 1997) with the given training dataset is tested using independent experimental data (**Figures 4**, **7–9**).

CLUSTER ANALYSIS OF THE miRNAs

In this work, we used standard statistical tools available in the STATISTICA system (Afifi et al., 2003), which has the "Joining (tree clustering)" mode in "Cluster" section under the "Multivariate/Exploratory" option in the "Statistics" part. Under this mode, we clustered all the RNAs being studied using all $42 = 7 \times 6$ possible combinations of seven Linkage rules: "Single linkage," "Complete linkage," "Unweighted pair-group average," "Weighted pair-group average," "Unweighted pair-group centroid," "Weighted pair-group centroid," and "Ward's methods," and each from among six "Distance measures": "Squared Euclidian distance," "Euclidian distance," "City-block (Manhattan) distance," "Chebychev distance metric," "Power," "Percent disagreement," and "1-Pearson r." The color-coded results obtained from the most widely used (predefined) combination of the "Single linkage" rule and the "Euclidian distance" metric are shown in **Figure 9**. The results obtained from each of the other 41 combinations are not shown, because they have only minor deviations (less than 1% of RNAs) in the vicinity of the intercluster boundary caused by heterogeneity in experimental data.

RESULTS AND DISCUSSION

miRNA ABUNDANCE IN ARABIDOPSIS

We ran ACTIVITY (Ponomarenko et al., 1997) on the experimental data (Axtell and Bartel, 2005) on the mean abundance of mature canonical ubiquitous miRNAs in Arabidopsis (**Table 1**, ln[miRNA]). We composed a training dataset (**Table 1**) for ACTIVITY (Ponomarenko et al., 1997) consisting of all the variants that had the lowest and highest values for miRNA abundance in seven organs of this plant (inflorescence, stem, silique, seedling, root, cauline and rosette leaves) and the occurrence of nucleotides A, U, G, C, W, R, and K in miRNAs (IUPAC-IUB, 1971). The resulting 17 out of 27 miRNAs in the training dataset (**Table 1**) represent the ranges of values of miRNA properties rather than data heterogeneity (Azuma-Mukai et al., 2008). The other 10 miRNAs were used as an independent experimental control dataset (**Figure 3**).

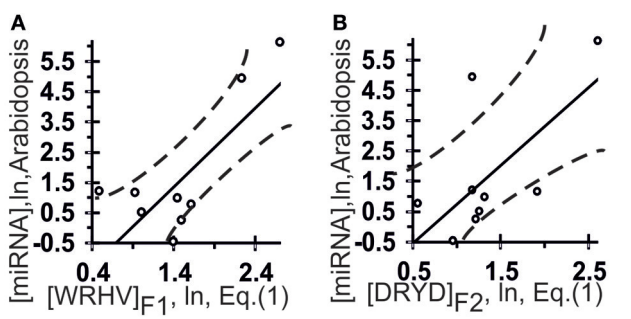

FIGURE 3 | Control test of the patterns (the *x*-axis) found by ACTIVITY (Ponomarenko et al., 1997) in the training dataset (Table 1) using independent experimental data (the *y*-axis) taken from the same data source (Axtell and Bartel, 2005). Two linear correlations in Arabidopsis: one **(A)** between miRNA abundance in the plant and $[WRHB]_{F3}$ abundance in these miRNAs and one **(B)** between miRNA abundance and $[DRYD]_{F4}$ abundance. Both are statistically significant in the control dataset of 11 miRNAs (Axtell and Bartel, 2005). Dashed curves depict 95% confidence intervals for linear regression (solid lines) built using STATISTICA (Afifi et al., 2003).

The highest estimated value, $\Xi = 0.48$, was assigned [Equation (2)] to the correlation between the abundance of miRNAs (ln[miRNA]) in Arabidopsis and the abundance of WRHW tetranucleotides ($[WRHW]_{F1}$) with its highest weight, F1(i), in the center of the miRNA (**Figure 1**, short-dashed line). In the control dataset, the correlation between $[WRHW]_{F1}$ and ln[miRNA] was statistically significant (**Figure 3A**: $r = 0.74$, $\alpha < 0.025$).

Two more Ξ values were equal to 0.46 at the same WRHW tetranucleotide with narrower peaks (Ponomarenko et al., 2008). No other values $\Xi > 0$ were found in the training set. The next eight $z_1z_2z_3z_4$ tetranucleotides that had the highest Ξ-values were RYHV ($\Xi_{MAX} = -0.01$), RHWV (-0.05), DYDR (-0.07), SNKH (-0.07), RHWR (-0.08), SWBH (-0.09), BRHR (-0.11), and SVKH (-0.12) (in descending order).

For the S-shaped weights, the highest estimated value, $\Xi = 0.47$, was assigned [Equation (2)] to the abundance of the DRYD tetranucleotide ($[DRYD]_{F2}$) with its highest weight, F2(i), at the 3′-end of the miRNA (**Figure 1**, dotted line). In the control dataset, the correlation between $[DRYD]_{F2}$ and ln[miRNA] was statistically significant (**Figure 3B**: $r = 0.66$, a < 0.05). No other values $\Xi > 0$ were found in the training set (Ponomarenko et al., 2008). The next nine $z_1z_2z_3z_4$ tetranucleotides that had the highest Ξ-values were SNYW ($\Xi_{MAX} = -0.02$), WVVM (-0.04), RVYR (-0.05), VAHS (-0.06), VRDS (-0.07), RDMW (-0.08), DYDR (-0.09), RHWK (-0.18), and SVKH (-0.23) (in descending order).

Because $[WRHW]_{F1}$ and $[DRYD]_{F2}$ were independent ($r = 0.39$, $\alpha > 0.25$), we skipped the optimization procedure and derived the following formula:

$$[\text{miRNA}]\{\xi_j\} = 0.78 + 1.31\,[\text{WRHW}]_{F1}\{\xi_j\} + 0.76\,[\text{DRYD}]_{F2}\{\xi_j\}. \quad (3)$$

Estimates made with Equation (3) were statistically significantly (**Figure 4**, **Table 3**: $r = 0.59$, $\alpha < 0.0025$) correlated with data reported by Axtell and Bartel (2005). As can be seen from **Figure 4**, the *in vivo* measured abundances of most miRNAs with calculated abundances ranging from 3.0 to 5.0 ln range from -0.5 to 5.5 ln (almost the full range of the graph). That is why the high *r*-value of the regression is probably due to the contribution of several anomalies like the most abundant miRNAs. To see if it is as it appears to be, we additionally estimated Spearmen's rank correlation coefficient (**Table 3**: $R = 0.54$; $\alpha < 0.005$) and Kendall's rank correlation coefficient (**Table 3**: $\tau = 0.38$; $\alpha < 0.01$) (they

