Published online 29 November 2016 *Nucleic Acids Research, 2017, Vol. 45, No. 3* ***1079–1090***
doi: 10.1093/nar/gkw1031

Large-scale analysis of microRNA expression, epi-transcriptomic features and biogenesis

Dimitrios M. Vitsios, Matthew P. Davis, Stijn van Dongen and Anton J. Enright*

EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received September 22, 2016; Revised October 17, 2016; Editorial Decision October 18, 2016; Accepted October 18, 2016

ABSTRACT

MicroRNAs are important genetic regulators in both animals and plants. They have a range of functions spanning development, differentiation, growth, metabolism and disease. The advent of next-generation sequencing technologies has made it a relatively straightforward task to detect these molecules and their relative expression via sequencing. There are a large number of published studies with deposited datasets. However, there are currently few resources that capitalize on these data to better understand the features, distribution and biogenesis of miRNAs. Herein, we focus on Human and Mouse for which the majority of data are available. We reanalyse sequencing data from 461 samples into a coordinated catalog of microRNA expression. We use this to perform large-scale analyses of miRNA function and biogenesis. These analyses include global expression comparison, co-expression of miRNA clusters and the prediction of miRNA strand-specificity and underlying constraints. Additionally, we report for the first time a global analysis of miRNA epi-transcriptomic modifications and assess their prevalence across tissues, samples and families. Finally, we report a list of potentially misannotated miRNAs in miRBase based on their aggregated modification profiles. The results have been collated into a comprehensive online repository of miRNA expression and features such as modifications and RNA editing events, which is available at: http://wwwdev.ebi.ac.uk/enright-dev/miratlas. We believe these findings will further contribute to our understanding of miRNA function in animals and benefit the miRNA community in general.

INTRODUCTION

Non-coding regulators such as miRNAs have been a significant avenue of research since their discovery and the realization that these molecules are both widespread in animals and plants and also frequently highly conserved. The main mode of regulation by miRNAs in animals is translational repression and degradation of target transcripts (1). This regulation involves the binding of a mature 19–22 nt miRNA to a target transcript through direct formation of a double stranded duplex driven by complementarity between the miRNA and the target site (2). This binding event is initiated through the so-called 'seed' region of the miRNA (nts 2–8) which requires for the most part perfect complementarity. The biogenesis of miRNAs is now relatively well characterized (3). They are encoded by long non-coding transcripts or as passengers in the introns and UTRs of protein-coding transcripts. They are formed as 70–120 nt stem-loop structures on the host molecule and are recognized and excised by enzymes including Drosha and DGCR8. The resulting cleaved hairpin molecule is referred to as a miRNA precursor and these pre-miRNAs are exported from the nucleus to the cytoplasm where they enter the RNA silencing machinery. The enzyme Dicer with cofactors excises a double-stranded duplex from the pre-miRNA which is unwound. In general one of the strands is degraded (the passenger strand) and the other strand becomes a mature single-stranded miRNA capable of being loaded into the RNA-induced silencing complex (RISC) and capable of silencing target transcripts.

Large-scale cloning and sequencing of small RNAs using capillary sequencing allowed the initial detection of large sets of animal miRNAs (4). However, the advent of next-generation sequencing (NGS) allowed these molecules to be rapidly detected in different tissues and organisms. The primary repository for miRNAs is the miRBase database (5). Initially, miRNAs were usually confirmed by northern blot or similar assay prior to their inclusion in miRBase. However, the advent of large-scale NGS studies has meant that it is impractical to confirm every single sequence detected via targeted amplification. Given that the genome is replete with putative stem-loop structures and that small RNA sequencing detects many short molecules and degradation products, there are many putative miRNA sequences in miRBase which may in fact not be canonical miRNAs but instead may be other functional ncRNAs or the degradation products of longer molecules.

*To whom correspondence should be addressed. Tel: +44 1223 492 668; Fax: +44 1223 492 620; Email: aje@ebi.ac.uk

While NGS has greatly increased our power to detect and catalog miRNA expression these data are usually complex and are processed differently from laboratory to laboratory. Hence, while there are currently over 850 deposited small RNA sequencing datasets (ENA, GEO (6)) there isn't a comprehensive database or catalog of where and when these miRNAs have been detected. Additionally as each experiment has been processed with different criteria and filters the results may be difficult or impractical to compare directly. We sought to address these issues by building a comprehensive catalog of miRNA expression from large numbers of previously published small RNA sequencing datasets for both Human and Mouse, for which raw FASTQ data are available. For each dataset we have performed automated barcode demultiplexing, 5′/3′ adapter detection using *de Bruijn graph* analysis followed by adapter excision and computational size selection (15–32 nts). Additionally, some samples require the removal of poly-A or poly-C tracts. This data pre-processing step has been performed by a pipeline which was based on the already published pipelines Kraken (7) and Chimira (8). Each dataset has been mapped to known miRNA precursor sequences using a single computational pipeline (Chimira). This pipeline not only represents a cohesive platform for the collation and analysis of small RNA NGS data but also allows the detection of events such as 5′/3′ modification of miRNAs via enzymes such as terminal uridylyl transferases (tutase) (9) or adenosine deaminase RNA (ADAR) editing (10). The raw count data obtained was normalized and annotated according to each experiment, providing a comprehensive catalog of miRNA expression in Human and Mouse together with a variety of complementary data that can assist us in the analysis of miRNA function and biogenesis.

Using this comprehensive dataset we have performed the largest analysis to-date on miRNA expression, expression of miRNA clusters and the prevalence of miRNA modifications. Additionally, we provide a database of miRNA expression and modifications for Human and Mouse accompanied with tools for advanced query searches, that we believe will prove useful to the community.

MATERIALS AND METHODS

Dataset annotation

Annotation for the *miratlas* database has been generated manually based on the information that is available in the original databases for each dataset. The curated annotation classes may refer to either a cell line/type/tissue (e.g. liver) or a condition/disease (e.g. cancer). In case both a cell line/type/tissue and a condition/disease are provided for a dataset, only the condition/disease information is used for the annotation of that dataset. Additional information is provided in the *miratlas* repository as well as the links to the original resources.

Data normalization

Raw expression data and modification count tables have been normalized using DESeq2 (11). Each dataset was provided as a distinct condition at the design formula of the DESeq2 normalization method.

Identification process for complex read geometry inference

Input datasets used for this analysis have been prepared, in general, by different experimental protocols using a variety of barcodes, 3′ adapters and/or 5′ adapters. Thus, it is imperative first of all to infer the read geometry of each input dataset in order to later clean the sequences from barcodes/adapters and further process the samples. We have developed a pipeline that is deciphering the presence or not of barcode sequence in the input samples by looking for the enrichment of any sequence of 3–6 nt long at the 5′ end of the first 2 million sequences of an input sample file. Inference of the 3′ adapter is accomplished through the command-line version of the 3′ adapter detection feature of Chimira (8), which integrates minion and swan (7). In that case though, the position of the suggested adapter relative to the input sequences is also defined and thus the inferred adapter may either be a 5′ or 3′ adapter. In case the suggested adapter sequence does not match at least 90% with a known Illumina adapter sequence (without any mismatches), input files are also manually checked in order to identify any potential sequences that are attached to already known highly expressed miRNAs, such as the let-7 miRNAs. The full pipeline for the inference of barcode and adapter sequences is described at the flowchart at Supplementary Figure S1. Datasets from ENA/GEO that were detected with ambiguous adapter sequences or barcode annotation were excluded from the analysis. Eventually, we compiled a set of 52 datasets with a well characterized read geometry that we used for our analysis.

Adapter trimming

Following inference of the 3′ adapter sequence, trimming is performed by Chimira or SequenceImp (for the loci-specific expression analysis) that both integrate the tool *Reaper*.

Genomic clusters definition

Genomic clusters are defined as follows (Supplementary Figure S2B.a): let *mir_1, mir_2* and *mir_3* be three miRNA genes in neighboring locations on the genome without any other miRNA genes interfering at the genomic space between *mir_1* and *mir_3*. Then, *mir_1*, *mir_2* belong to the same genomic cluster (GC-1) if and only if: $d_{12} \leq W$. *mir_3* also belongs to GC-1 if and only if $d_{23} \leq W$ (d_{13} may be greater than W but it will be less than or equal to 2W). Thus, a genomic cluster may contain pairs of miRNAs whose distance is greater than W but for each miRNA there is at least another miRNA in that cluster that is closer to it less than W base pairs.

