9956–9964 *Nucleic Acids Research, 2016, Vol. 44, No. 20* *Published online 19 September 2016*
doi: 10.1093/nar/gkw793

Structural dynamics control the MicroRNA maturation pathway

Paul Dallaire[1], Huiping Tan[2], Keith Szulwach[2], Christopher Ma[2], Peng Jin[2] and François Major[1,*]

[1]Institute for Research in Immunology and Cancer, and Department of Computer Science and Operations Research, Université de Montréal, PO Box 6128, Downtown Station, Montréal, Québec H3C 3J7, Canada and [2]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA

Received June 19, 2015; Revised August 26, 2016; Accepted August 29, 2016

ABSTRACT

MicroRNAs (miRNAs) are crucial gene expression regulators and first-order suspects in the development and progression of many diseases. Comparative analysis of cancer cell expression data highlights many deregulated miRNAs. Low expression of miR-125a was related to poor breast cancer prognosis. Interestingly, a single nucleotide polymorphism (SNP) in miR-125a was located within a minor allele expressed by breast cancer patients. The SNP is not predicted to affect the ground state structure of the primary transcript or precursor, but neither the precursor nor mature product is detected by RT-qPCR. How this SNP modulates the maturation of miR-125a is poorly understood. Here, building upon a model of RNA dynamics derived from nuclear magnetic resonance studies, we developed a quantitative model enabling the visualization and comparison of networks of transient structures. We observed a high correlation between the distances between networks of variants with that of their respective wild types and their relative degrees of maturation to the latter, suggesting an important role of transient structures in miRNA homeostasis. We classified the human miRNAs according to pairwise distances between their networks of transient structures.

INTRODUCTION

The secondary (2D) structure of a RNA sequence is defined by the Watson–Crick interactions forming between complementary bases within itself, A•U and G•C (1). However, functional RNAs in cells are rarely fully complementary, giving rise to single-stranded dynamic regions that ease conformational changes inherent to RNA function (2,3). Recently, transitions between low-abundance excited structural states (ESs) were visualized by nuclear magnetic resonance (NMR) relaxation dispersion (4). RNA dynamics were linked to function by showing that sequence mutations affecting the relative abundance of the ESs result in modulations of function (4).

The recent NMR studies were made on rather short RNAs of lengths less than 40 nucleotides. The number of ESs increases with RNA size (4), and it is uncertain whether NMR studies will soon be applicable to larger RNAs. On the other hand, computational methods are free from the limitations of experimental techniques. For instance, RNA 2D structure prediction programs are known to generate and evaluate very large numbers of alternative conformations (5–8). RNA structures have been studied computationally, providing new insights and steering further functional investigations (9). Interestingly, the ESs observed by NMR are also predicted computationally, and there is a high correlation between computational and experimental energy changes between ground states (GSs) and ESs (4). This correlation was observed in sets of suboptimal conformations that include non-canonical base pairs (10), which were found to be the main differences between the ESs of the studied sequences (4).

These results from NMR studies prompted us to formalize the impact of RNA sequence changes on function in terms of the ground and transient states. Consider the transition network of the 12 most stable structures of the *Escherichia coli* 16S ribosomal RNA A-site (Figure 1A). In this network, a pair of structures is linked if it is connected by a single basic transition, either a base pair formation or a bulge migration (Figure 2A). The A-site structure bound to a mRNA::tRNA complex (11), labeled BS, is crucial (12), but the NMR conditions without the ribosomal context favored the path through the bulge migration and toward the ES (4). The crystallographic structure of the *E. coli* ribosome bound to a mRNA::tRNA complex induces the base pair transition towards the BS. Interestingly, the GS, which corresponds to the A-site conformation of the ribosome unbound from the complex, is the most stable state observed, by NMR and X-ray crystallography, as well as computa-

[*]To whom correspondence should be addressed. Tel: +1 514 343 6752; Fax: +1 514 343 5839; Email: Francois.Major@UMontreal.CA

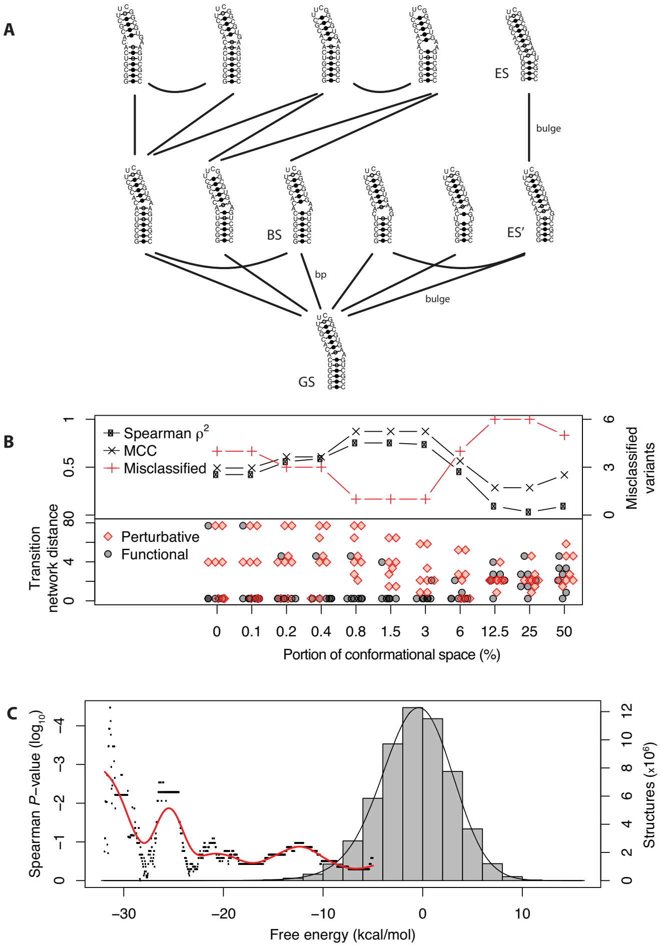

Figure 1. Transition network. (**A**) The 12 most stable predicted structures of the *Escherichia coli*; 16S ribosomal RNA A-site include the ground (unbound) state (GS), the mRNA::tRNA bound state (BS) and the excited state (ES) observed by NMR relaxation dispersion. ES′ is a necessary transition state between the GS and the ES. The GS and BS are linked by a base pair formation transition (bp); a A∘A bp in the GS is lost in the BS. The GS and ES′, as well as the ES′ and ES are linked by single bulge migration transitions (bulge); from 5′ to 3′, A bulge in the GS, then G in the ES′ and then U in the ES. Canonical Watson–Crick base pairs are indicated by black dots; Wobble base pairs by grey dots; and, non-canonical base pairs by white dots. (**B**) Top: The likelihood to predict the functionality of 15 A-site sequence variants using increasing portions of their conformational space, expressed in square Spearman correlation (ρ^2) (two-sided configuration), Matthews correlation coefficient (MCC) and number of misclassified variants. Bottom: Separation between functional (grey circles) and perturbative (red diamonds) sequence variants at different portions of the conformational space considered. (**C**) Spearman correlation p-values as the portion of conformational space considered increases (black dots). The red line is a smoothing of the *P*-values. The histogram indicates the number of structures at different energy levels.

tionally. It is noteworthy that the computed network in Figure 1A includes the structures visualized in different experimental conditions.