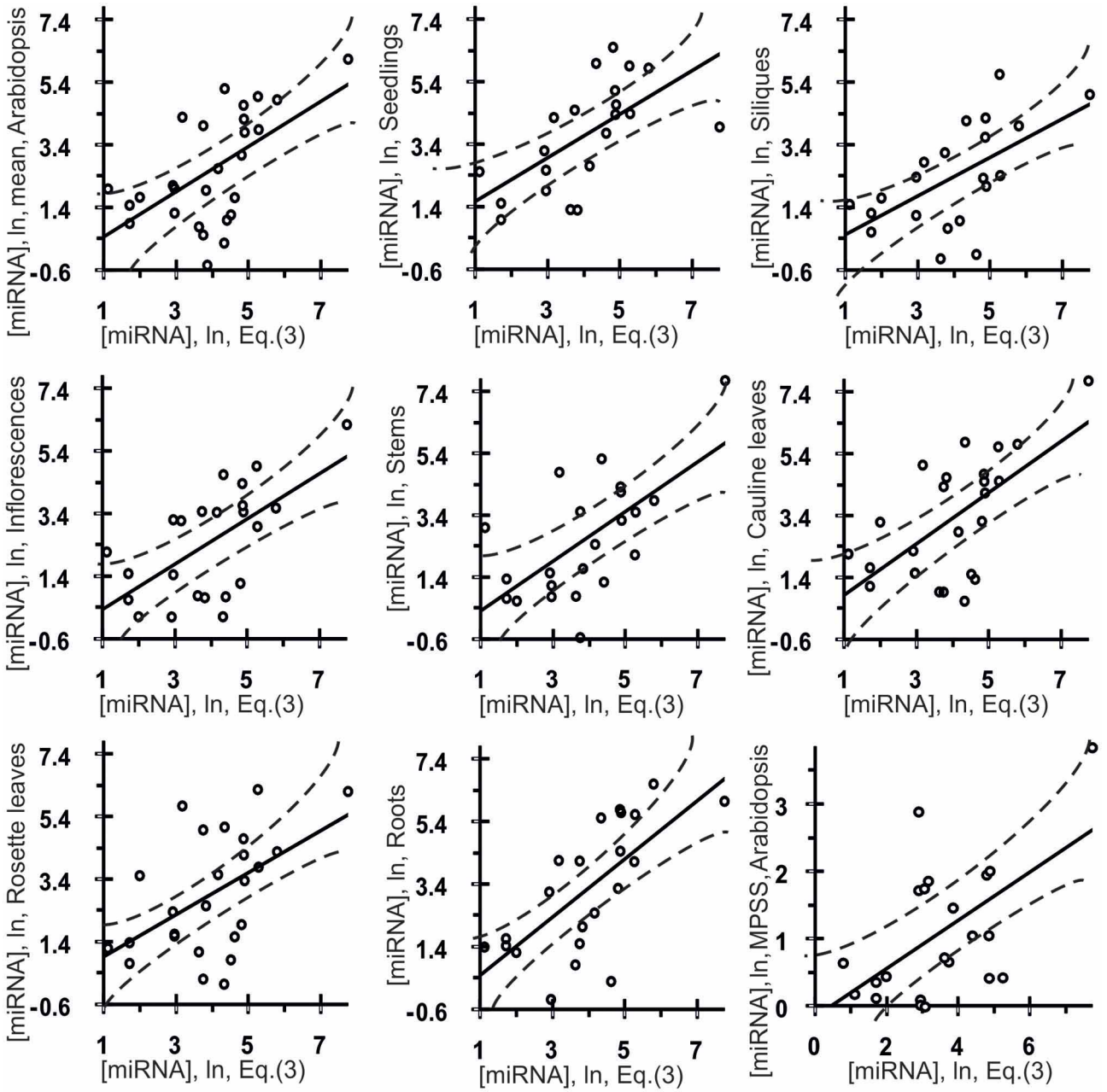

FIGURE 4 | Control test of the final equation (3), the *x*-axis, derived with ACTIVITY outputs (Ponomarenko et al., 1997) on the training dataset (Table 1) and using independent experimental data (the *y*-axis) taken from various data sources. Twenty seven correlations in Arabidopsis (**Table 3**) between the miRNA abundance estimated by Equation (3), the *x*-axis, and those measured experimentally, the *y*-axis, namely: the mean abundances of 28 canonical miRNAs used above, the organ-specific abundance of these miRNAs in seedlings, siliques, inflorescences, stems, cauline leaves, rosette leaves, and roots as independently measured *in vivo* (Axtell and Bartel, 2005), and, finally, the abundances of 22 miRNAs obtained using Massively Parallel Signature Sequencing (MPSS) by Lu et al. (2005) as independent experimental control datasets. Dashed curves and solid lines as in the legend to **Figure 3**.

consider the ranks of [miRNA] values rather than their true values) for *in silico* [miRNA] values (the *x*-axis) and *in vivo* [miRNA] values (the *y*-axis).

Also, **Figure 4** shows seven statistically significant linear correlations between the estimates obtained using Equation (3) and the abundances of miRNA in seven organs of Arabidopsis (Axtell and Bartel, 2005), namely: inflorescences, stems, siliques, seedlings, roots, cauline leaves and rosette leaves. The dashed curves in this figure depict the boundaries of the 95% confidence interval for the mean miRNA abundance in Arabidopsis estimated by

Table 3 | Twenty-seven correlations in Arabidopsis (Figure 4) between the miRNA abundance estimated by Equation (3) and those measured experimentally.

Experimental dataset		Linear correlation		Spearmen's rank correlation		Kendall's rank correlation	
No.	**(Reference) organ**	**Coefficient *r***	**Significance α**	**Coefficient *R***	**Significance α**	**Coefficient τ**	**Significance α**
	Axtell and Bartel, 2005						
1	Means, plant	0.63	<0.001	0.54	<0.005	0.37	<0.01
2	Seedlings	0.63	<0.0025	0.69	<0.001	0.47	<0.005
3	Siliques	0.59	<0.005	0.60	<0.005	0.43	<0.01
4	Inflorescences	0.62	<0.0025	0.62	<0.0025	0.42	<0.005
5	Stems	0.64	<0.0025	0.59	<0.005	0.40	<0.01
6	Cauline leaves	0.63	<0.001	0.54	<0.005	0.37	<0.01
7	Rosette leaves	0.52	<0.01	0.51	<0.001	0.35	<0.025
8	Roots	0.69	<0.0005	0.73	<0.00025	0.55	<0.0005
	Lu et al., 2005						
9	Whole plant, MPSS	0.56	<0.01	0.46	<0.05	0.30	<0.05

MPSS, Massively Parallel Signature Sequencing.