Functional clusters generation

We created a correlation matrix for the co-expression of all miRNAs detected in this study. This matrix defines a weighted graph and weights of its edges correspond to the correlation of expression between two miRNAs. In order to obtain the functional clusters, we clustered this graph using MCL (12) setting the value of the *filter threshold* parameter to 0.8.

Small RNA-Seq data processing

Chimira has been used in order to trim input sequences from sequencing adapters, align them against the human or mouse hairpin precursors and extract miRNA expression data and their associated modifications. In case of more complex read geometry SequenceImp has been used for cleaning the input data before alignment. Besides, SequenceImp (7) has been used in order to get miRNA expression data with loci-specific information at the genomic level in order to correlate miRNA co-expression with genomic proximity.

Modification analysis using collapsed patterns

Modification counts were extracted by Chimira. Default output contains all identified modification patterns at the 3′, 5′ ends of miRNAs as well as internal modifications (single-nucleotide polymorphisms, ADAR-edits) within the miRNA sequences. Each pattern is associated with the exact location of this modification relative to the original sequence. In order to study the expression and distribution of very specific variants, i.e. adenylated, uridylated, guanylated and cytocylated, we have simplified Chimira's output for each dataset into collapsed modification tables. These data format contains only miRNA variants with a single-nt or poly-nt modification only, where nt = {A, U, G, C} and poly-nt may refer to any sequence of two or more identical nucleotides. All other variants are ignored and their counts are collapsed with the counts of the respective wild-type miRNA forms.

Detection of Illumina sequencing biases patterns

We filtered the reads from 12 human samples sequenced by Illumina, retaining only those that were at least 10 nt shorter than the maximum length among all the reads, which is the length that occurs more frequently among the reads of the sample. This filtering process allows us to retain only the reads that may correspond to 3′ exons of actual mRNA transcripts. The filtered reads were then aligned against the 3′ human exons of the reference database we have constructed, allowing the identification of sequence artefacts that are appended to the 3p end of the transcripts that probably represent sequencing artefacts (Supplementary Figure S3E).

Strand selection analysis

Free energy calculations have been performed using the *RNAfold* program from the *RNASoft* suite (13) and free-energy parameters for predictions of RNA duplex stability were based on previously published work (14). Complementary miRNA counts of a miRNA contain counts from all possible loci at the genome, in case the id of that miRNA is not indicative of its genomic location origin and there are multiple paralogs for the same mature sequence. For instance, if '*hsa-let-7a-5p*' is the miRNA of interest and $counts_{(arm)}$ *is its expression depth*, then the expression depth of its complementary miRNAs ($counts_{(compl\_arm)}$) will include the counts from both '*hsa-let-7a-3p*' and '*hsa-let-7a-2-3p*', since there is no loci specific information at the id of the '*hsa-let-7a-5p*' miRNA.

Detection of mis-annotated miRNAs

We used the modification counts extracted from Chimira and their accompanied modification index positions to construct the overall coverage profile for each miRNA, that includes template counts and mismatches at each nucleotide position. From the entire set of miRNAs for each species, we retained only those that were expressed in at least 5% of all datasets of that species. We then parsed all coverage profiles and highlighted those that had four or more nucleotide positions with a mismatch ratio >70%. The filtered list of modification profiles was then compared manually to the profiles of the control miRNAs in order to retain only the miRNAs with a non-canonical modification profile.

RESULTS

A total of 52 NGS datasets were obtained from both ENA and GEO covering in total 461 biological samples including biological replicates (Supplementary Table S1). For each dataset, FASTQ raw data were downloaded and annotation information was manually curated according to tissue, cell type, disease state or cell line (see 'Materials and Methods' section). These raw data serve as the foundation for all subsequent analyses described below. Of particular note in the case of small RNA datasets is that the molecule being sequenced is usually shorter than the sequence read obtained from an NGS experiment. This means that most captured sequences contain both small RNA sequence and some amount of the 3′ sequencing adapter. Because the 3′ adapter sequence used in the original experiment was not readily available, we automatically inferred the most likely adapter used, based on 3′ *de Bruijn* graph assembly and removed the adapter sequences using *Reaper* (see 'Materials and Methods' section). Finally, these adapter purged sequences (representing small RNAs and contaminants) were de-duplicated, using *Tally*, such that each sequence was only represented once in the final input FASTA file accompanied with its respective coverage depth. These cleaned and de-duplicated sequences were the primary input into the miRNA analysis pipeline (Chimira). This pipeline automatically scans each sequence against all known miRBase precursor sequences from a selected species and detects the likely miRNA, which arm of the precursor it originated from (5′/3′) and searches for non-canonical nucleotides which may be the result of editing and/or modification by enzymes such as Tutases. All miRNA counts, annotations and features detected are stored in a MySQL database for further analysis.

Global analysis of microRNA expression

In order to validate the initial results and to evaluate how well the automated small RNA analysis performs we normalize the count data (see 'Materials and Methods' section) and perform sample-to-sample unsupervised clustering based on co-expression correlation analysis. This allows us to explore both the sample to sample variation of miRNAs and to identify clusters of miRNAs which are significantly over-represented in certain datasets. Additionally, it allows us to identify groups of samples with very similar miRNA profiles. Our aim is hence to explore

miRNA expression across this heterogeneous pool of data and to characterize patterns among datasets of similar or different conditions. This analysis (Figure 1A, B, D and E) clearly demonstrates clustering of both miRNAs and samples across the datasets.

For miRNAs, the data clearly show a disparity between highly tissue specific and ubiquitously expressed miRNAs (Supplementary Figure S2A). For example, the let-7 family of miRNAs are among the most abundant and widespread detected miRNAs as expected, together with miR-21, miR-191 and miR-92a. Some highly expressed miRNA clusters also show distinct expression, including the miR-106b-25 cluster and the miR-17-92 cluster. Two miRNAs, hsa-miR-147a and hsa-miR-518a-5p, were expressed only in placenta tissue samples, which may imply that their functionality is exclusively influencing embryonic development in humans. Moreover, six miRNAs (hsa−miR−3689b−3p, hsa−miR−5707, hsa−miR−4534, hsa−miR−5583−5p, hsa−mir−3529−3p, hsa−mir−603) are expressed only in a particular dataset from lymphoma cell lines. The miR-302 cluster, thought to be important for pluripotency and cell-cycle regulation was among the most specifically expressed clusters, being predominantly expressed only in embryonic stem cells and in brain cancer. Overall, however these data are complex and it is convenient to instead perform pairwise clustering of miRNAs and samples separately to better detect significant commonalities and differences between miRNAs in one analysis and samples in the second analysis. This functionality is available in the web-based interface of *miratlas*.

For sample to sample correlations (Figure 1B and E) we observe specific groups of tissues and conditions clustering together for example cancer cell lines, B-cells and similar tissues. For some tissues and cell types multiple experiments from different sources are available. These would ideally have extremely correlated results with differences being explained by differing NGS platforms or experimental strategies. We observe on average 0.79 correlation of miRNA counts across seven human datasets where the same tissue or cell type has been profiled (0.82 respectively across 11 samples in Mus Musculus). In contrast, taking random comparisons of different tissues resulted in an average correlation of 0.68 in human and 0.69 in mouse. Clearly, although RNA extraction protocols, sequencing platform and sample treatment account for variation between samples from the same tissue, the correlations remain highly significant ($P \leq 0.018$ for human and $P \leq 0.005$ for mouse).

Three sample types show much lower expression than others (Saliva, Spermatozoa and Serum from pulmonary tuberculosis). These samples do not cluster effectively as they are difficult to normalize due to low sequence counts. In these cases it is likely that the correlation observed is spurious and primarily due to low-counts and/or contamination with RNA degradation products. However, the spermatozoa sample likely has low counts due to the previously observed paucity of small RNAs detectable in sperm (15,16). Clearly, one of the most defined features of the miRNA expression level correlation within Human and Mouse is due to the fact that many miRNAs are co-expressed from the same host transcript. We next sought to explore the expression of miRNAs while taking into consideration their genomic context and likely transcriptional unit.

microRNA clusters derived from genomic proximity

It is well known that many groups of miRNAs are encoded by a single transcript (coding or non-coding). These miRNA clusters are usually predicted by virtue of their close proximity on the genome. Previous computational studies have suggested that miRNA hairpins lying within 10 kb (17) are likely to be co-transcribed. We sought to update these findings from earlier studies, based on EST and cDNA data, with the data described here. Additionally we use both the genomic location and also miRNA co-expression analysis to re-evaluate these predictions and to generate novel miRNA clusterings. For this analysis we assess the accuracy of genomic clusters of miRNAs predicted using different genomic distance thresholds and miRNA co-expression as a measure of their co-regulation.