The basic transitions that define the network are fully determined by the set of structures that constitute it, and thus this set of structures contains the same information than does the network. Here, we considered miR-125a and miR-125a-G22, two RNA sequences that induce different sets of structures and networks. A direct comparison of these two sets provides information about their relative dynamics. For such comparisons, we developed a metric that returns the distance between two sets of structures. We evaluated if this distance can provide any information about the relative processing efficiencies of miRNA variants.

MATERIALS AND METHODS

Distance and similarity metrics

The metrics calculate the distance and degree of similarity between two RNA sequences by comparing their sets of predicted 2D structures. To accomplish this, we: (i) predict the sets of stable 2D structures, including non-canonical base pairs, using the MC-Fold software (10); (ii) compute a statistical description of the base pairing patterns of each nucleotides from these sets (signature); and (iii) compute a similarity value between pairs of signatures. In some instances we were required to compute the local minima from the MC-Fold predicted sets of structures. These were obtained by brute force calculation (see below), in $O(Ns^2)$, where N is the sequence length for s enumerated structures. A fast algorithm for this problem has been reported (13).

We predict either fixed numbers of structures (typically a few thousands), or all structures below an energy threshold expressed as a percentage of the minimum free energy. We build for each sequence a signature composed of seven values for each nucleotide, two for the paired (i.e. either the (i) 5′ or (ii) 3′ partner) and five for the unpaired conformations. When a nucleotide is involved in a base pairing conformation, it can interact with different partners, yielding N-1 counts of base pairing events for a sequence of length N and we compute a statistic that summarizes these base-pairing counts. The unpaired nucleotides can occur in various structural contexts, and we distinguish between: (iii) terminal loop; (iv) loop in the 5′ arm of a stem; (v) loop in the 3′ arm of a stem; (vi) multi-loop junction; or, (vii) dangling at either end of the structure. The seven values for a nucleotide were normalized to frequencies between 0 and 1, so that the signature is independent of sequence length. We defined the single strandedness of a nucleotide as the sum of its five unpaired values.

Consider the values from the signatures of miR-125a (major allele labeled GC) and miR-125a-G22U (minor allele labeled UC). Nucleotide 16 has values $GC_{16} = [0.881, 0, 0.118, 0, 5e\text{-}5, 0, 0]$ and $UC_{16} = [0.917, 0, 0.082, 0, 8e\text{-}4, 0, 0]$. It is stabilized in UC and other variants (see the **b** region in Figure 4F), indicated by increasing values in the first entry of the vector (5′ base-pairing partner; $0.917 > 0.881$) and decreasing values of the unpaired entries (e.g. $0.082 < 0.118$). Nucleotide 72 has values $GC_{72} = [0, 0.82, 0, 0.18, 4e\text{-}5, 0, 0]$ and $UC_{72} = [3e\text{-}6, 0.75, 0, 0.25, 5e\text{-}5, 0, 0]$. It is destabilized in UC and other variants (see the **h** region in Figure 4F), indicated by decreasing paired values in the second vector entry (3′ base-pairing partner; $0.75 < 0.82$) and increasing unpaired values (e.g. $0.25 > 0.18$).

Measuring the distance between two positions in these signatures can be done in several non-equivalent ways, and we chose to use the sum of the absolute differences in corresponding conformations. The result of comparing nucleotides represented as normalized counts in vectors **v** and

w is thus simply:

$$\sum_{i\,=\,1\ldots7} |\mathbf{v}_i - \mathbf{w}_i|$$

and gives a distance score. $|GC_{16} - UC_{16}|$ is [0.036, 0, 0.036, 0, 7.5e-4, 0, 0] whose sum is 0.073 and $|GC_{72} - UC_{72}|$ is [3e-6, 0.07, 0.18, 0.25, 5e-5, 0, 0] whose sum is 0.50. These values are indicative of a larger distance in the latter than the former nucleotide. To compare two signatures of equal length, we sum the results of these sums at every position. For instance, GC and UC sequences are 3.7 apart on this scale. Alternatively, we used a similarity score:

$$\sum_{i\,=\,1\ldots7} \mathrm{MIN}\,(\mathbf{v}_i\,, \mathbf{w}_i)$$

in a dynamic programming routine to compare signatures obtained from unequal lengths using an algorithm that is similar to sequence alignment. Here, MIN(GC_{16}, UC_{16}) is [0.0.881, 0, 0.082, 0, 5e-5, 0, 0] whose sum is 0.96 and MIN(GC_{72}, UC_{72}) is [0, 0.75, 0, 4e-5, 0, 0, 0] whose sum is 0.75. These values are indicative of a larger similarity in the former than the latter nucleotide. The similarity between GC and UC sequences is 84.14. The largest possible similarity for these sequences is 86, i.e. the number of nucleotides. This yields the best possible nucleotide pairwise assignments and optimal score. We used the affine gap model, where a gap has two distinct penalty costs, gap initiation (−4.5) and gap elongation (−2.5), that are added (negative values) to the score, i.e. removed from the overall similarity. This scheme allows for the selection of solutions with minimum numbers of gaps.

In the case of highly different sequence lengths (not used in this study), we developed a version where a target length can be specified. Our computational scheme accepts a number of variations in the similarity score calculation, some of which have been found useful previously to compare complete base pairing partition tables for computing guide trees for RNA structure alignments (14), and to detect changes in the partition table induced by sequence variation (15). These variations and relative pros and cons were recently discussed (16).

To a certain extent, our calculations aim at comparing truncated partition tables, in the same spirit as proposed by Lorenz and Clote (17). However, there is no software that computes the necessary partition table for MC-Fold as of yet. The use of MC-Fold was required because it was shown to predict correctly sparsely populated transient structural states that were visualized by NMR. This was possible in particular because of the MC-Fold's ability to predict the set of non-canonical base pairs (10). To obtain large numbers of structures, we developed a very fast version of the MC-Fold program using dynamic programming. The algorithm of the new version is similar to that developed by Wuchty *et al* (18).