Equation (3). As can be seen, despite the statistical significance in the correlations between the value defined by Equation (3) and miRNA abundance, a large part of data points in **Figure 4** exist outside of the dashed lines. This implies that Equation (3) is an adequate source of rough estimates of miRNA abundances in Arabidopsis organs; however, there is a high variability of their organ-specific values (the coefficient of variation, $C_V = \sigma/M_0 \times 100\%$, expressed as the percentage of the ratio between the standard deviation and the mean, ranging from 7 to 72%, the mean being 31 ± 19%) which was ignored by Equation (3) due to lack of data.

Finally, the three above mentioned correlations between *in silico* and *in vivo* [miRNA] values were statistically significant in the independent experimental dataset (Lu et al., 2005) [**Table 3**: $r = 0.56$ ($\alpha < 0.01$), $R = 0.46$ ($\alpha < 0.05$), $\tau = 0.30$ ($\alpha < 0.05$)]. Therefore, the statistical significance of 27 independent tests (**Figure 4** and **Table 3**) is rather an argument for than against a dependence of miRNA abundance on tetranucleotide abundance in these miRNAs.

However, the molecular mechanism that Equation (3) is consistent with remains unclear. Admittedly, Hwang et al. (2007) established experimentally that the sequence of a mature miRNA is a factor for the efficiency of its export from the nucleus to the cytoplasm, and Gantier et al. (2011) explored effects of Dicer1 on the miRNA half-life in a context dependent manner (Gantier et al., 2011). If we were to consider Equation (3) together with the results of the experiments performed by Winter and Diederichs (2011b) and by Martinez and Gregory (2013) suggesting that miRNAs and Ago2 are likely to stabilize each other, it could be admitted that Equation (3) implies miRNA/Ago2 affinity.

miRNA/Ago AFFINITY IN MAN

We ran ACTIVITY (Ponomarenko et al., 1997) simultaneously on two libraries (**Table 2**) of mature human miRNAs specific for either Ago2 or Ago3 (Azuma-Mukai et al., 2008). To this end, instead of using the affinity magnitude [miRNA/Ago2] and [miRNA/Ago3], we heuristically constructed two auxiliary estimates:

$$\Sigma = \big([\text{miRNA/Ago2}] + [\text{miRNA/Ago3}]\big)/2;$$
$$\Delta = \big([\text{miRNA/Ago2}] - [\text{miRNA/Ago3}]\big)/2. \quad (4)$$

We composed a training dataset (**Table 2**) for ACTIVITY (Ponomarenko et al., 1997) consisting of all the variants that had the lowest and highest values for Σ, Δ, the abundance of nucleotides A, U, G, C, W, R, and K in miRNAs (IUPAC-IUB, 1971). The resulting 12 miRNAs in the training dataset (**Table 2**) represent the ranges of values of miRNA properties rather than data heterogeneity (Azuma-Mukai et al., 2008).

The highest estimated value, $\Xi = 0.36$, was assigned [Equation (2)] by ACTIVITY (Ponomarenko et al., 1997) to the correlation between Δ and the abundance, $[RHHK]_{F3}$, of the RHHK tetranucleotide (IUPAC-IUB, 1971) with its highest weight, F3(i), in the center of the miRNA (**Figure 1**, broken line). This corresponds to the difference that Ago2 and Ago3 have in cleaving the mRNA in the center of its complementarity with the miRNA-Ago2(3)-RISC complex (Song et al., 2004). The $[RHHK]_{F1}$ values for all the 12 miRNAs of the training dataset are presented in **Table 2**. The correlation between $[RHHK]_{F3}$ and Δ was statistically significant ($r = 0.75$, $a < 0.005$), and so was that in the control dataset (**Figure 5A**: $r = 0.51$, $a < 0.05$). Another Ξ value, 0.34, indicated at the same tetranucleotide, RHHK, with a narrower peak, F(i), in the center of the miRNA. No other higher-than-zero Ξ values were found in the training dataset (Omelyianchuk et al., 2011). The next eight tetranucleotides that had the highest Ξ-values were WRHH ($\Xi_{MAX} = -0.01$), RBBM (−0.01), RDDK (−0.01), ABMD (−0.02), YWBM (−0.02), RBMD (−0.02), DSSV (−0.04), and WRMH (−0.06) (in descending order).

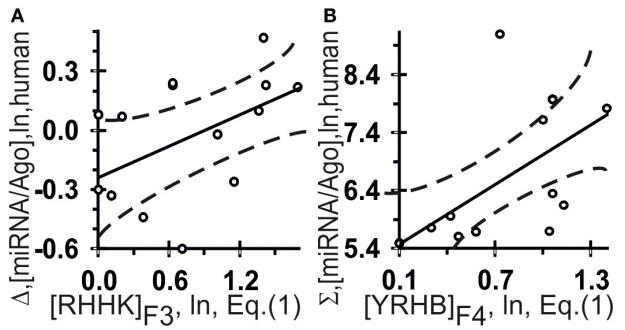

FIGURE 5 | Control test of the patterns (the *x*-axis) found by ACTIVITY (Ponomarenko et al., 1997) in the training dataset (Table 2) using independent experimental data (the *y*-axis) taken from same data source (Azuma-Mukai et al., 2008). Two statistically significant linear correlations: one **(A)** between the miRNA/Ago-affinity estimate (Δ) and $[RHHK]_{F3}$; and one **(B)** between another estimate (Σ) and $[YRHB]_{F4}$. Both are statistically significant in the control dataset of 16 canonical miRNAs (Azuma-Mukai et al., 2008). Dashed curves and solid lines as in the legend to **Figure 3**.

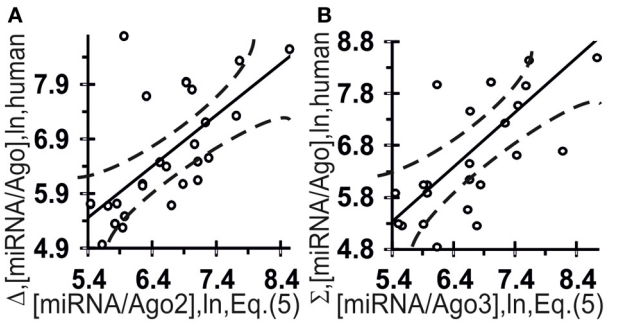

FIGURE 6 | Control test of the final Equation (5), the *x*-axis, derived with ACTIVITY outputs (Ponomarenko et al., 1997) on the training dataset (Table 2) and using independent experimental data (the *y*-axis) on the canonical miRNAs taken from the same data source (Azuma-Mukai et al., 2008). miRNA/Ago affinity as measured *in vivo* (Azuma-Mukai et al., 2008) and as estimated *in silico* and expressed in logarithms are statistically significantly correlated for Ago2 **(A)** and Ago3 **(B)**. Dashed curves and solid lines as in the legend to **Figure 3**.