We first define all possible miRNA genomic clusters using a custom window of size W. The W parameter has been selected as large as possible while still retaining the number of clusters with negative intrinsic correlation at a relatively low level (Supplementary Figure S2Bb and c). Based on these criteria and the relevant literature (5,17), we assign 10 000 nt as our window size for further analysis. This value produces a total of 153 genomic clusters in human and 92 clusters in mouse. After the genomic clusters have been constructed, we calculate the average correlation of miRNAs co-expression within each genomic cluster (Supplementary Figure S2Bd and e). The number of clusters with positive intrinsic correlation compared to those with negative correlation is statistically significant ($P \leq 10^{-5}$), based on a model that is constructed as the average consensus of 10 runs with random genomic cluster assignments to the miRNAs of our study. We additionally observe that 33.3% of all genomic clusters in human datasets demonstrate a significant average intra-cluster correlation of >0.7 ($P \leq 2.7 \times 10^{-6}$). Interestingly, there are 18 clusters in the human datasets and 12 in the mouse datasets that have non-significant negative correlation values (-0.3 to 0.0). In these instances the small RNAs detected likely are transcribed from separate transcriptional units, products of alternative splicing, possibly mis-annotated RNA degradation products or under some other form of complex regulation. One interesting example with poor expression correlation is the cluster containing hsa-miR-1306 and hsa-miR-3618. These miRNAs are products of the DGCR8 transcript with the miR-3618 hairpin present in the 5′UTR being processed by the microprocessor complex as part of DGCR8s complex transcriptional control mechanism (18).

Clusters derived from miRNA co-expression

Another way to explore the clustering of miRNAs is to look for functional clustering of miRNAs based solely on their co-expression. The assumption here is that miRNAs with high expression correlation are likely to be involved in similar biological systems. We expect that clusters defined in this manner should show considerable overlap with clusters derived from the genomic proximity analysis above. However,

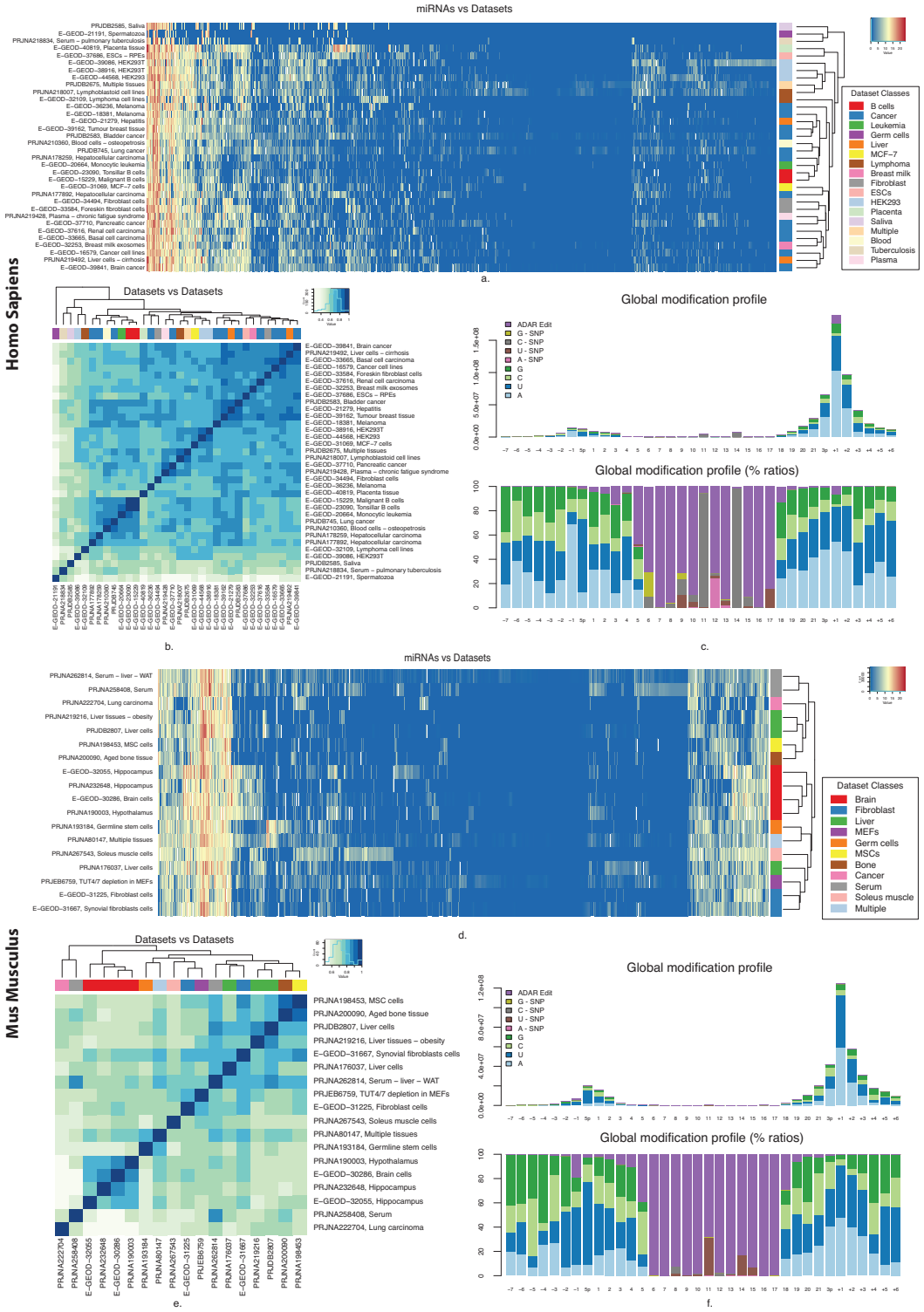

Figure 1. Global miRNA expression and modification profiles. (**A** and **D**) miRNAs expression profile across all 34 human/18mouse datasets. (**B** and **E**) Sample to sample clustering based on the global miRNAs expression. (**C** and **F**) Aggregate modification profiles from the human and mouse datasets positioned with reference a mature miRNA sequence.

we may also be able to identify groups of miRNAs encoded by transcripts at different genomic loci that still exhibit correlated expression of their host transcripts and may well be functionally linked. As expected, results show (Figure 2) significant overlap between clusters derived from proximity and those derived by expression correlation. However, we also observe a subset of functional links between groups of miRNAs expressed at different genomic loci with significant expression correlation. For example transcriptional cluster 3 (Figure 2A) is comprised of a number of genomic clusters including those on chrX, 19 and 13. Close inspection of these transcriptionally linked clusters (Figure 3) indeed indicates a preponderance of EGR1 transcription factor motifs, coupled with SP1 and NRF1. These results indicate that the high degree of transcriptional correlation observed between these three genomic clusters is a result of their being driven by the same transcriptional inputs. The high majority of the rest of transcriptional clusters with divergent genomic origin content contain either miRNA paralogs or let-7 family miRNAs, in both human and mouse.

Calling and analysis of the prevalence of microRNA modifications

In the past few years it has been widely demonstrated that miRNAs go through post-transcriptional alterations that can modify their 3′ ends, mainly via mono- or poly-uridylation (9,19). Such epi-transcriptomic alterations can have tremendous regulatory impact including how the small RNA machinery in the cell processes these molecules or whether or not they are degraded. In this study, we present for the first time a global profile of miRNA modifications occurring at both 3′ and 5′ ends. In order to identify the modifications in both ends of each miRNA we have employed additional analysis steps where all primary miRNA sequences are mapped against miRNA precursors using Chimira (8). Chimira scans the aligned regions in order to detect bases in the miRNA sequence that are not encoded in the genomic sequence. These unalignable nucleotides can be any of the following classes: (i) base-calling errors, (ii) single nucleotide polymorphisms or (iii) post-transcriptional miRNA modifications (e.g. via tutases). Base-calling errors are pseudo-random, platform-dependent and are more likely to occur at the 3′ end of a sequencing read, although at relatively low frequencies. SNPs are easier to detect as they will be present in a significant fraction of all reads observed. Finally, modifications such as uridylation or ADAR editing can be detected due to their being highly skewed toward particular modifications (e.g. mono-U, poly-U or A→U).