Transition networks and local minima

A 2D RNA structure is represented as a vector r, where each nucleotide is assigned a number according to its pairing partner. If a nucleotide, x, is unpaired, then r[x] = −1. If nucleotide i is paired with nucleotide j, then r[i] = j and r[j] = i. A base pair breakage or formation implies that exactly two values in r are altered. Conversely, two vectors r and s that differ at exactly two indices are linked by a single base pair change. If a bulge migration occurs (Figure 2A), then three values in r are changed. However, events other than bulge migration, for instance a loop migration or stem jump also cause three values to change in r and their detection requires verification of positional invariants: two vectors differ solely by a bulge migration if and only if the values at three indices are changed and one pair of indices involves adjacent nucleotides interchanging their values. The detection of such differences between two structures from a set is computed in time proportional to the length of the vectors, O(n). We incrementally compute groups of structures from a set by comparing all its member pairs so that all members in a group are connected by a series of single changes, and such a suite of changes connects no pair of structures taken from different groups. In each group, the most energetically stable structure is a local minimum (Figure 2B).

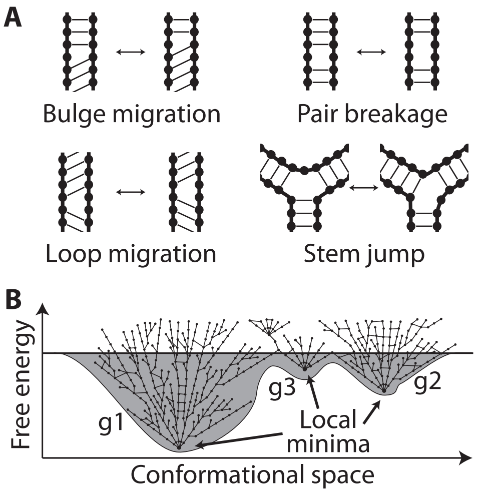

Figure 2. Simple transformations and local minima. (**A**) Simple transformations considered in the building of transition networks. Like for the bulge migration, three values change in the vector representation of two 2D structures connected by a loop migration or a stem jump. They differ from the bulge migration by the indices they modify in the vector representation. (**B**) Using only structures below a fixed energy threshold (grey area) form groups (g1, g2 and g3). The most stable structure in each group defines a local minimum.

DNA plasmids

To generate pSM2-miR-125a-G, sequence corresponding to the miR-125a precursor and 125 bp flanking region on each side was amplified from normal human genomic DNA and inserted into the pSM2 vector (Open Biosystems). Plasmid pSM2-miR-125a-U was generated by introducing a G-T substitution at the sequence corresponding to the eighth position of the mature miR-125a (Stratagene, La Jolla, CA, USA). A double mutant (pSM2-miR-125a-U-M) contained the G-T polymorphism in addition to a complementary C-A base change, reported previously (19) and other mutants were prepared in the same manner.

Cell culture, transfection and luciferase assay

Human HEK293 cells were grown in Dulbecco's modified Eagle's medium (GIBCO) supplemented with 10% fetal bovine serum and penicillin/streptomycin. Plasmids were transfected into cells by Lipofectamine 2000 (Invitrogen) following the manufacturer's protocol. Equal amounts of pSM2-shRNA against EGFP were co-transfected with each pSM2-miR-125a construct, and each sample was transfected in triplicate.

Quantitative RT–PCR

Total RNA from cultured cells was isolated with Trizol (Invitrogen). RNA samples were reverse-transcribed into cDNA with the ThermoScript First-Strand Synthesis System (Invitrogen). Real-time PCR was performed with gene-specific primers and Power SYBR Green PCR Master Mix using the 7500 Standard Real-Time PCR System (Applied Biosystems). The primers were:

- 5′-AATGTCTCTGTGCCTATCTCCATCT-3 ’(F1)
- 5′-GTCCCTGAGACCCTTTAACC-3′ (F2)
- 5′-AACCTCACCTGTGACCCTG-3′ (R)

These primers were used for the detection of pri- and pre-miR-125a as described previously (19). F1 and R are for pri-miRNA. Since the primer for pre-miRNA (F2) also recognizes the corresponding pri-miRNA, F2 and R therefore measure both forms.

Relative quantities of mature miRNA were detected using Ncode miRNA detection kit (Invitrogen), with primers specific to each allele. Each allele was detected with a primer complement to full-length mature miRNA with variants specific to each allele, and a Universal qPCR Primer provided with the kit. The amplification efficiency of each pair of primers was determined using the chemically synthesized mature miR-125a and used for the quantification of mature miRNAs. The siEGFP produced from the co-transfected pSM2-shRNA-EGFP was also detected by Ncode miRNA detection kit and used as endogenous control in ddCt relative quantification. siEGFP was used for normalization in the calculation of maturation efficiency in the subsequent analyses.

RESULTS

Transition networks

An interesting question is how many structures must be considered to get an accurate comparison of transition networks? In the absence of *a priori* knowledge, we used increasing portions of the conformational space of 15 A-site sequence variants, some of which are known to perturb translation, by either affecting binding to the translation initiation factor or promoting read-through and frameshifting (20). Figure 1B shows the separation of the A-site variants as a function of conformational portion. As we see, the curves reach a plateau where all but one sequence variant is misclassified. Interestingly, the left and right sides of the plateau indicate, respectively, smaller and larger portions of the conformational space where more variants are misclassified. This indicates that the structures defined within 1% of the conformational space contain the information needed to classify properly the variants, while taking too many structures introduces noise. Figure 1C shows the exploration of a much larger partition of the conformational space. Again, we see that small sets of stable structures contain more information about the function of the variants (small Spearman *P*-values), than larger sets of structures that include less and less stable structures (higher Spearman *P*-values). Increasing the considered portion of the conformational space incurs noise and reduces the signal.

Maturation efficiency

MiRNAs are transcribed from genomic DNA in primary transcripts (pri-miRNA), which are recognized and transformed by many proteins and complexes before they result in mature products of about 22 nucleotides (21–23). Each one of the miRNA protein partners requires a suitable and possibly induced-fit structure that exhibits various structural features for both specific recognition and processing. Among the first miRNA processing steps, the pri-miRNA is recognized and cleaved to a precursor form (pre-miRNA) by the Drosha/DGCR8 microprocessor complex. Then, EXPORTIN-5 exports pre-miRNAs to the cytoplasm (24,25), where the RNA-Induced Silencing Complex, which includes DICER and an Argonaute protein, cleaves their loops and loads the mature products (26–29). Other proteins have been shown to be involved in the maturation of specific miRNAs, such as LIN28 that mediates the terminal uridylation and blockage of let-7 (30,31), SMAD that promotes miRNA maturation by Drosha (32) and hnRNP A17 that is required for the processing of miR-18a (33,34). The complete maturation process has not yet been fully characterized, and additional steps and factors may yet be discovered.