For Σ, the highest estimated value, $\Xi = 0.36$, was assigned [Equation (2)] to the abundance, $[YRHB]_{F4}$, of the YRHB tetranucleotide (IUPAC-IUB, 1971) with its highest weight F4(i) at the 3′-end of the miRNA (**Figure 1**, solid line). This corresponds to the contact of the miRNA and the Ago protein in the 3D structure of the mRNA:miRNA-Ago-RISC complex (Song et al., 2004). The correlation between $[YRHB]_{F4}$ and Σ was statistically significant in the control dataset (**Figure 5B**: $r = 0.61$, $\alpha < 0.025$). No other values $\Xi > 0$ were found in the training set (Omelyianchuk et al., 2011). The next nine tetranucleotides that had the highest Ξ-values were RBMB ($\Xi_{MAX} = (-0.01)$, ANKK (−0.01), DKSM (−0.12), MHKR (−0.13), YHKD (−0.14), HASH (−0.16), WNNS (−0.17), DDSM (−0.19), KMDK (−0.21) (in descending order).

Figure 6 shows independent estimates made on the basis of these two correlations for the affinity of miRNAs for Ago2 ($[miRNA/Ago2] = \Sigma + \Delta$) and Ago3 ($[miRNA/Ago3] = \Sigma - \Delta$) derived without optimization:

$$\begin{aligned}[miRNA/Ago2]\{\xi_j\} &= 4.97 + 0.52[RHHK]_{F3}\{\xi_j\} \\ &\quad +1.35[YRHB]_{F4}\{\xi_j\}; \\ [miRNA/Ago3]\{\xi_j\} &= 6.11 - 0.52[RHHK]_{F3}\{\xi_j\} \\ &\quad +1.35[YRHB]_{F4}\{\xi_j\}. \end{aligned} \quad (5)$$

They are statistically significantly [**Figure 6**: (A) $r = 0.66$ and (B) $r = 0.66$, $\alpha < 0.00025$] correlated with all the experimental data (Azuma-Mukai et al., 2008).

Figure 7 shows independent *in silico* estimates obtained using Equation (5) for 48 miRNAs named the "individual variants" by Azuma-Mukai et al. (2008) because of their 5′- and/or 3′-terminal differences from canonical mature miRNAs, which were associated by Azuma-Mukai et al. (2008) with (i) alternative maturation (Azuma-Mukai et al., 2008) or (ii) post-maturation

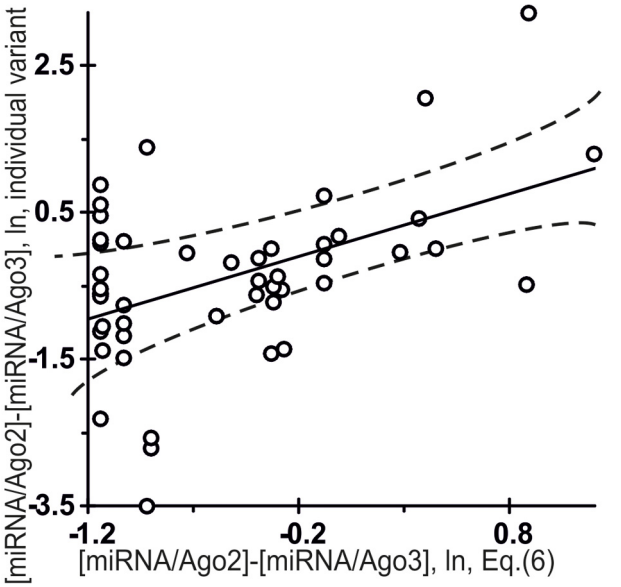

FIGURE 7 | Control test for the difference between the affinity of miRNAs for Ago2 and that for Ago3 (Equation (6), the *x*-axis) estimated using final Equation (5) and independent experimental data (the *y*-axis) on the miRNA individual variants taken from the same data source (Azuma-Mukai et al., 2008). The differential affinities of the Ago2 and Ago3 proteins for 48 miRNAs named the "individual variants" by Azuma-Mukai et al. (2008) because of their 5′- and/or 3′-terminal differences from canonical mature miRNAs, which were associated by Azuma-Mukai et al. (2008) with (i) alternative maturation (Azuma-Mukai et al., 2008) or (ii) post-maturation processing (Azuma-Mukai et al., 2008), as independently measured *in vivo* (Azuma-Mukai et al., 2008) and as estimated *in silico* [Equation (6)] and expressed in natural logarithms are statistically significantly correlated (**Table 4**). Dashed curves and solid lines as in the legend to **Figure 3**.

processing (Azuma-Mukai et al., 2008). The estimated value was statistically significant ($r = 0.49$, $\alpha < 0.001$) for the difference between the affinity of the miRNAs for Ago2 and that for Ago3:

$$[\text{miRNA/Ago2}] - [\text{miRNA/Ago3}] = 1.04[\text{RHHK}]_{\text{F3}}\left\{\xi_j\right\} - 1.14. \quad (6)$$

This is consistent with the commonly accepted view that an individual miRNA variant forms complexes with Ago2 and Ago3 depending on its affinity for each of them, because specific interactions that normally occur due to evolutionary selection for affinity for these proteins are not there.

AN ILL-POSED INVERSE PROBLEM SOLUTION

The values of the abundances of 96 mature miRNAs in an extract from human embryonic kidney cells, HEK293T, under normal conditions (A) and following preincubation for 8 h with the transcription inhibitor actinomycin D (Bail et al., 2010) (B) are on the *y*-axis in **Figure 8**. Let us see whether these values can be predicted using Equation (5) with miRNA nucleotide sequences known from the miRBase database (Kozomara and Griffiths-Jones, 2010).