Overall, we find that 3′ modifications are far more prevalent than detected 5′ modifications (Figure 1C and F). In total, 95 human (4.4%) and 142 mouse (7.8%) miRNAs showed on average significant levels of 3′ modification (i.e. more than 25%). Similarly, 23 human (1.1%) and 24 mouse (1.3%) miRNAs showed on average significant levels (i.e. more than 25%) of 5′ modification. Mono and dinucleotide additions are the most common modifications, although longer modifications were observed too, albeit at lower frequencies (Supplementary Figure S3A and B). In both Human and mouse we observe a preponderance of Adenosine and Uracil modifications (Figure 1C) suggesting that both adenylation and uridylation by TUTases are likely the primary modifications made to miRNAs at least in animal systems. Both cytoplasmic adenylation by GLD-2 (20) and terminal uridylation by Tut4/Tut7 have been reported before as important for miRNA stability and degradation (19). However, we believe this is the first large scale detection and analysis of these events across animal tissues.

In order to investigate the significance of the presence of 3′ Guanine and Cytosine modifications, we performed an analysis in 12 human samples from mRNA-Sequencing experiments that were derived from Illumina Sequencing instruments to identify whether these G:C modifications may result from known sequencing biases present in the instrument (see 'Materials and Methods' section). To evaluate this we assume that G:C sequencing biases for mRNA samples will be largely similar to those obtained from small RNA sequencing. However, we would not expect any terminal modifications to occur within sequencing reads derived from exonic mRNA sequence, any non-genomic nucleotides observed are more likely to be sequencing errors.

The derived profile of sequencing biases is very rich in Gs and Cs (Supplementary Figure S3E) and greater than 65% of all observed errors for mono and dinucleotide errors. With regards to the datasets that have been used in our large-scale miRNA analysis, some of them are derived from Illumina Sequencing instruments, while others are derived from different types of instruments and another significant percentage among them do not provide in their annotation any information about the sequencing instruments that have been used during the experiment. The lack of annotation makes it difficult to computationally model and filter these likely G:C biases, however the data suggest that for the most part they are largely sequencing artefacts. Besides, this strongly suggests that the observed A:U enrichments are highly unlikely to be due to such sequencing artefacts and instead represent valid biological effects.

The prevalence of 5′ modifications is far lower than that observed for 3′ changes. Although some tRNAs are known to have 5′ modifications, we are not aware of any reported biochemical experiments of 5′ modification of small RNAs. For both human and mouse however a preponderance of 5′ A and U modifications are observed but are extremely rare as compared to 3′ modifications. It has already been reported that the 5p ends of miRNAs are generally post-processing stable in contrast with the 3p ends (21). Additionally, addition of 5′ nucleotides would dramatically alter the targeting of a miRNA loaded into the RISC complex. This may explain the lower count numbers and also the lower variability of the 5p modifications in comparison with the 3p modifications. However, certain datasets, among those from spermatozoa, monocytic leukemia and saliva samples (Supplementary Figure S3F–H) as well as two cancer datasets (E-GEOD-39841: brain cancer and E-GEOD-36236: skin cancer; profiles available in *miratlas*) exhibit a high ratio of 5p modifications, especially at the first nt upstream to the 5′ end. These modifications essentially redefine the seed region patterns of the modified miRNAs and consequently change the repertoire of the mRNAs that are being targeted by them. This may be affecting the functionality of some or all of these tissues by causing irregularities related with disease conditions.

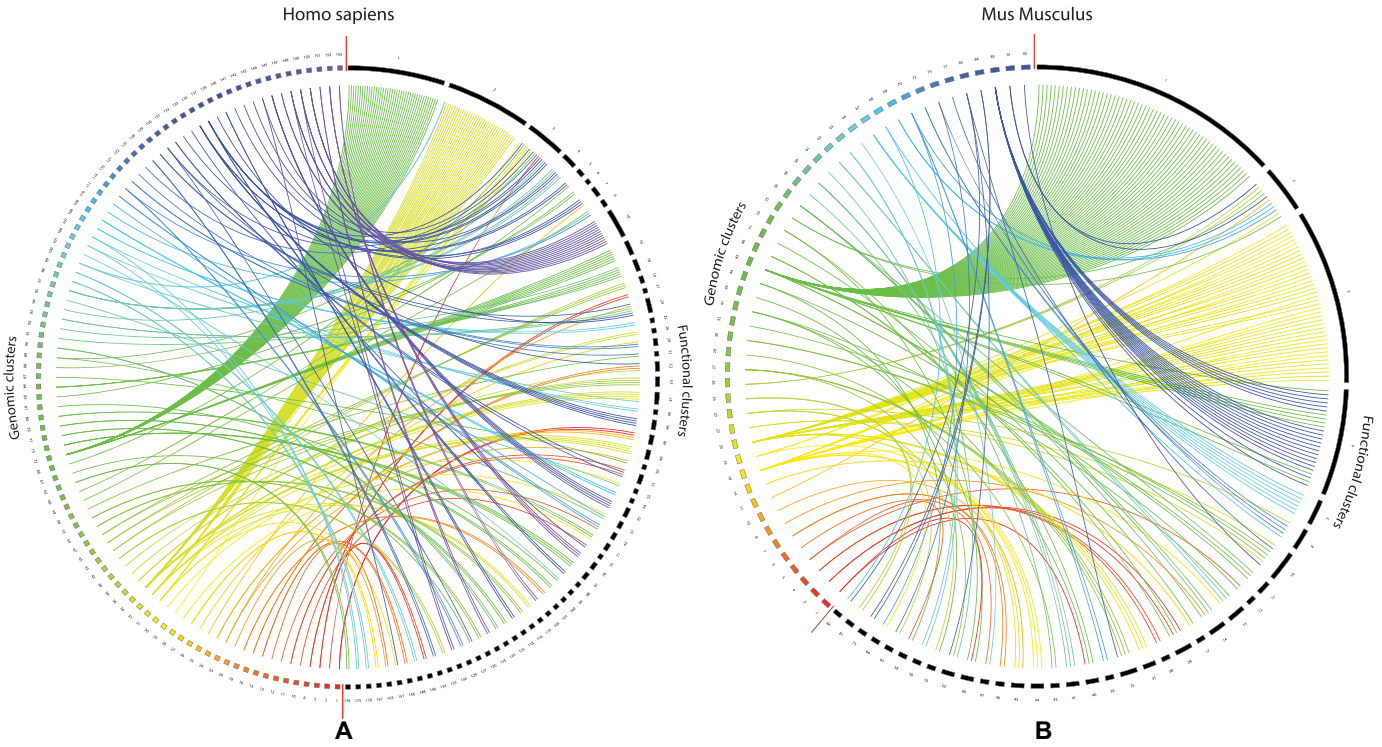

Figure 2. Associations of functional miRNA clusters with the respective genomic clusters. (**A**) Human samples. (**B**) Mouse samples. Each functional cluster is denoted by a black colored arc with numeric id. The length of the arc is proportional to the size of the cluster it represents. MicroRNAs that don't have any genomic cluster assignment have been omitted from this analysis for the sake of clarity of the figure. Genomic clusters correspond to the arcs of fixed length, colored with a non-black hue and they are sorted clockwise based on their proximity at the genome.

In the majority of cases, modifications affect less than or around 20% of the total expression depth while there are very few cases that they reach 30-40% of the total depth (Figure 4A and D). Moreover, the prevalence of 3′ modifications is significantly higher than 5′ modifications (Figure 4). A strong enrichment for 3′ A and U modifications is observed in both human and mouse which is in agreement with previous studies implicating tutase enzymes. For 5′ modifications a smaller enrichment is observed for 5′ Adenylation, however this enrichment is only observed to be significant in Human samples and a corresponding shift is not observed in mouse. The mild enrichment for 5′ Adenosine is puzzling and possibly reflects the presence of 5′Methyl adenosine sites known to be important for primary miRNA processing.

We now explore the distribution of adenylated and uridylated variants across all datasets (Figure 5 and Supplementary Figure S4). We focused on the expression of the let-7 family miRNAs and we observed a high resemblance of their modification profiles for these particular variants. Only mir-98 has a markedly different profile, potentially due its lower expression compared to other members of the let-7 family. We also projected the modification distribution of a set of highly expressed miRNAs. Some of these profiles show high similarity with the respective let-7 profiles while others demonstrate a relatively low modification depth (Supplementary Figure S5). This finding suggests that the frequency of modification events is not always associated with miRNA abundance but may be driven by other factors related to a particular condition, cell type or tissue. Several dataset types also tend to cluster together based on the overall expression of different types of variants in both species (Supplementary Figure S6) or the modification frequencies of the most highly modified miRNAs (Supplementary Figures S7 and 8). However, although overall expression of let-7 miRNAs is very similar across samples of the same cell type or condition, the expression of individual let-7 variants (e.g. adenlyated, guanylated) seems to deviate even for samples of the same annotation class.