It was recently shown that lower miR-125a expression in breast cancer patients is indicative of shorter survival rates (35). Interestingly, a single-nucleotide polymorphism (SNP) found in a minor allele expressed by breast cancer patients was shown to considerably affect the maturation of miR-125a and thus abundance of mature products (36). This prompted us to investigate the ability of a cell to process a primary transcript into a mature miRNA. We defined the maturation efficiency of this process as the instantaneous molar ratio of mature product to the corresponding primary transcript abundance. The SNP was found in the seed region of the mature miRNA, and was initially thought to hinder the miRNA interaction with its target mRNAs. Surprisingly, it was shown to block Drosha processing instead (19). The major and minor alleles of pri-miR-125a fold into the same GS. In the minor allele, a G•C base pair is substituted by an isosteric non-canonical U•C base pair that slightly affects the energy and stability of the GS (Figure 3A). The loss in maturation is not due to the loss of a canonical base pair, because a substitution into the A•U base pair also leads to weak maturation efficiency (Supplementary Table S1; see also Figure 4A bottom).

Recently, the structural requirements for DROSHA processing of a few pri-miRNAs were shown to be subject to

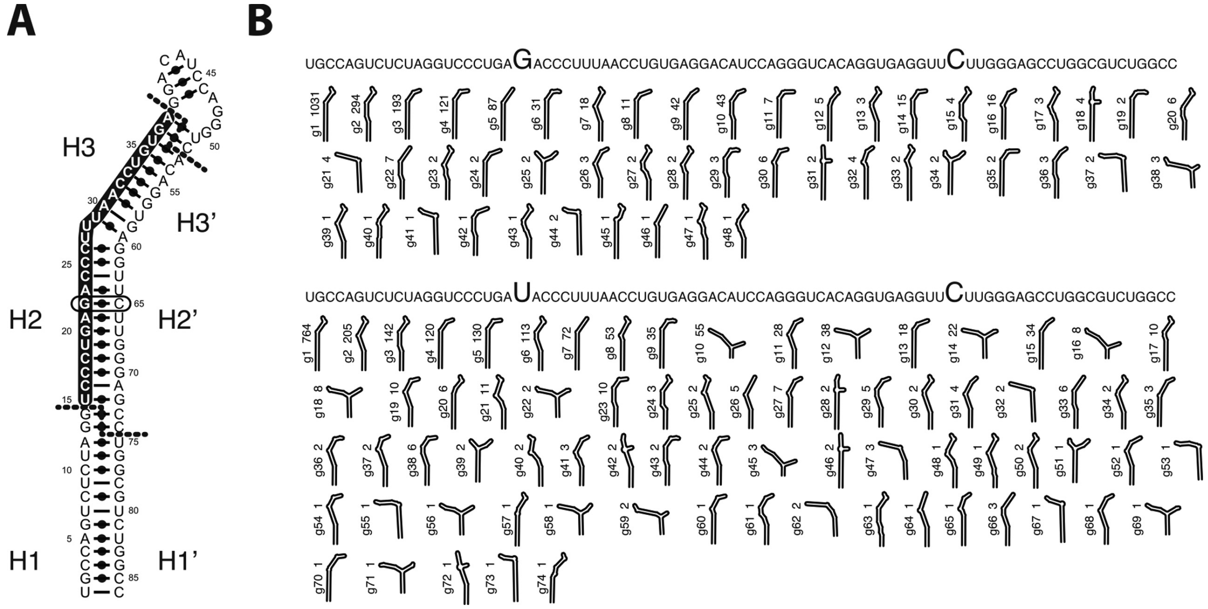

Figure 3. Energy minima of miR-125a major and minor allele hairpins. (**A**) Minimum free energy structure (global minimum) of the human miR-125a allele hairpins. The base pair that contains the position of the G22U SNP is encircled. The mature miRNA sequence is highlighted. The Drosha (lower dotted line) and DICER (upper dotted lines) cuts are shown. (**B**) The 2000 most stable predicted structures of the major allele define 48 energy minima (top), while those of the minor allele define 74 (bottom). Each basin is defined by its local minimum 2D structure, energy rank (g1, g2, ...) and number of conformations in it.

the flexibility of a specific region (37). We asked whether we could link miRNA processing to transition networks and use the metrics described above to predict variations in overall maturation efficiencies. We computed the first 2000 most stable structures of the major and minor alleles of miR-125a. A comparison of the transition networks defined by these sets of structures indicated huge differences, both in the size and energy of local minima (Figure 3B). Two of the most striking differences between the major and minor alleles are the change in the numbers of minima, from 48 to 74, and reduced number of structures connected to the GS, from 1031 to 764. We also noted that the seventh minimum of the major allele, g7 (Figure 3B top), ranks 2nd among the minima of the minor allele, g2 (Figure 3B bottom), with an important increase in the number of structures connected to it, from 18 to 205.

To determine if we can use transition network comparison to predict maturation efficiencies, we engineered the two miR-125a alleles, as well as the 14 other variants to address all possible base pairs at the position of the SNP. We expressed each variant in cultured cells, and measured their maturation levels by qRT-PCR (Supplementary Table S1; see also Figure 4A bottom). We then compared the maturation data with the distance between transition networks for various portions of their conformational space (Figure 4B). The maturation levels were lower for each variant when compared to the major allele. We observed a strong link between maturation levels and distances of transition networks, revealed by the high correlation between the two factors ($\rho = -0.75$). We used pairwise distances of transition networks to classify the variants of similar maturation levels (Figure 4A). Except for variants GG and GU, all variants whose processing is proficient (>20% compared to the major allele) share a cluster that includes the major allele. The minor allele, which is one of the least processed, shares a cluster with other lowly expressed variants that are located far from the major allele.

We applied the same hierarchical clustering approach to other data sets used in previous studies of miRNA processing by the Drosha/DGCR8 microprocessor. The results corroborated with those of the miR-125a variants, i.e. we observed a strong correlation between distances of transition networks and maturation efficiency: miR-30a (23) (Figure 4C) and miR-21 (38) (Figure 4D). In the dendrogram of miR-30a, the clustering result shows the two unprocessed variants (labeled M4 and M5) clearly set apart from the processed variants. The seven variants of miR-21 were cloned, expressed in HEK293 cells and their processing measured by Northern blotting. The normalized distances of the miR-21 variants' transition networks to the major allele were computed using the precursor sequence, flanked by the complete expressed primary transcript of 312 nucleotides. The dendrogram clearly separated the variants into two clusters, one of which contains the two lowly processed variants (labeled G1 and GGU), and the other contains the highly processed variants (all but G2).