On the one hand, under the normal experimental conditions in (Bail et al., 2010), a total amount of a certain miRNA was measured so that the experimental value [miRNA] should be described by the linear-additive approximation as follows:

$$\begin{aligned}[\text{miRNA}]^{(\#)}\left\{\xi_j\right\} = {} & \beta_1^{(\#)}\left([\text{Ago1}]\right)[\text{miRNA/Ago1}]^{(\#)}\left\{\xi_j\right\} \\ & + \beta_2^{(\#)}\left([\text{Ago2}]\right)[\text{miRNA/Ago2}]^{(\#)}\left\{\xi_j\right\} \\ & + \beta_3^{(\#)}\left([\text{Ago3}]\right)[\text{miRNA/Ago3}]^{(\#)}\left\{\xi_j\right\} \\ & + \beta_4^{(\#)}\left([\text{Ago4}]\right)[\text{miRNA/Ago4}]^{(\#)}\left\{\xi_j\right\} + \varepsilon; \quad (7)\end{aligned}$$

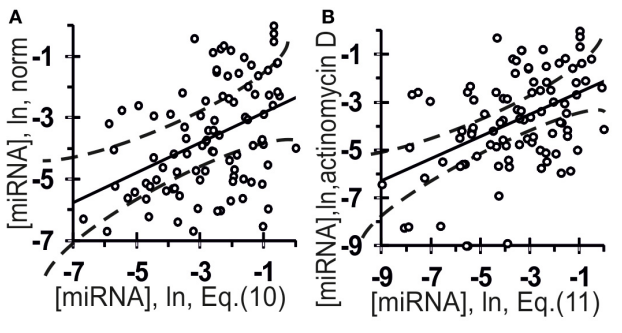

FIGURE 8 | Verification of ill-posed inverse problem solutions (Equation (10) and (11), the *x*-axis) using the final Equation (5) and independent experimental data (the *y*-axis) taken from another data source (Bail et al., 2010). The abundance of 96 mature miRNAs in an extract from the human embryonic kidney cell line HEK293T: **(A)** norm; **(B)** preincubation for 8 h with the transcription inhibitor actinomycin D (Bail et al., 2010). Independent experimental control data (*y*-axes) and *in silico* estimates and expressed on the same measurement scale [Equation (10) and Equation (11), respectively; *x*-axes] are statistically significantly correlated (**Table 4**). Dashed curves and solid lines as in the legend to **Figure 3**.

where: $\beta_1^{(\#)}$, $\beta_2^{(\#)}$, $\beta_3^{(\#)}$, and $\beta_4^{(\#)}$ represent occupancies of the corresponding Ago1, Ago2, Ago3, and Ago4 proteins given an equilibrium of the miRNA and Ago molecule turnover in normal (#) HEK293T cells; ε is the prediction error, which inevitably creeps in due insufficient experimental data.

On the other hand, only two variables, $[\text{miRNA/Ago2}]^{(\#)}$ and $[\text{miRNA/Ago3}]^{(\#)}$, out of 8 can be estimated by Equation (5) if the experimental conditions (&) are different (Azuma-Mukai et al., 2008), $[\text{miRNA/Ago2}]^{(\&)}$ and $[\text{miRNA/Ago3}]^{(\&)}$. In addition, Ago1 and Ago2, but not Ago3, are the major Ago proteins in human, and the expression of Ago3 and Ago4 is low (Valdmanis et al., 2012). Therefore, it seems quite difficult, or probably impossible, to estimate the total miRNA amount from $[\text{RHHK}]_{\text{F3}}\{\xi_j\}$ and $[\text{YRHB}]_{\text{F4}}\{\xi_j\}$ since too many ambiguities exist and too small contribution of [miRNA/Ago3] to the total amount of the miRNA is logically expected for Equation (7). In this sense, Equation (7) is an "ill-posed inverse problem."

We have recently proposed a solution to an ill-posed inverse problem (Mironova et al., 2013) using STATISTICA (Afifi et al., 2003) and considering the existing additional information given in the frames. In our case, this additional information is represented by two results of the experiments performed by Winter and Diederichs (2011b) and by Martinez and Gregory (2013) suggesting that miRNAs and Ago2 are likely to stabilize each other and that Ago2 is one of two major Ago proteins in human (Valdmanis et al., 2012). This representation substantiates the use of STATISTICA (Afifi et al., 2003) as a means of assessing the statistical significance of the linear-additive contribution of $[\text{miRNA/Ago2}]^{(\#)}$ estimates using Equation (5) in the linear-additive approximation by Equation (7) for experimental values $[\text{miRNA}]^{(\#)}$ as follows:

$$[\text{miRNA}]^{(\#)}\left\{\xi_j\right\} = \gamma^{(\&)\to(\#)}[\text{miRNA/Ago2}]^{(\&)\to(\#)}\left\{\xi_j\right\} + \delta^{(\&)\to(\#)}; \quad (8)$$

where: $\gamma^{(\&)\to(\#)}$ and $\delta^{(\&)\to(\#)}$ are the numerical values of the coefficients of dimension (constriction/compression and shifting, respectively) required for setting up a correspondence between the ranges of experimental variables found by Azuma-Mukai et al. (2008), (&), and by Bail et al. (2010) for normal HEK293T cells (#) without any optimization; $[\text{miRNA/Ago2}]^{(\&)\to(\#)}\{\xi_j\}$ is a heuristic estimate of an unknown $[\text{miRNA/Ago2}]^{(\#)}\{\xi_j\}$ value using the $[\text{miRNA/Ago2}]^{(\&)}\{\xi_j\}$ estimate and the final Equation (5).

The following formula was used as a heuristic estimate of $[\text{miRNA/Ago2}]^{(\&)\to(\#)}\{\xi_j\}$:

$$\begin{aligned}[\text{miRNA/Ago2}]^{(\&)\to(\#)}\left\{\xi_j\right\} = {} & [\text{miRNA/Ago2}]^{(\&)}\left\{\xi_j\right\}/3 \\ & -[\text{miRNA/Ago3}]^{(\&)}\left\{\xi_j\right\}; \quad (9)\end{aligned}$$

where: 1/3 is the heuristic coefficient that takes into account the normalization of experimental measurements (Azuma-Mukai et al., 2008) for Ago2 only in miRNA/Ago2 complexes within RISC without reference to Ago2 involvement in the regulation of transcription initiation or miRNA

biogenesis;—[miRNA/Ago3]$^{(\&)}\{\xi_j\}$ is a heuristic correction, which takes into account a negative effect of the competition between Ago2 and Ago3 for miRNA binding and reduces [miRNA/Ago2]$^{(\&)\rightarrow(\#)}\{\xi_j\}$ in the measurements taken without Ago3 (&).

After all intermediate calculations, the final Equation (7) assumed the following form:

$$[\text{miRNA}]\left\{\xi_j\right\} = -2.74 + 1.32[\text{RHHK}]_{F3}\left\{\xi_j\right\} - 1.71[\text{YRHB}]_{F4}\left\{\xi_j\right\}; \quad (10)$$

where: −2.74, 1.32, and −1.71 are the numerical values of the regression coefficients in Equation (7) via Equation (5).

The estimates obtained using Equation (10) were statistically significantly correlated with the measured [miRNA] values in normal HEK293T cells (**Figure 8A**, **Table 4**: $r = 0.42$, $\alpha < 0.000025$; $R = 0.43$, $\alpha < 0.000025$; and $\tau = 0.30$ at $\alpha < 0.000025$).