For ADAR editing events, we observe an enrichment for brain in both Human and Mouse (Supplementary Figures S7 and 8), in correspondence with previous studies (10). Additionally, we observe an enrichment in serum and some cancer samples of non neuronal origin. The rate of ADAR editing observed in brain samples is (2%) and it occurs most predominantly in the seed region of miRNAs, also in line with previous studies. We also observe that two cancer samples from human and mouse have very similar profiles and that is also the case for another pair of serum samples from the two species (Supplementary Figure S9). This may imply that ADAR edits for those particular conditions are preserved across these two species.

Finally, we have built the global maps of modification expression for all distinct types of modification and aggregated variants (Supplementary Figures S10 and 11) and we observe again that adenylation and uridylation are the most predominant modification types and they tend to occur sig-

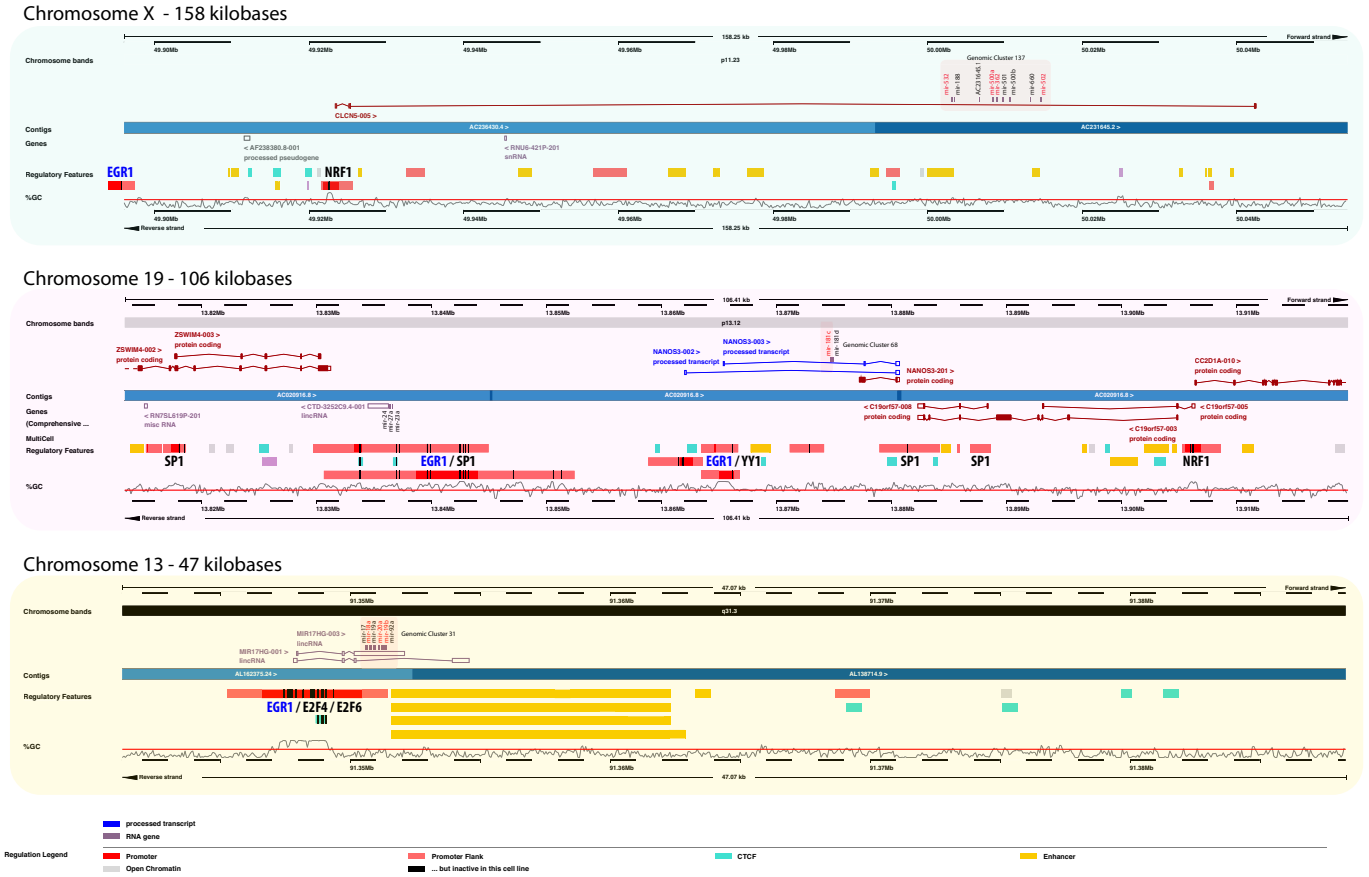

Figure 3. Genomic cluster example with transcriptional correlation of clusters from chromosomes 13, 19 and X, due to co-regulation mainly by the transcription factor EGR1, coupled with SP1 and NRF1.

nificantly more frequently at the 3′ end of miRNAs, both in human and mouse.

MicroRNA strand-specificity analysis and characterization

The dataset we obtain is extremely useful for exploring other aspects of miRNA biogenesis. In particular, because we obtain sequencing counts for both the 5′ and 3′ strands from a miRNA precursor we can use these data to globally explore mature strand selection of miRNAs. During the miRNA maturation process in general only one strand of the miRNA duplex is assembled into RISC while the complementary strand is degraded. This phenomenon has been studied in the past and the prevailing theory is that the asymmetry in the selection of the dominant miRNA strand may be explained by the difference in the stability of the bonds of the miRNA duplex at 5′ ends of each strand. This hypothesis has been proved experimentally for a small number of miRNAs (22). However, there is no global analysis so far that evaluates and models strand selection for miRNAs. We sought to both test these hypotheses and extract a global model of strand-selection for miRNAs based on the Gibbs free energies (ΔG) of the bonds present in the double stranded pre-miRNA.

During the formation of a double stranded RNA molecule, low ΔG values indicate that the reaction can occur spontaneously and lead to a stable form. Conversely, high ΔG values, calculated with reference to a ds-RNA segment, indicate high likelihood for that segment to unwind without the intervention of an external energy source.

For all miRNAs, we calculated the ΔG free energies for short double stranded segments of their hairpin structures around the 5′ end of the miRNA from each strand. We tested a variety of definitions for these segments. In each case, the window used for the definition of the segments focuses on a ds-RNA region of the hairpin, starting upstream, downstream or right at the 5′ end of each mature miRNA and extending for N (N≥1) nt overall toward the 3′ end of the miRNA. Specifically, we calculated the ΔG for each segment starting at the 5′ end of the 5′ mature product (ΔG1) and at the 5′ end of the 3′ mature product (ΔG2) and set their difference as $\Delta\Delta G = \Delta G2 - \Delta G1$ (Figure 6A.i).

Based on expression data from this analysis, all let-7 family miRNAs turn out to be very highly 5′-strand specific. Looking closely at the secondary structure of the let-7 family hairpin precursors, let us assume that the 5′ end regions of the 5′ miRNA products are more unstable than the corresponding ends of the 3′-miRNA products (e.g. due to prevalence of A:U bonds, gaps or wobbles). So, based on the conventions for the calculated free energies we used before, we would expect that ΔG1 > ΔG2, since ΔG1 refers to a more unstable structure. As a result, we would expect that ΔΔG