The strong link we observe between distances of miRNA transition networks and maturation efficiencies suggests that: (i) multiple transient structures are involved in miRNA maturation; (ii) these transient structures are encoded in miRNA primary transcripts; and, (iii) the transitions between these structures determine miRNA fate. As we can see in Figure 4F, the single strandedness profiles of the miR-125a variants exhibit variability. At the moment, it is unclear which miRNA regions participate in which maturation step. The G22•C65 base pair in the major allele sta-

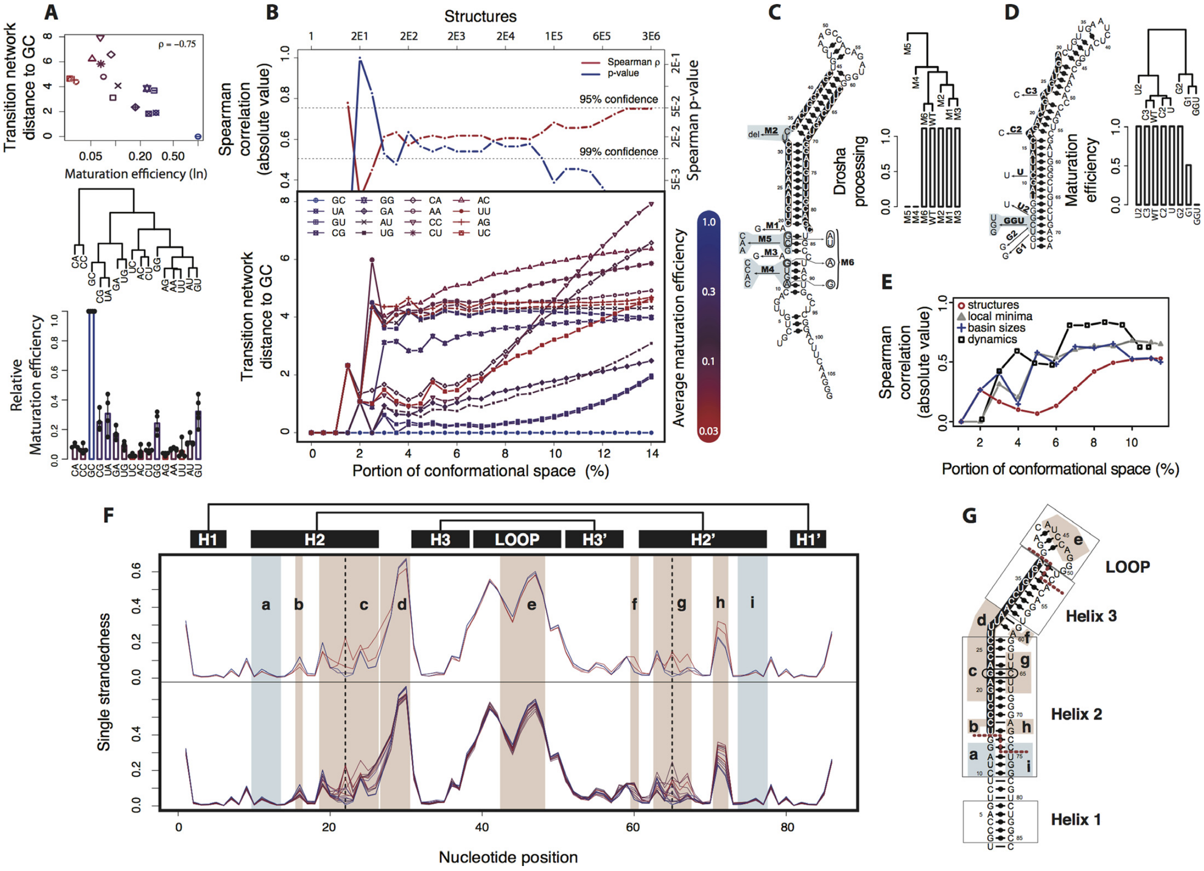

Figure 4. MiRNA transition networks and maturation efficiencies. The pair of nucleotides at positions 22 and 65 identifies each miR-125a variant: the major allele GC, the deleterious allele UC and so on. The mutants are color coded according to the $\log_{10}$ of their maturation efficiencies from red (low efficiency) to blue (high efficiency). The symbols and colors used in panel A are also used in panels B and F. (**A**) Top: the transition network distances to GC increase as maturation efficiencies decrease. Middle: hierarchical clustering based on pairwise transition network distances. Bottom: bar plot of the maturation efficiencies. The heights of the bars represent the average of four replicates, all shown by dots. (**B**) Bottom: transition network distance to GC of miR-125a variants at increasing portions of conformational space (up to 14%, 3 × 106 structures). Top: corresponding Spearman correlations (ρ) and P-values. (**C**) Secondary structure and, hierarchical clustering and relative levels of Drosha processing of seven miR-30a variants using the top 1000 predicted structures; the miR-30a major allele is labeled WT. (**D**) Secondary structure and, hierarchical clustering and relative levels of maturation efficiencies of eight miR-21 variants using the top 1000 predicted structures; the miR-21 major allele is labeled WT. (**E**) Spearman correlation at increasing portions of conformational space using the number of structures, number of local minima, average number of structures connected to local minima (basin sizes) and transition network distances using local minima. (**F**) Single strandedness profiles of the miR-125a variants. Top: single strandedness of the two most (GC and UA) and least (AG and UC) processed variants. Bottom: all variants. The vertical dotted lines indicate the mutated base pair (nucleotides 22 and 65). The A-helix regions are shown (5′ strands H1, H2 and H3; 3′ strands H1′, H2′ and H3′). The regions affected by the mutations are shown in pink (**b** to **h**, where **b**, **d**, **e**, **f** and **h** are distant from the mutation sites). The region under the Drosha cut is show in blue (**a** and **i**). (**G**) The minimum free energy secondary structure of the miR-125a major and minor alleles (the sequence of the major allele is shown). The A-helix regions are labeled as in (F). The Drosha (lower red line) and Dicer (upper red line) cuts are shown.

bilizes the surrounding regions of each partner. The blue line in the low single strandedness area centered at the dotted lines indicates them. The variants that stabilize these regions (blue lines) are processed more efficiently by the Drosha/DGCR8 complex than those that destabilize them (red lines).