Nevertheless, there is an absolutely required additional stage in addressing an ill-posed inverse problem (Mironova et al., 2013), namely, verification using independent experimental data. To include this stage, we additionally reproduced all the calculations for experimental data under conditions that included ($) preincubation of HEK293T cells for 8 h with the transcription inhibitor actinomycin D (Bail et al., 2010). Because actinomycin D inhibits transcription elongation, the main difference between these conditions ($) and the normal conditions (#) is that no primary pri-miRNA transcripts are present and, consequently, Ago2-mediated miRNA biogenesis does not go. That is why we used 1/2 instead of 1/3 in Equation (9). After all intermediate calculations, the final Equation (10) derived from Equation (5) assumed the following form:

$$[\text{miRNA}]\left\{\xi_j\right\} = -4.56 + 2.20[\text{RHHK}]_{F3}\left\{\xi_j\right\} - 1.90[\text{YRHB}]_{F4}\left\{\xi_j\right\}. \quad (11)$$

The estimates obtained using Equation (11) were statistically significantly correlated with the measured [miRNA] values at the HEK293T cells preincubated for 8 h with the transcription inhibitor actinomycin D (Bail et al., 2010), as shown in **Figure 8B** and **Table 4**: $r = 0.46$, $\alpha < 0.000005$; $R = 0.43$, $\alpha < 0.000025$; and $\tau = 0.30$ at $\alpha < 0.000025$). Nevertheless, despite the statistical significance in the correlations between the value defined by Equation (10) and (11) and miRNA abundance, a large part of data points in **Figure 8** exist outside of the dashed lines. This implies that Equations (10) and (11) are adequate sources of rough estimates of miRNA abundances in human embryonic kidney cells, HEK293T, under proper experimental conditions consistent with Ago2 protein affinity for miRNAs; however there must be the Ago1 protein, which is another major Ago protein in man (Valdmanis et al., 2012) and which was ignored by Equations (10) and (11) due to lack of experimental data [miRNA/Ago1].

Collectively, all these results imply that Equation (5) produces adequate estimates ([miRNA]) for miRNA abundance in the given human cell line with an account of the biochemical features of the method used for experimental measurements [Equations (10) and (11) as examples of an ill-posed inverse problem solution].

Thus, Equation (5) found *in silico* in one experiment (Azuma-Mukai et al., 2008) readily applies to the next (Bail et al., 2010), at least within the limits of applicability of the theory that underlies these experiments. We had previously demonstrated this possibility (Ponomarenko et al., 2001), and its value is that it allows previously found patterns to be used for planning conditions of future experiments (for example, see Savinkova et al., 2013).

OUR HYPOTHESIS ON miRNA ABUNDANCE IN THE HUMAN BRAIN

Thus, all the different types of correlation shown in **Figures 3–8** fit each other like pieces of a puzzle, which allowed us to heuristically generalize all of them and state that the miRNA abundance in the human brain regions or neocortical areas may be roughly described by the function of YRHB and RHHK abundances in these miRNAs for their practical consideration by cancer and neurodegeneration researchers.

Let us check this hypothesis.

AN INDEPENDENT CONTROL TEST

Figure 9 presents the results of a cluster analysis performed using STATISTICA (Afifi et al., 2003) on the data on miRNA abundance in the human brain taken from the Sestan Brain Atlases (Kang et al., 2011), the *y*-axis, vs. *in silico* estimates obtained using [Equation (5)] within the framework of the roughest

Table 4 | Nine correlations in human (Figures 7, 8) between the miRNA abundance estimated by Equation (5) and those measured experimentally.

Experimental dataset	Linear correlation		Spearmen's rank correlation		Kendall's rank correlation	
(Reference) conditions	Coefficient *r*	Significance α	Coefficient *R*	Significance α	Coefficient τ	Significance α
Azuma-Mukai et al., 2008						
Individual variants	0.49	<0.001	0.30	<0.05	0.22	<0.05
Bail et al., 2010						
HEK293T, norm	0.42	<0.000025	0.43	<0.00005	0.30	<0.00025
HEK293T, actinomycin D	0.46	<0.000005	0.43	<0.00005	0.30	<0.00025

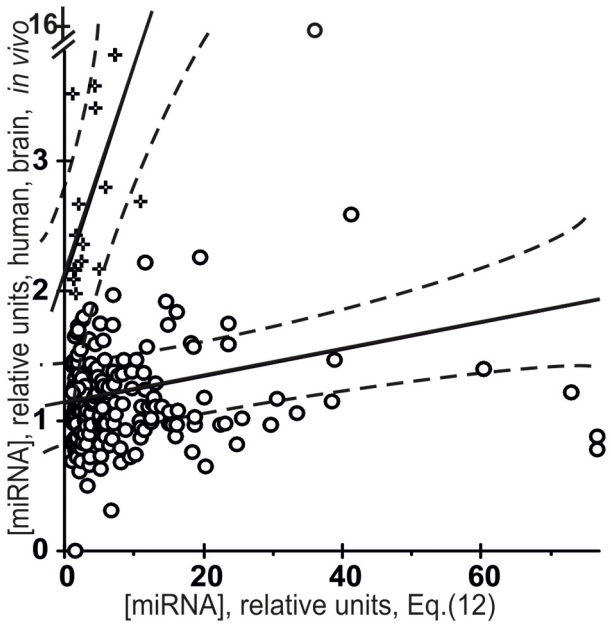

FIGURE 9 | Verification of the "limiting stage" approximation [Equation (12), the *x*-axis] using the final Equation (5) and independent experimental data (the *y*-axis) taken from the Sestan Brain Atlases (Kang et al., 2011). Results of a cluster analysis performed using STATISTICA (Afifi et al., 2003) on the miRNA abundance in the human brain taken from the Sestan Brain Atlases (Kang et al., 2011) vs. *in silico* estimates [Equation (12)] within the framework of the "limiting stage" approximation. The major cluster (◯) includes 294 miRNAs (92%), which have the lowest mean miRNA abundance *in vivo*, [miRNA] = 1.2 ± 1.0, and a high mean [miRNA/Ago] affinity estimated [Equation (12)], 9.1 ± 12.7, and the lowest statistically significant correlation coefficient $r = 0.14$ ($\alpha < 0.025$). The minor cluster (+) includes 24 miRNAs (8%) with the highest mean miRNA abundance *in vivo*, 2.7 ± 0.7, and a lowest mean estimate of [miRNA/Ago] affinity *in silico* [Equation (12)], 3.5 ± 2.8, and the highest correlation coefficient $r = 0.66$ ($\alpha < 0.001$). Dashed curves and solid lines as in the legend to **Figure 3**.

approximation possible, the so-called "limiting stage" approximation (the *x*-axis):

$$[\text{miRNA}]\{\xi_j\} = \exp\left[\text{MIN}\left(0.52[\text{RHHK}]_{F3}\{\xi_j\} + 1.35[\text{YRHB}]_{F4}\{\xi_j\};\ 1.14 - 0.52[\text{RHHK}]_{F3}\{\xi_j\} + 1.35[\text{YRHB}]_{F4}\{\xi_j\}\right)\right]. \quad (12)$$