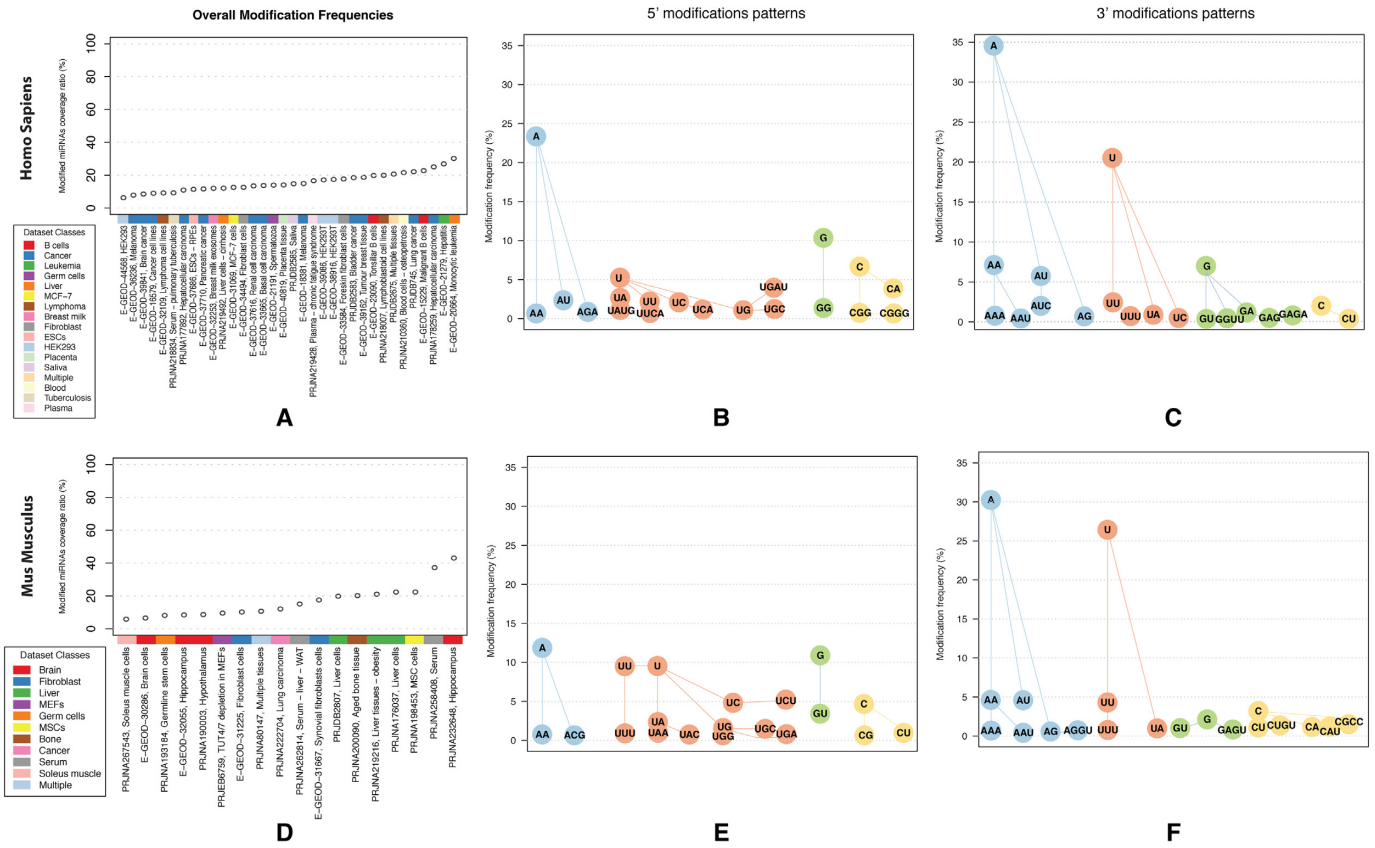

Figure 4. Overall extent of modification events and most dominant patterns. (**A** and **D**) Modification ratios coverage across all human and mouse datasets. (**B** and **E**) Prevalence of top-20 most frequent modifications patterns at the 5′ end of miRNAs in human and mouse. (**C** and **F**) Prevalence of of top-20 most frequent modifications patterns at the 3′ end of miRNAs in human and mouse.

< 0 for the 5′-strand specific miRNAs, $\Delta\Delta G > 0$ for the highly 3p-strand specific miRNAs and $\Delta\Delta G \approx 0$ for the non-strand-specific miRNAs.

In order to test our hypothesis, we classified all miRNAs based on their strand specificity. For this analysis, only miRNAs with two mature products, one for each strand of the hairpin precursor, have been taken into account. We first calculated the expression ratio of each miRNA strand product using the following formula:

$$expression\_ratio_{(arm)} = \frac{counts_{(arm)}}{counts_{(arm)} + counts_{(compl\_arm)}} \quad (1)$$

where:

- $(arm, compl\_arm) = (3p, 5p)$ or $(5p,3p)$
- $counts_{(arm)}$: is the total normalized depth of the '(arm)' mature miRNA product across all datasets and
- $counts_{(compl\_arm)}$: is the total normalized depth of the '$(compl\_arm)$' mature miRNA product, at all possible loci of the genome (see Methods).

Based on the *expression_ratio* scores calculated using the formula above, we grouped all miRNA precursors into three groups:

- highly 5′-strand specific: $a \leq expression\_ratio_{5p} \leq b$
- highly 3′-strand specific: $a \leq expression\_ratio_{3p} \leq b$
- non-strand specific: $0.4 \leq (expression\_ratio_{5p} \parallel expression\_ratio_{3p}) \leq 0.6$

for different sets of increasing strand specificity thresholds: (a,b) = {(0.7, 0.85), (0.85, 0.93), (0.93, 0.97), (0.97, 1)}.

We then test our hypothesis by calculating the $\Delta\Delta G$ values for all three types of strand specific groups with reference to a different segment of the ds-RNA hairpin structure each time. We have used increasing strand specificity thresholds for the highly 5′ and 3′ strand specific groups in order to examine if there is any shift in the $\Delta\Delta G$ values as the strand specificity criteria become more stringent. Moreover, we checked if the $\Delta\Delta G$ values from each strand specific group were distinguishable for the other groups implying that free energies calculated for a specific window of a ds-RNA hairpin segment are correlated with the strand selection process. Gibbs free energies have also been calculated for an additional group of 1000 'random' miRNAs. This group of 'random' miRNAs is formed by selecting randomly 10 non-strand specific miRNAs identified in our study and generating for each of them 100 permutations of their hairpin precursor sequences, permitting only permutations that fold to hairpin-like structures in the end.