Mutating the nucleotides at positions 22 and 65 affects the flexibility of multiple regions, as revealed by variations in their single strandedness profiles computed from the transition networks: helix 2 (H2) (segments **b**, **c**, **g** and **h** in Figure 4F and G), the hinge between H2 and helix 3 (H3) (segments **d** and **f** in Figure 4F and G) and the loop (segment **e** in Figure 4F and G). Relative to lowly expressed variants (red lines in Figure 4F), changing the G22•C65 base pair in the major allele destabilizes C16, region **b**, and the first nucleotide of region **c**, U19, but destabilizes A71 and G72, in the corresponding 3′ arm, region **h** (Figure 4F). It further stabilizes region **d** and destabilizes region **h** that directly affects the hinge area between H2 and H3 (Figure 4F and G).

The differences in single strandedness between the two most and least processed variants are highlighted in the top panel of Figure 4F. Even if we observe a great variation in single strandedness among all the variants in the loop (region **e** in the bottom panel of Figure 4F), the small variation observed in the top panel suggests that these changes in flexibility are not likely to be involved in the loss of maturation. The stem segment highlighted by the **b** and **h** regions in Figure 4F that is next to the Drosha cut displays a greater difference in single strandedness. One could expect that such a difference would be a better candidate for further studying the loss of maturation. Similarly, we observe a significant change in the single strandedness of the variants in the stem segment highlighted by the **c** and **g** regions in Figure 4F. The change occurs at the base pair that is mutated and propagates to the hinge between H2 and H3 (regions **d** and **f** in Figure 4F).

Note that the stem segment defined by the regions labeled **a** and **i** (blue region in Figure 4F), which is next to the Drosha cut site, was observed to be flexible in many miRNAs (39). This was confirmed experimentally by selective 2′-hydroxyl acylation by primer extension chemistry in miR-16-1, miR-30a and miR-107 (37). Furthermore, variants of these three miRNAs that stabilize this region were shown to mature less efficiently than their corresponding endogenous miRNAs *in vitro* (37). This flexible hot spot in the vicinity of the Drosha cleavage site is required for processing and its exact location can vary in different miRNAs. According to our analysis of the single strandedness, in miR-125a the **b** and **h** regions would play the role of this flexible region. The blue segment in Figure 4F and G indicate the flexible hot spot based on miR-30a (37).

Finally, to determine which statistical aspects of the sets of predicted structures best captured the maturation signal of miR-125a variants, we questioned whether the correlation is observed using simpler statistics or less data. Surprisingly, we found that the local minimum structures alone reveal some information about the classification of the miR-125a variants according to their maturation efficiencies (Figure 4E). This further emphasizes that a small number of the most stable transient structures control the maturation fate of pri-miRNA sequences.

Transient states and group II miRNAs

Some pri-miRNA are processed by the microprocessor to an alternate form, where Drosha cut sites exhibit a 1-nucleotide instead of the normal 2-nucleotide 3′ overhang. These pre-miRNAs require urydilation for further processing (40). To find whether these group II miRNAs differ from other miRNAs in their transient state networks, we compared pairwise all human miRNAs from miRBase (41) using the transition network metric. We then used principal components analysis to determine a suitable visualization plane, which is shown in Figure 5. The resulting density map clearly indicates that group II miRNAs do not cluster with the main density areas, but they do not form a set of similar points either as they spread over lowly populated areas (Figure 5A). The density map shows two peaks and a number of possible clusters, as well as a continuous path of dense areas. The central zone consists of a low density spread, and it is toward this valley that we find the group II miRNAs, indicating that they depart significantly from mainstream miRNA transition networks. The same data were submitted to cluster analysis using the DBSCAN algorithm that determines cluster membership according to the variation in density of point coordinates (42) (Figure 5B and Supplementary Table S2).

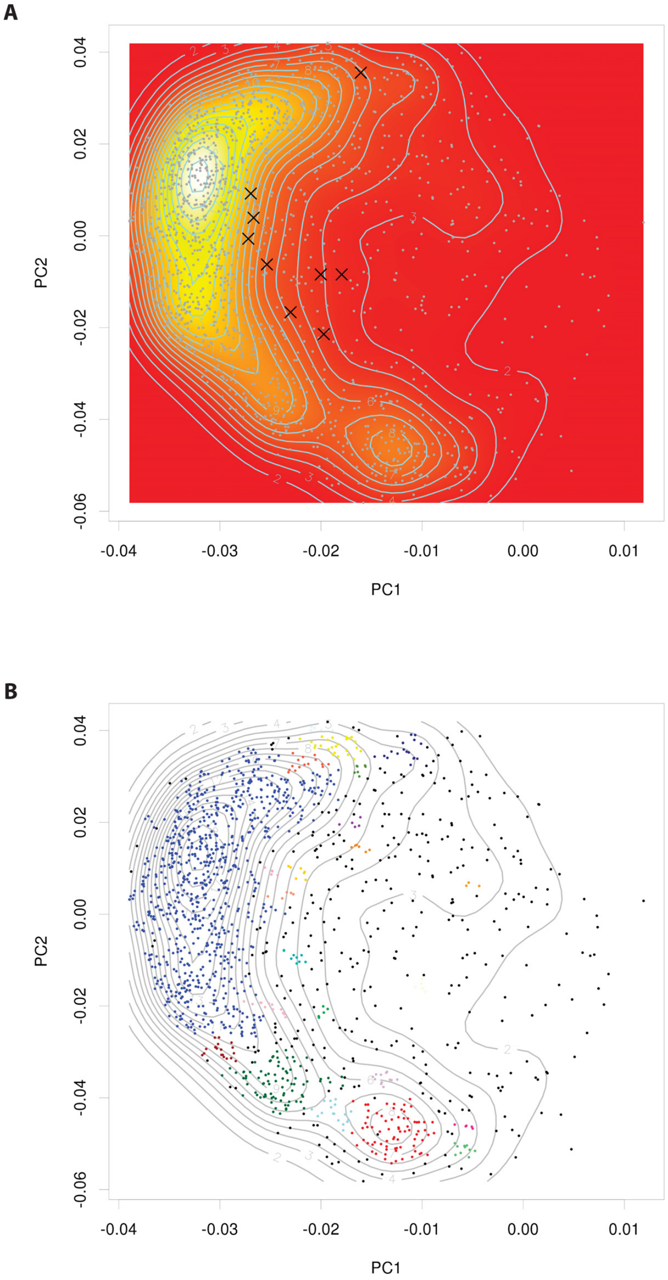

Figure 5. Groups of human pri-miRNAs. The transition network distance metric was applied pairwise to all human pri-miRNA hairpins forming segments from miRBase and the distribution of their distances was analyzed using principal components analysis. (**A**) Projection of miRNAs on the first two principal components (PC1 and PC2). The point density is highlighted from red (sparse) to white (dense), and emphasized by 20 contour lines. Most of the variability (79% SD) is captured in this plane, with the rest being orthonormal to the plane. (**B**) Same as in A, but the points were clustered, and each cluster is represented by a different color.