It was applied due to lack of data on the preference of any of these miRNAs for any of the Ago1, Ago2, Ago3, or Ago4 protein in the *Argonaute* family (Gagnon and Corey, 2012) and also due to lack of data on the affinity of any miRNA for two of these four proteins (50%), Ago1 and Ago4. Moreover, not only cells specific for the central nervous system can influence the mean abundance of miRNAs in the human brain, but many more tissue-specific cells such as neurons, glia (microglia, oligodendrocytes, astrocytes, etc.), meninges (connective tissues covering the brain and containing a large number of blood vessels), cells of choroid plexus (capillaries, simple cuboidal epithelium, ependymal cells) and other can do as much. This diversity should increase the variance of [miRNA] values even more than in the case of the organ-specific abundance of miRNAs in Arabidopsis (**Figure 4**). Indeed, while the C_V-values for miRNA abundance in Arabidopsis inflorescences, stems, siliques, seedlings, roots, cauline and rosette leaves ranged from 7 to 72% (the mean being $31 \pm 19\%$), the C_V-values for miRNA abundance in 95 human brain regions or neocortical areas ranged from 9 to 281% (the mean being $73 \pm 39\%$), possibly due to very high levels of expression of unique miRNAs in a limited number of these regions or areas.

First of all, the major cluster (○) includes 294 miRNAs (92%), which have a low mean miRNA abundance *in vivo*, [miRNA] = 1.2 ± 1.0, and a high mean [miRNA/Ago] affinity estimated *in silico* [Equation (12)], 9.1 ± 12.7. This cluster comprises miRNAs that have no preference for binding to any of the four proteins in the *Argonaute* family. This result is consistent with the statement used in the derivation of Equation (5) and made by Azuma-Mukai et al. (2008): most human miRNAs have no preference for binding to any particular Ago protein. We were surprised to see that even the roughest estimates were nevertheless statistically significantly linearly correlated ($r = 0.14$, $\alpha < 0.025$) with *in vivo* measurements.

Finally, the minor cluster (+) includes 24 miRNAs (8%) with a high mean miRNA abundance *in vivo*, 2.7 ± 0.7, and a low mean estimate of [miRNA/Ago] affinity *in silico* [Equation (12)], 3.5 ± 2.8. This cluster contains miRNAs, each of which has a preference for binding to one particular Ago protein. This result is consistent with conclusions made by Azuma-Mukai et al. (2008): a few miRNAs in man have preference for binding to any particular Ago protein. As can be seen, these roughest estimates are, again, statistically significantly ($r = 0.66$, $\alpha < 0.001$) correlated with *in vivo* values. Importantly, a higher *r*-value for specific than non-specific miRNA/Ago affinity ($0.66 > 0.14$) is in agreement with the most common view of the interactions between molecules. Nevertheless, despite the statistical significance in the correlations between the value defined by Equation (12) and miRNA abundance in the human brain, a large part of data points in **Figure 9** exist outside of the dashed lines. Therefore, Equation (12) is an adequate source of only roughest estimates of miRNA abundances in the human brain; however, there is a wealth of relevant information on the Ago1 and Ago4 proteins (**Figure 8**), on the tissue-specific patterns of miRNA and Ago gene expression (**Figure 4**), which was ignored by Equation (12) due to lack of experimental data.

CONCLUDING REMARKS

We have now established that miRNA abundances depend on taxon-specific tetranucleotides in miRNAs.

First of all, specific tetranucleotides in a given miRNA seem to be responsible for the selectivity of miRNA binding to the proper Ago protein, which determines the biological function of the RISC containing this miRNA/Ago complex: (i) the RISC interacts with promoter DNAs or messenger RNAs (mRNAs) as it searches them for a complementary target of these miRNAs;

and (ii) the RISC binds to or cleave this target within the mRNAs (Gagnon and Corey, 2012).

Based on these facts, we have for the first time obtained quantitative *in silico* estimates for miRNA abundances in the human embryonic kidney cells HEK293T by roughly solving an ill-posed inverse problem, and, also, in the human brain regions or neocortical areas, which are statistically significantly correlated with data from independent experiments on measuring these values *in vivo* taken from a work by Bail et al. (2010) and from the Sestan Brain Atlases (Kang et al., 2011), respectively. These two correlations are consistent with the results of two experiments, one performed by Winter and Diederichs (2011b) and another, by Martinez and Gregory (2013), and demonstrated that the affinity of miRNAs for Ago proteins is an influence on the abundance of both miRNAs and Ago proteins due to their mutual co-stabilization in cells.

In summary, we have found evidence that *in silico* estimates like these can reach an acceptable accuracy level for their practical consideration by cancer and neurodegeneration researches once the preference of these miRNAs for the proteins in the *Argonaute* family has become known, and so have yet unknown values of the affinity of any miRNA for two of the four proteins (50%), Ago1 and Ago4, which is absolutely required for a more accurate approximation. In any case, because the abundance estimates [Equations (5)—(12)] for most miRNAs were more statistically significant in a particular human cell line (**Figure 8**) than in the human brain as a whole (**Figure 9**), the more specifically a target for estimation is defined (the entire human organism, an organ, a part, a tissue, a cell type, or a cell line), the more suitable these estimates are for practical use.

ACKNOWLEDGMENTS

This work was in part supported by grants 11-04-01254, 12-04-01584, and 12-04-33112 from RFBR; NSc-5278.2012.4 from the President of Russia, Project 8740 from the Russian Ministry of Education and Science, Integration Project #136 from the SB RAS, Integration Project 6.8 and 30.29 from the RAS Presidium, and Program 28 from the RAS.

REFERENCES

Afifi, A. A., Clark, V. A., and May, S. (2003). *Computer-Aided Multivariate Analysis*. New York, NY: CRC Press.

Axtell, M. J., and Bartel, D. P. (2005). Antiquity of microRNAs and their targets in land plants. *Plant Cell* 17, 1658–1673. doi: 10.1105/tpc.105.032185

Azuma-Mukai, A., Oguri, H., Mituyama, T., Qian, Z. R., Asai, K., Siomi, H., et al. (2008). Characterization of endogenous human Argonautes and their miRNA partners in RNA silencing. *Proc. Natl. Acad. Sci. U.S.A.* 105, 7964–7969. doi: 10.1073/pnas.0800334105

Bail, S., Swerdel, M., Liu, H., Jiao, X., Goff, L. A., Hart, R. P., et al. (2010). Differential regulation of microRNA stability. *RNA* 16, 1032–1039. doi: 10.1261/rna.1851510

Barbato, C., Ruberti, F., and Cogoni, C. (2009). Searching for MIND: microRNAs in neurodegenerative diseases. *J. Biomed. Biotechnol.* 2009, 871313. doi: 10.1155/2009/871313

Fishburn, P. (1970). *Utility Theory for Decision Making*. New York, NY: John Wiley and Sons.