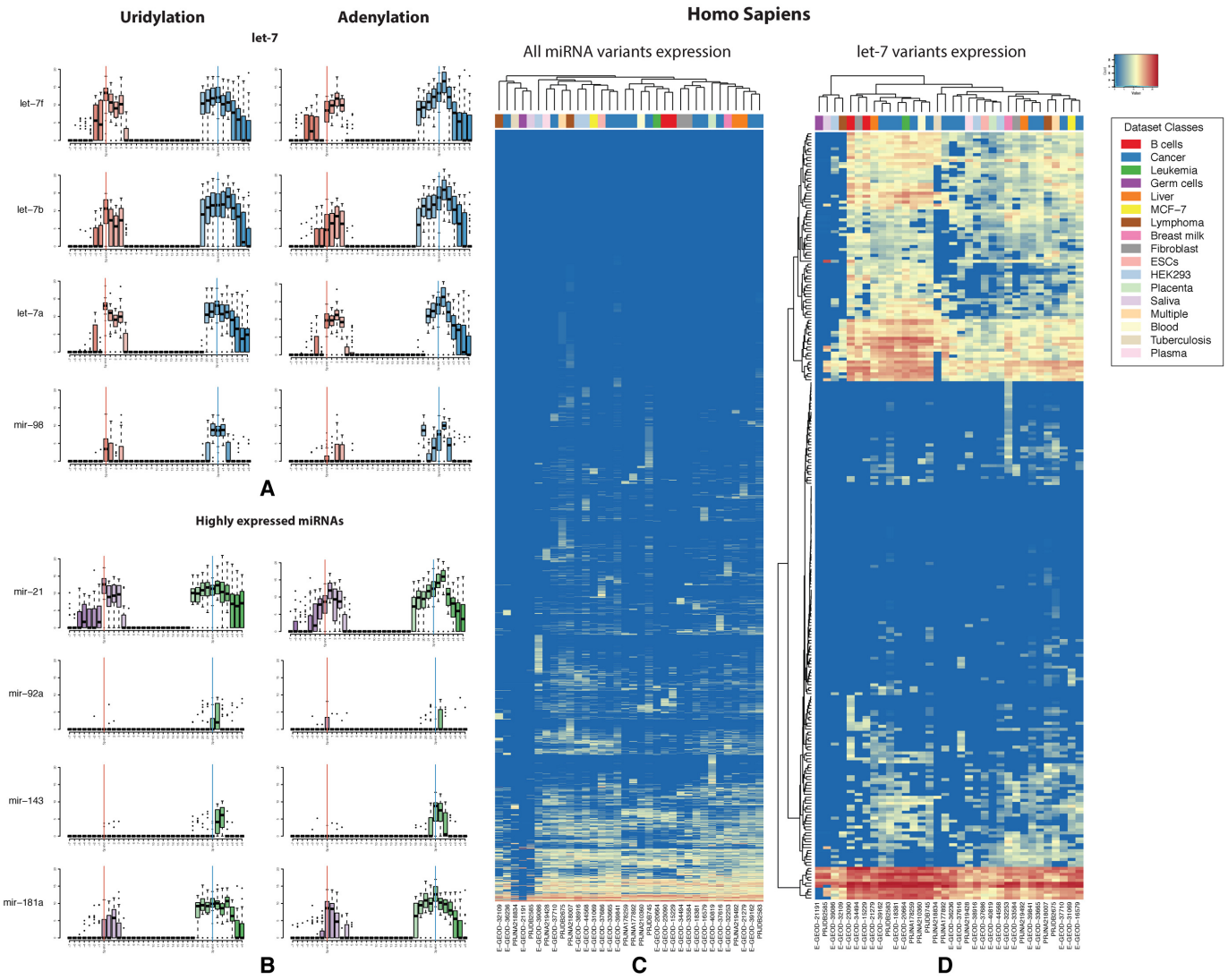

Figure 5. Modification analysis in human based on collapsed modification variants (i.e. wild-type and mono/poly-: adenylated, uridylated, gunaylated, cytocylated variants). (**A**) Distribution profiles of uridylated and adenylated miRNA variants for a subset of the let-7 family across all human datasets. (**B**) Distribution profiles of uridylated and adenylated variants for a set of four highly expressed miRNAs across all human datasets. (**C**) Expression profiles of all miRNA collapsed modification variants across all human datasets. (**D**) Expression profiles of the let-7 family miRNA collapsed modification variants across all human datasets.

After rigorous testing, we identified that there is a strong separation between the highly 5′-strand specific, highly 3′-strand specific and the non strand specifics groups of miRNAs only when the $\Delta\Delta G$ is calculated using a window that contains the first $N = 2$ nucleotides of each strand (Figure 6A.i). In this case, the $\Delta\Delta G$ distribution complies remarkably well with the assumption we have made earlier with regards to the expected $\Delta\Delta G$ values for different types of miRNAs in terms of their strand specificity (Figure 6A.ii–v). Besides, the group of random miRNAs, that is used as a control to examine the variance of $\Delta\Delta G$ across a large number of hypothetical hairpins, follows quite precisely the $\Delta\Delta G$ profile of the non-strand specific miRNAs that originate from real hairpins. Hence, these results indicate that, in general, the stability of the first 2 nt at the 5′ end of each strand plays the most crucial role for mature miRNA strand selection.

Furthermore, we made another observation that refers to the $\Delta\Delta G$ calculations for a window of 3 nt, starting at the 3′ end of each miRNA strand and extending for an extra nt in both sides (Figure 6B.i). This window contains nucleotides that are not present in the ds-RNA that is extracted to the cytoplasm but exist only in the hairpin precursor. However, we can notice again that there is a quite clear, although milder than in the previous case, separation of the miRNAs based on the strand specificity of their mature products (Figure 6B.ii–v,). In this case though, $\Delta\Delta G$ distribution for the highly 5′-strand and 3′-strand specific is reversed compared to the previous distribution. That may indicate that the unstable 2 nt long ds-RNA segment at the 5′ end of each strand is reinforced by the adjacent 3nt long ds-RNA seg-

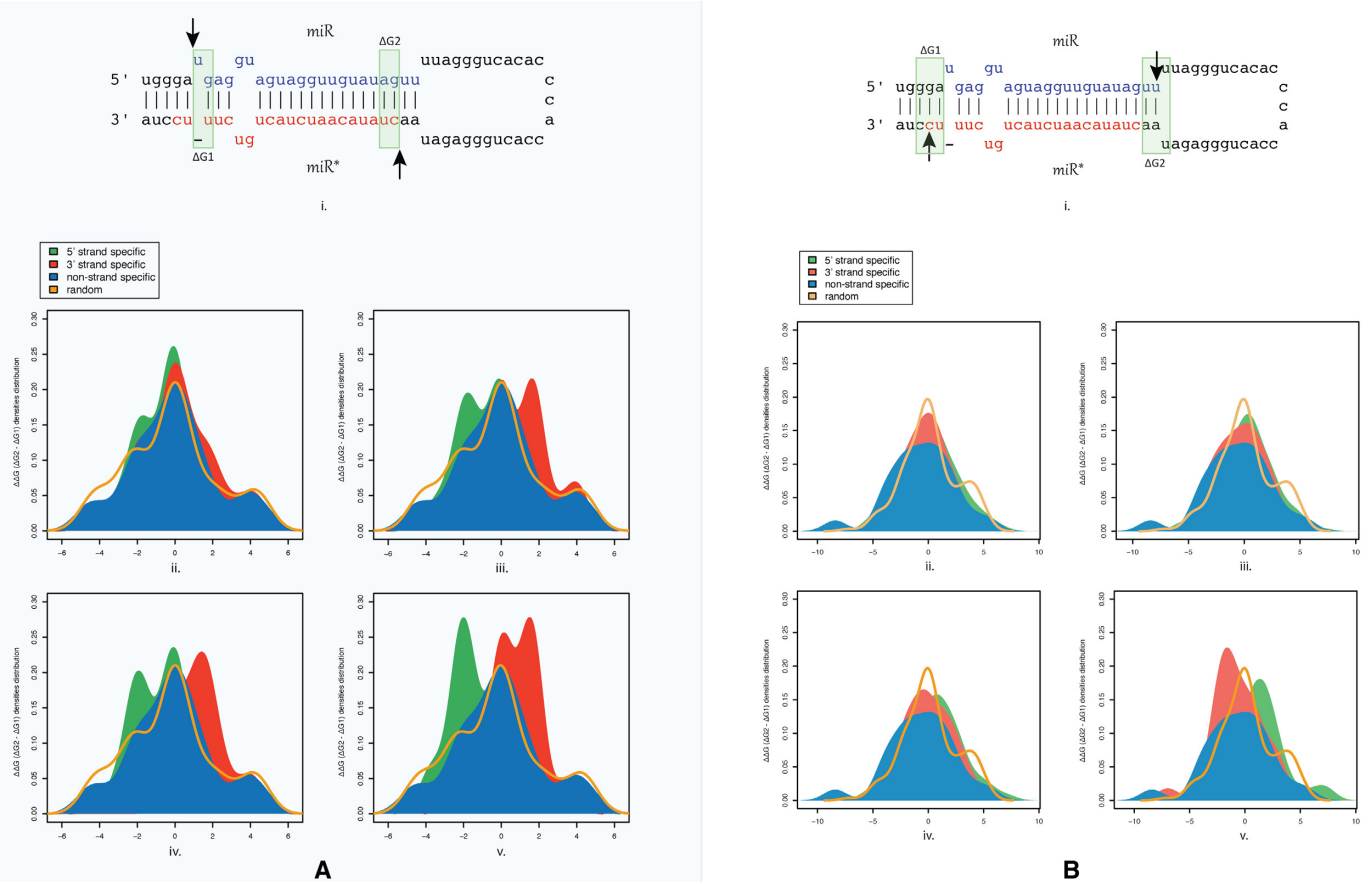

Figure 6. Separation of strand specificity classes based on free energies difference. (**A**) Optimal classes segregation based on free energy calculations at the 5′ ends of the two potential strand products. (**A**.i) Window used for the free energy calculations that leads to a clear separation of the strand specificity classes based on the energy difference: $\Delta\Delta G = \Delta G2 - \Delta G1$. The start of the window is set to the first nucleotide of the 5′ end of both strands and the window extends for an extra nucleotide toward the rest of each miRNA strand. (**B**) Free energy calculations at a 3 nt long ds-RNA segment adjacent to the optimal window. (**B**.i) 3-nt long window starting with the 2-nt overhang at the 3′ end of each strand used for the ΔG calculations in the hairpin sequences. The start of the window is set to the first nucleotide of the 2-nt overhang at the 3′ end of each strand and the window extends for another two nucleotides toward the 3′ end of the whole hairpin in both cases. (**A** and **B**)ii–v: ΔΔGs densities for the groups of the highly 5′-strand specific, highly 3′-strand specific, non-strand specific miRNAs and random miRNAs, by progressively setting stricter criteria of strand specificity, based on the $expression\_ratio_{5p/3p}$ values (Equation 1). (ii) $0.7 \leq expression\_ratio_{5p/3p} < 0.85$, (iii): $0.85 \leq expression\_ratio_{5p/3p} < 0.93$, (iv) $0.93 \leq expression\_ratio_{5p/3p} < 0.97$, (v) $0.97 \leq expression\_ratio_{5p/3p} \leq 1$.

ment that contains the 2 nt overhang at the 3′ end of the complementary strand. So, the asymmetry of miRNA duplexes in their 5′ ends is balanced by an opposite asymmetry in their 3′ ends which contributes to the preservation of the hairpin structure energy equilibrium.