DISCUSSION

Since miRNAs were discovered about two decades ago (43), characterizing the mechanisms behind miRNA biogenesis has been the focus of a lot of research (44). The realization of the role of miRNA dynamics in their mechanisms of action brings a new perspective into this puzzling challenge. Using a quantitative model, we captured changes in miRNA maturation that are linked to their transition networks, which could be seen as a primitive model of RNA dynamics. Our results suggest that the miRNA maturation pathway is highly sensitive to sequence variations that modulate the topology of a miRNA's network of transient structures. We focused on a SNP in miR-125a that interferes with maturation and is highly associated with breast cancer patients (36), the ribosomal A-site, and a few other miRNA examples. In general, our model possibly applies to many, if not all, RNAs, and links RNA transition networks and function (45,46). Besides, this model suggests similarities between RNA and intrinsically disordered proteins, which fold and function upon specific chemical signals shown crucial in ligand binding (47).

The computational determination of RNA structure is often understood as a search for the most stable, native or active conformation among a huge conformational space. Many researchers are now hypothesizing that changes inducing variations in this space can be linked to perturbations in RNA function. We provide new insights indicating that, (i) the most stable transient structures included in a small fraction of the conformational space are relevant to function; and, (ii) non-canonical base pairs characterize the transitions between these transient conformational states. Understanding the effect of sequence variations on RNA function thus relies on the analysis of such networks of transient structures.

The method introduced here does not allow us yet to identify which miRNA structures are involved in maturation. Further analysis will most certainly require solving and/or modeling the 3D structures of the complexes involved. Until then, we can only speculate about which miRNA binding partners are affected by the flexibility changes that occur in the miRNA variants. Our analysis of the single strandedness of miR-125a variants allows us to propose the immediate vicinity of the Drosha cleavage site, in H2, or the hinge between H2 and H3. Whether these variable regions interfere with the interaction between the miRNA and the Drosha/DGCR8 complex, or favor an interaction with another yet unidentified factor in the nucleus is unclear and will need to be determined experimentally. Also, we cannot confirm whether one altered or many regions synergistically mediate the observed modulation of miRNA maturation efficiency.

Secondary structure prediction does not yet fully capture the entropic contribution of all chemical bonds and degrees of freedom. However, much progress in this direction has been made, in particular the development of the energy model in MC-Fold was an important step toward reducing the structural uncertainty of large loops and bulges. It provides a new way to enumerate their transient conformations by extending the set of base pairing types. The enumeration of these conformations brings 2D structure prediction closer to molecular dynamics than ever before.

There is no doubt that the identification and comparison of the conformations induced by the interactions with various cofactors, which are likely enumerated by MC-Fold, are necessary steps towards studying the relative biological activity of sequence variants. It is remarkable, given that 2D structure prediction derives from studies of static experimental RNA structures (10), that it is now possible to venture into the dynamic behavior of RNA molecules at the cellular level using sequence data alone.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Bryan R. Cullen for providing the flanking sequences of miR-21.
Authors contributions: Method design (P.D. and F.M.), method implementation (P.D.), experimental work (H.T., K.S., C.M. and P.J.) and data analysis and writing (P.D. and F.M.).

FUNDING

Natural Sciences and Engineering Research Council of Canada (NSERC) [RGPIN/04539-2014 to F.M.]; Canadian Institutes of Health Research (CIHR) [MOP-93679 to F.M.]; National Institutes of Health [R01GM088813 to F.M.]; Human Frontier Science Program [RGP0002/2009-C to F.M.]. Funding for open access charge: NSERC.
Conflict of interest statement. None declared.

REFERENCES

1. Gralla,J., Steitz,J.A. and Crothers,D.M. (1974) Direct physical evidence for secondary structure in an isolated fragment of R17 bacteriophage mRNA. *Nature*, **248**, 204–208.
2. Mandal,M. and Breaker,R.R. (2004) Gene regulation by riboswitches. *Nat. Rev. Mol. Cell Biol.*, **5**, 451–463.
3. Williamson,J.R. (2000) Induced fit in RNA-protein recognition. *Nat. Struct. Biol.*, **7**, 834–837.
4. Dethoff,E.A., Petzold,K., Chugh,J., Casiano-Negroni,A. and Al-Hashimi,H.M. (2012) Visualizing transient low-populated structures of RNA. *Nature*, **491**, 724–728.
5. Ding,Y. (2006) Statistical and Bayesian approaches to RNA secondary structure prediction. *RNA*, **12**, 323–331.
6. Hofacker,I.L., Bernhart,S.H. and Stadler,P.F. (2004) Alignment of RNA base pairing probability matrices. *Bioinformatics*, **20**, 2222–2227.
7. Mathews,D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178–1190.
8. McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
9. Sim,A.Y., Minary,P. and Levitt,M. (2012) Modeling nucleic acids. *Curr. Opin. Struct. Biol.*, **22**, 273–278.
10. Parisien,M. and Major,F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.
11. Kondo,J. and Westhof,E. (2008) The bacterial and mitochondrial ribosomal A-site molecular switches possess different conformational substates. *Nucleic Acids Res.*, **36**, 2654–2666.
12. Zeng,X., Chugh,J., Casiano-Negroni,A., Al-Hashimi,H.M. and Brooks,C.L. 3rd (2014) Flipping of the ribosomal A-site adenines provides a basis for tRNA selection. *J. Mol. Biol.*, **426**, 3201–3213.
13. Flamm,C., Fontana,W., Hofacker,I.L. and Schuster,P. (2000) RNA folding at elementary step resolution. *RNA*, **6**, 325–338.
14. Bonhoeffer,S., McCaskill,J.S., Stadler,P.F. and Schuster,P. (1993) RNA multi-structure landscapes. A study based on temperature dependent partition functions. *Eur. Biophys. J.*, **22**, 13–24.
15. Halvorsen,M., Martin,J.S., Broadaway,S. and Laederach,A. (2010) Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet.*, **6**, e1001074.