Gagnon, K. T., and Corey, D. R. (2012). Argonaute and the nuclear RNAs: new pathways for RNA-mediated control of gene expression. *Nucleic Acid Ther.* 22, 3–16.

Gantier, M. P., McCoy, C. E., Rusinova, I., Saulep, D., Wang, D., Xu, D., et al. (2011). Analysis of microRNA turnover in mammalian cells following Dicer1 ablation. *Nucleic Acids Res.* 39, 5692–5703. doi: 10.1093/nar/gkr148

Havens, M. A., Reich, A. A., Duelli, D. M., and Hastings, M. L. (2012). Biogenesis of mammalian microRNAs by a non-canonical processing pathway. *Nucleic Acids Res.* 40, 4626–4640. doi: 10.1093/nar/gks026

Hayes, K. G., Perl, M. L., and Efron, B. (1989). Application of the bootstrap statistical method to the tau-decay-mode problem. *Phys. Rev. D. Part. Fields* 39, 274–279. doi: 10.1103/PhysRevD.39.274

Hwang, H. W., Wentzel, E. A., and Mendell, J. T. (2007). A hexanucleotide element directs microRNA nuclear import. *Science* 315, 97–100. doi: 10.1126/science.1136235

IUPAC-IUB (1971). Commission on biochemical nomenclature (CBN). Abbreviations and symbols for nucleic acids, polynucleotides and their constituents. *J. Mol. Biol.* 55, 299–310. doi: 10.1016/0022-2836(71)90319-6

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., et al. (2011). Spatio-temporal transcriptome of the human brain. *Nature.* 478, 483–489. doi: 10.1038/nature10523

Kozomara, A., and Griffiths-Jones, S. (2010). MiRBase: integrating microRNA annotation and deep-seencing data. *Nucleic Acids Res.* 39, D152–D157. doi: 10.1093/nar/gkq1027

Lu, C., Tej, S. S., Luo, S., Haudenschild, C. D., Meyers, B. C., and Green, P. J. (2005). Elucidation of the small RNA component of the transcriptome. *Science* 309, 1567–1569. doi: 10.1126/science.1114112

Martinez, N. J., and Gregory, R. I. (2013). Argonaute2 expression is post-transcriptionally coupled to microRNA abundance. *RNA* 19, 605–612. doi: 10.1261/rna.036434.112

Mironova, V. V., Omelyanchuk, N. A., Savina, M. S., Ponomarenko, P. M., Ponomarenko, M. P., Likhoshvai, V. A., et al. (2013). How multiple auxin responsive elements may interact in plant promoters: a reverse problem solution. *J. Bioinform. Comput. Biol.* 11, 1340011. doi: 10.1142/S0219720013400118

Omelyianchuk, N. A., Ponomarenko, P. M., and Ponomarenko, M. P. (2011). The nucleotide sequence features of the mature microRNA seem to be responsible for the affinity to human Ago2 AND Ago3 proteins. *Mol. Biol. (Mosk).* 45, 327–336. doi: 10.1134/S0026893311020130.

Ponomarenko, J. V., Furman, D. P., Frolov, A. S., Podkolodny, N. L., Orlova, G. V., Ponomarenko, M. P., et al. (2001). ACTIVITY: a database on DNA/RNA sites activity adapted to apply sequence-activity relationships from one system to another. *Nucleic Acids Res.* 29, 284–287. doi: 10.1093/nar/29.1.284

Ponomarenko, M. P., Kolchanova, A. N., and Kolchanov, N. A. (1997). Generating programs for predicting the activity of functional sites. *J. Comput. Biol.* 4, 83–90. doi: 10.1089/cmb.1997.4.83

Ponomarenko, M. P., Omelyanchuk, N. A., Katokhin, A. V., Savinskaya, S. A., and Kolchanov, N. A. (2008). The abundance of microRNA in Arabidopsis thaliana correlates with the presence of tetranucleotides WRHW and DRYD in their sequences. *Dokl. Biochem. Biophys.* 420, 150–154. doi: 10.1134/S1607672908030149

Savinkova, L., Drachkova, I., Arshinova, T., Ponomarenko, P., Ponomarenko, M., and Kolchanov, N. (2013). An experimental verification of the predicted effects of promoter TATA-box polymorphisms associated with human diseases on interactions between the TATA boxes and TATA-binding protein. *PLoS ONE* 8:e54626. doi: 10.1371/journal.pone.0054626

Sohn, I., Owzar, K., George, S. L., Kim, S., and Jung, S. H. (2009). A permutation-based multiple testing method for time-course microarray experiments. *BMC Bioinformatics* 10:336. doi: 10.1186/1471-2105-10-336

Song, J. J., Smith, S. K., Hannon, G. J., and Joshua-Tor, L. (2004). Crystal structure of Argonaute and its implications for RISC slicer activity. *Science* 305, 1434–1437. doi: 10.1126/science.1102514

Valdmanis, P. N., Gu, S., Schuermann, N., Sethupathy, P., Grimm, D., and Kay, M. A. (2012). Expression determinants of mammalian argonaute proteins in mediating gene silencing. *Nucleic Acids Res* 40, 3704–3713. doi: 10.1093/nar/gkr1274

Winter, J., and Diederichs, S. (2011a). MicroRNA biogenesis and cancer. *Methods Mol. Biol.* 676, 3–22. doi: 10.1007/978-1-60761-863-8_1

Winter, J., and Diederichs, S. (2011b). Argonaute proteins regulate microRNA stability:

Increased microRNA abundance by Argonaute proteins is due to microRNA stabilization. *RNA Biol.* 8, 1149–1157. doi: 10.4161/rna.8.6.17665

Zadeh, L. (1965). Fuzzi sets. *Inform. Control* 8, 338–353. doi: 10.1016/S0019-9958(65)90241-X

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 August 2012; accepted: 06 June 2013; published online: 01 July 2013.

Citation: Ponomarenko MP, Suslov VV, Ponomarenko PM, Gunbin KV, Stepanenko IL, Vishnevsky OV and Kolchanov NA (2013) Abundances of microRNAs in human cells can be estimated as a function of the abundances of YRHB and RHHK tetranucleotides in these microRNAs as an ill-posed inverse problem solution. Front. Genet. ***4****:122. doi: 10.3389/fgene.2013.00122*

This article was submitted to Frontiers in Non-Coding RNA, a specialty of Frontiers in Genetics.