Detection of mis-annotated miRNAs

The data obtained clearly shows that miRNAs have distinct patterns of expression, modification, strand-selection and genomic localization. The many hundreds of thousands of miRNA to precursor alignments obtained from NGS data also allow us to detect miRNAs which do not appear to illustrate the hallmarks of well characterized miRNAs. Previous reports have described many such molecules present in the *miRBase* database and suggest they represent mis-annotated sequences likely derived from other non-coding RNAs or degradation products of longer molecules (e.g. tRNAs). We used the alignment data obtained to automatically scan for miRNAs whose alignments and modification profiles did not fit those of well characterized miRNAs. We identify 22 Human and 21 Mouse miRNAs whose profiles clearly differ from miRNAs such as let-7 (Supplementary Figures S12 and S13). Scanning this set of miRNAs against miRBase shows that 11 of 43 identified miRNAs show similarity to annotated non miRNA molecules in the Rfam database (23). These miRNAs together with a comparison of miRNAs whose provenance is well established, e.g. via northern blot (Supplementary Figure S14), is also available in *miratlas*.

Finally, we wanted to examine if the prevalent form of miRNAs expressed in the miratlas datasets is equivalent with the miRBase canonical annotation. Specifically, we are interested in miRNAs whose predominant form detected in our data was longer or shorter than the annotated version. Thus, we reanalyzed all human and mouse miratlas registered samples and extracted the average length of the expressed template miRNA sequences across them, normalized by their overall expression depth. We then calculated the difference in length between each expressed miRNA and

its corresponding annotated sequence in miRBase (Supplementary Figure S15). We detected that for both Human and Mouse the prevalent form of expressed miRNAs is on average 1nt shorter than the accepted canonical sequence in miRBase. We also detect a small number of miRNAs which appear to be longer than their annotated mature sequence. Examples of both are shown (Supplementary Figures S16 and 17).

CONCLUSION

We present a comprehensive analysis of miRNA expression across multiple tissues and cell lines in Human and Mouse. These data are derived from high-throughput sequencing experiments from public resources. We have used these data to build a comprehensive miRNA expression dataset for Human and Mouse that takes into account both expression levels and detected modifications to miRNAs (e.g. 3′ uridylation or ADAR editing). This combined data resource allows us to explore the complex features of miRNA transcription across tissues and to group miRNAs into clusters based on their expression correlation. Additionally, we use these data to explore the likely transcriptional coupling of miRNAs in co-expressed clusters. We explore in detail, for the first time, the prevalence of both 5′ and 3′ nucleotide modifications to miRNAs and show that mono and dinucleotide 3′ modifications are the primary modifications observed in both human and mouse, with ADAR editing mostly restricted to brain and cancer cell types. Finally, we use these data to build a thermodynamic model for how the mature strand of a miRNA precursor is selected by exploring structural constraints around the ends of miRNA precursors derived from large-scale NGS data. We believe these findings and associated data will be of benefit to our understanding of miRNA function in animals and will also prove useful to the miRNA community in general.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank members of the Enright and Marioni laboratories (EMBL) for interesting and useful discussions and support.

FUNDING

EMBL core funding; MRC Methodology Research Fellowship [MR/L012367/1 to M.P.D.]. Funding for open access charge: EMBL core funding.
Conflict of interest statement. None declared.

REFERENCES

1. Tuschl,T., Zamore,P.D., Lehmann,R., Bartel,D.P. and Sharp,P.A. (1999). Targeted mRNA degradation by double stranded RNA in vitro. *Genes Dev.*, **13**, 3191–3197.
2. Lewis,B.P., Shih,I.H., Jones-Rhoades,M.W., Bartel,D.P. and Burge,C.B. (2003) Prediction of mammalian MicroRNA targets. *Cell.*, **115**, 787–798.
3. Krol,J., Loedige,I. and Filipowicz,W. (2010). The widespread regulation of microRNA biogenesis, function and decay. *Nat. Rev. Genet.*, **11**, 597–610.
4. Landgraf,P., Rusu,M., Sheridan,R., Sewer,A., Iovino,N., Aravin,A., Pfeffer,S., Rice,A., Kamphorst,A.O., Landthaler,M. *et al.* (2005) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414.
5. Griffiths-Jones,S., Kaur Saini,H., van Dongen,S. and Enright,A. (2008). mirbase: tools for microrna genomics. *Nucleic Acids Res.*, **36**, D154–D158.
6. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**(Database issue), D991–D995.
7. Davis,M.P., van Dongen,S., Abreu-Goodger,C., Bartonicek,N. and Enright,A.J. (2013). Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods*, **63**, 41–49.
8. Vitsios,D.M. and Enright,A.J. (2015) Chimira: analysis of small RNA sequencing data and microRNA modifications. *Bioinformatics*, **31**, 3365–3367.
9. Heo,I., Joo,C., Kim,Y.K., Ha,M., Yoon,M.J., Cho,J., Yeom,K.H., Han,J. and Kim,V.N. (2009). TUT4 in concert with Lin28 suppresses mi- croRNA biogenesis through pre-microRNA uridylation. *Cell*, **138**, 696–708.
10. Blow,M.J., Grocock,R.J., van Dongen,S., Enright,A.J., Dicks,E., Futreal,P.A., Wooster,R. and Stratton,M.R. (2006) RNA editing of human microRNAs. *Genome Biol.*, **7**,R27.
11. Love,M., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
12. van Dongen,S. and Abreu-Goodger,C. (2012) Using MCL to extract clusters from networks, in bacterial molecular networks: methods and protocols. *Methods Mol. Biol.*, **804**, 281–295.
13. Andronescu,M., Aguirre-Hernández,R., Condon,A. and Hoos,H.H. (2003) RNAsoft: a suite of RNA secondary structure prediction and design software tools. *Nucleic Acids Res.*, **31**, 3416–3422.
14. Freier,S.M., Kierzek,R., Jaeger,J.A., Sugimoto,N., Caruthers,M.H., Neilson,T. and Turner,D.H. (1986) Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. U.S.A.*, **83**, 9373–9377.
15. Krawetz,S.A., Kruger,A., Lalancette,C., Tagett,R., Anton,E., Draghici,S. and Diamond,M.P. (2011) A survey of small RNAs in human sperm. *Hum. Reprod.*, **26**, 3401–3412.
16. Suh,N., Baehner,L., Moltzahn,F., Melton,C., Shenoy,A., Chen,J. and Blelloch,R. (2010) MicroRNA function is globally suppressed in mouse oocytes and early embryos. *Curr. Biol.*, **20**, 271–277.
17. Saini,H.K., Griffiths-Jones,S. and Enright,A.J. (2007) Genomic analysis of human microRNA transcripts. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 17719–17724.
18. Triboulet,R., Chang,H.M., Lapierre,R.J. and Gregory,R.I. (2009). Post-transcriptional control of DGCR8 expression by the Microprocessor. *RNA*, **15**, 1005–1011.
19. 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.
20. Katoh,T., Sakaguchi,Y., Miyauchi,K., Suzuki,T., Kashiwabara,S., Baba,T. and Suzuki,T. (2009) Selective stabilization of mammalian microRNAs by 3′ adenylation mediated by the cytoplasmic poly(A) polymerase GLD-2. *Genes Dev.*, **23**, 433–438.
21. Hibio,N., Hino,K., Shimizu,E., Nagata,Y. and Ui-Tei,K. (2012) Stability of miRNA 5′terminal and seed regions is correlated with experimentally observed miRNA-mediated silencing efficacy. Scientific Reports, 2, 996.
22. Schwarz,D.S., Hutvágner,G., Du,T., Xu,Z., Aronin,N. and Zamore,P.D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, **115**, 199–208.
23. Nawrocki,E.P., Burge,S.W., Bateman,A., Daub,J., Eberhardt,R.Y., Eddy,S.R., Floden,E.W., Gardner,P.P., Jones,T.A., Tate,J. *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, **43**, D130–D137.