16. Sabarinathan,R., Tafer,H., Seemann,S.E., Hofacker,I.L., Stadler,P.F. and Gorodkin,J. (2013) RNAsnp: efficient detection of local RNA secondary structure changes induced by SNPs. *Hum. Mutat.*, **34**, 546–556.
17. Lorenz,W.A. and Clote,P. (2011) Computing the partition function for kinetically trapped RNA secondary structures. *PLoS One*, **6**, e16178.
18. Wuchty,S., Fontana,W., Hofacker,I.L. and Schuster,P. (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–165.
19. Duan,R., Pak,C. and Jin,P. (2007) Single nucleotide polymorphism associated with mature miR-125a alters the processing of pri-miRNA. *Hum. Mol. Genet.*, **16**, 1124–1131.
20. O'Connor,M., Thomas,C.L., Zimmermann,R.A. and Dahlberg,A.E. (1997) Decoding fidelity at the ribosomal A and P sites: influence of mutations in three different regions of the decoding domain in 16S rRNA. *Nucleic Acids Res.*, **25**, 1185–1193.
21. Denli,A.M., Tops,B.B., Plasterk,R.H., Ketting,R.F. and Hannon,G.J. (2004) Processing of primary microRNAs by the Microprocessor complex. *Nature*, **432**, 231–235.
22. Gregory,R.I., Yan,K.P., Amuthan,G., Chendrimada,T., Doratotaj,B., Cooch,N. and Shiekhattar,R. (2004) The Microprocessor complex mediates the genesis of microRNAs. *Nature*, **432**, 235–240.
23. Lee,Y., Ahn,C., Han,J., Choi,H., Kim,J., Yim,J., Lee,J., Provost,P., Radmark,O., Kim,S. *et al.* (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature*, **425**, 415–419.
24. Lund,E., Guttinger,S., Calado,A., Dahlberg,J.E. and Kutay,U. (2004) Nuclear export of microRNA precursors. *Science*, **303**, 95–98.
25. Yi,R., Qin,Y., Macara,I.G. and Cullen,B.R. (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev.*, **17**, 3011–3016.
26. Bernstein,E., Caudy,A.A., Hammond,S.M. and Hannon,G.J. (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, **409**, 363–366.
27. Grishok,A., Pasquinelli,A.E., Conte,D., Li,N., Parrish,S., Ha,I., Baillie,D.L., Fire,A., Ruvkun,G. and Mello,C.C. (2001) Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell*, **106**, 23–34.
28. Hutvagner,G., McLachlan,J., Pasquinelli,A.E., Balint,E., Tuschl,T. and Zamore,P.D. (2001) A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*, **293**, 834–838.
29. Ketting,R.F., Fischer,S.E., Bernstein,E., Sijen,T., Hannon,G.J. and Plasterk,R.H. (2001) Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes Dev.*, **15**, 2654–2659.
30. Heo,I., Joo,C., Cho,J., Ha,M., Han,J. and Kim,V.N. (2008) Lin28 mediates the terminal uridylation of let-7 precursor microRNA. *Mol. Cell*, **32**, 276–284.
31. Viswanathan,S.R., Daley,G.Q. and Gregory,R.I. (2008) Selective blockade of microRNA processing by Lin28. *Science*, **320**, 97–100.
32. Davis,B.N., Hilyard,A.C., Nguyen,P.H., Lagna,G. and Hata,A. (2010) Smad proteins bind a conserved RNA sequence to promote microRNA maturation by Drosha. *Mol. Cell*, **39**, 373–384.
33. Guil,S. and Caceres,J.F. (2007) The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. *Nat. Struct. Mol. Biol.*, **14**, 591–596.
34. Michlewski,G., Guil,S., Semple,C.A. and Caceres,J.F. (2008) Posttranscriptional regulation of miRNAs harboring conserved terminal loops. *Mol. Cell*, **32**, 383–393.
35. Hsieh,T.H., Hsu,C.Y., Tsai,C.F., Long,C.Y., Chai,C.Y., Hou,M.F., Lee,J.N., Wu,D.C., Wang,S.C. and Tsai,E.M. (2015) miR-125a-5p is a prognostic biomarker that targets HDAC4 to suppress breast tumorigenesis. *Oncotarget*, **6**, 494–509.
36. Li,W., Duan,R., Kooy,F., Sherman,S.L., Zhou,W. and Jin,P. (2009) Germline mutation of microRNA-125a is associated with breast cancer. *J. Med. Genet.*, **46**, 358–360.
37. Quarles,K.A., Sahu,D., Havens,M.A., Forsyth,E.R., Wostenberg,C., Hastings,M.L. and Showalter,S.A. (2013) Ensemble analysis of primary microRNA structure reveals an extensive capacity to deform near the Drosha cleavage site. *Biochemistry*, **52**, 795–807.
38. Zeng,Y. and Cullen,B.R. (2003) Sequence requirements for micro RNA processing and function in human cells. *RNA*, **9**, 112–123.
39. Han,J., Lee,Y., Yeom,K.H., Nam,J.W., Heo,I., Rhee,J.K., Sohn,S.Y., Cho,Y., Zhang,B.T. and Kim,V.N. (2006) Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell*, **125**, 887–901.
40. Heo,I., Ha,M., Lim,J., Yoon,M.J., Park,J.E., Kwon,S.C., Chang,H. and Kim,V.N. (2012) Mono-uridylation of pre-microRNA as a key step in the biogenesis of group II let-7 microRNAs. *Cell*, **151**, 521–532.
41. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
42. Ester,M., Kriegel,H.-P., Sander,J. and Xu,X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis,E, Han,J and Fayyad,U (eds). *Proc. 2nd Int. Conf. Knowl. Discov. Data Mining*. AAAI Press, pp. 226–231.
43. 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.
44. Breving,K. and Esquela-Kerscher,A. (2010) The complexities of microRNA regulation: mirandering around the rules. *Int. J. Biochem. Cell Biol.*, **42**, 1316–1329.
45. Gkogkas,C.G., Khoutorsky,A., Ran,I., Rampakakis,E., Nevarko,T., Weatherill,D.B., Vasuta,C., Yee,S., Truitt,M., Dallaire,P. *et al.* (2013) Autism-related deficits via dysregulated eIF4E-dependent translational control. *Nature*, **493**, 371–377.
46. Kloc,M., Dallaire,P., Reunov,A. and Major,F. (2011) Structural messenger RNA contains cytokeratin polymerization and depolymerization signals. *Cell Tissue Res.*, **346**, 209–222.
47. Ferreon,A.C., Ferreon,J.C., Wright,P.E. and Deniz,A.A. (2013) Modulation of allostery by protein intrinsic disorder. *Nature*, **498**, 390–394.