

RESEARCH PAPER



miRBaseMiner, a tool for investigating miRBase content

Xiangfu Zhong , Fatima Heinicke , and Simon Rayner *

Department of Medical Genetics, Oslo University Hospital and University of Oslo, Oslo, Norway

ABSTRACT

microRNAs are small non-coding RNA molecules playing a central role in gene regulation. miRBase is the standard reference source for analysis and interpretation of experimental studies. However, the richness and complexity of the annotation is often underappreciated by users. Moreover, even for experienced users, the size of the resource can make it difficult to explore annotation to determine features such as species coverage, the impact of specific characteristics and changes between successive releases. A further consideration is that each new miRBase release contains entries that have had limited review and which may subsequently be removed in a future release to ensure the quality of annotation. To aid the miRBase user, we developed a software tool, miRBaseMiner, for investigating miRBase annotation and generating custom annotation sets. We apply the tool to characterize each release from v9.2 to v22 to examine how annotation has changed across releases and highlight some of the annotation features that users should keep in mind when using for miRBase for data analysis. These include: (1) entries with identical or very similar sequences; (2) entries with multiple annotated genome locations; (3) hairpin precursor entries with extremely low-estimated minimum free energy; (4) entries possessing reverse complementary; (5) entries with 3' poly(A) ends. As each of these factors can impact the identification of dysregulated features and subsequent clinical or biological conclusions, miRBaseMiner is a valuable resource for any user using miRBase as a reference source.

ARTICLE HISTORY

Received 16 January 2019
Revised 29 April 2019
Accepted 20 June 2019

KEYWORDS

Microrna; miRNA; miRBase; miRBaseMiner; annotation; characterization; NGS

Introduction



MicroRNAs (miRNAs) are a class of non-coding small RNAs with an average length of ~22 nucleotides [1]. They post-transcriptionally regulate gene expression in both animals and plants [2,3] via binding to the 3'UTR (untranslated region) of mRNAs, but some miRNAs can also regulate the expression of other miRNAs [4] and even themselves [5]. While there are multiple pathways by which miRNAs can be produced [6], they are most commonly generated via the canonical Drosha/Dicer pathway. This begins with transcription of a primary miRNA (pri-miRNA) that is cleaved into precursor hairpin (pre-miRNA) sequences. These pre-miRNAs are subsequently exported from the nucleus to the cytoplasm where they are cleaved by Dicer into a duplex of approximately 22 nucleotides. One (or both) of the strands is then loaded into the RNA-induced silencing complex (RISC) and guided to target one or more mRNAs [7].

With the identification of the regulatory potential of miRNAs, it was proposed that they should be defined in terms of a set of expression and biogenesis criteria [8] (1) expression criteria: detection of the miRNA-like transcript in a cDNA library and a hybridization experiment; (2) biogenesis criteria: a stable (i.e. a low free minimum energy (MFE)) predicted hairpin precursor structure; (3) accumulation of precursors under the condition of Dicer knockout or inhibition.


MiRNA annotation in miRBase

The miRNA registry represented the first attempt to collect miRNA data in a single location [9], and was subsequently developed into miRBase [10–15] to catalogue miRNA annotation (i.e., sequence and, when appropriate, genome location), pre-miRNA related hairpin sequence and supporting experimental evidence in a standard format. However, there remains variation in the quality of entries and different interpretations as to what defines a miRNA. Consequently, there have been several attempts to produce new reference sources. Some of these (e.g. RNAcentral [16,17] and miRCarta [18]) attempt to extend the existing annotation, whereas others (e.g. MirGeneDB [19]) attempt to refine annotation by classifying entries as either true positives or false positives according to a range of different criteria. However, the relative merits of these different criteria are an ongoing discussion as each can produce misclassification of miRNAs that are known to be functional. Moreover, as these analyses focus on a specific release of miRBase, this makes it difficult to compare findings. However, all of these approaches rely on miRBase for source annotation and the resource remains the most widely used and highly cited miRNA repository with more than 15 000 citations to date.

Thus, there is no straightforward way for a researcher to (i) compare his or her findings with studies using a different annotation repository, or (ii) generate a custom annotation

CONTACT Simon Rayner  simon.rayner@medisin.uio.no  Department of Medical Genetics, Oslo University Hospital and University of Oslo, Oslo, Norway

*Present address: Hybrid Technology Hub - Centre of Excellence, Institute of Basic Medical Sciences, P.O. Box 1110 Blindern, 0317 OSLO, Norway

 The supplementary data for this article can be accessed [here](#).

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

set, for example, removing potentially less reliable new entries that only have passed the first stage of review by the miRBase team. Consequently, users generally use the full annotation set in their analysis. Another point is the impact of using different versions of miRBase. For example, the various releases of the Affymetrix miRNA GeneChip print a full miRNA release on their arrays and Exiqon offer miRCURY™ LNA™ microRNA ISH Detection Probes for all miRBase entries. Also, the miRQC study [20], which compared reproducibility and detection amongst different miRNA detection technologies, used three different versions of miRBase but failed to consider, for example, sequence differences that may exist among common entries across all three miRBase versions.

MiRNA nomenclature

miRBase entries generally follow a standard nomenclature, although this has evolved over time and with subsequent releases. The current miRNA nomenclature scheme is summarized in Figure 1(a). An entry name comprises a series of fields. The first three fields provide progressively more specific information: Field 1 is a three- to seven-letter code prefix indicating the species; Field 2 may be set to ‘miR’ or ‘mir’, corresponding to a miRNA and hairpin precursor entry, respectively; Field 3 is usually a numeric identifier. (in plants,

field 2 and field 3 are not separated by a hyphen). Additional fields are used to indicate: (i) entries with identical mature sequence but different parent hairpin precursors (numerical); (ii) the same mature sequence from the same hairpin precursor sequence but at a different genome location (lower case letter); (iii) a 5p/3p suffix to specify the arm of origin. In earlier miRBase releases, a ‘*’ was used to indicate the miRNA was the minor product of the precursor maturation process and not functional. However, as product from both arms has been subsequently shown to be functional [7] this suffix is gradually being retired from miRBase, although its usage continues in the literature [21–26]. Additionally, the species code has been modified over time to reflect changes in the accepted species name; for example, the species code for *Capitella teleta* (NCBI-taxid: 283909) was updated from **cap** to **cte** in version 15 (to reflect the change from ‘*Capitella* sp’ to ‘*Capitella teleta*’ as the commonly accepted name). Similarly, *Merkel cell polyomavirus* (NCBI-taxid: 493803) was changed from **mcv** to **mcpv** in version 22.

Tracking changes in miRBase

Understanding the changes that have occurred between successive miRBase releases provides insight into how the resource has evolved. The current miRBase release includes a ‘change log’ for each hairpin precursor summarizing the

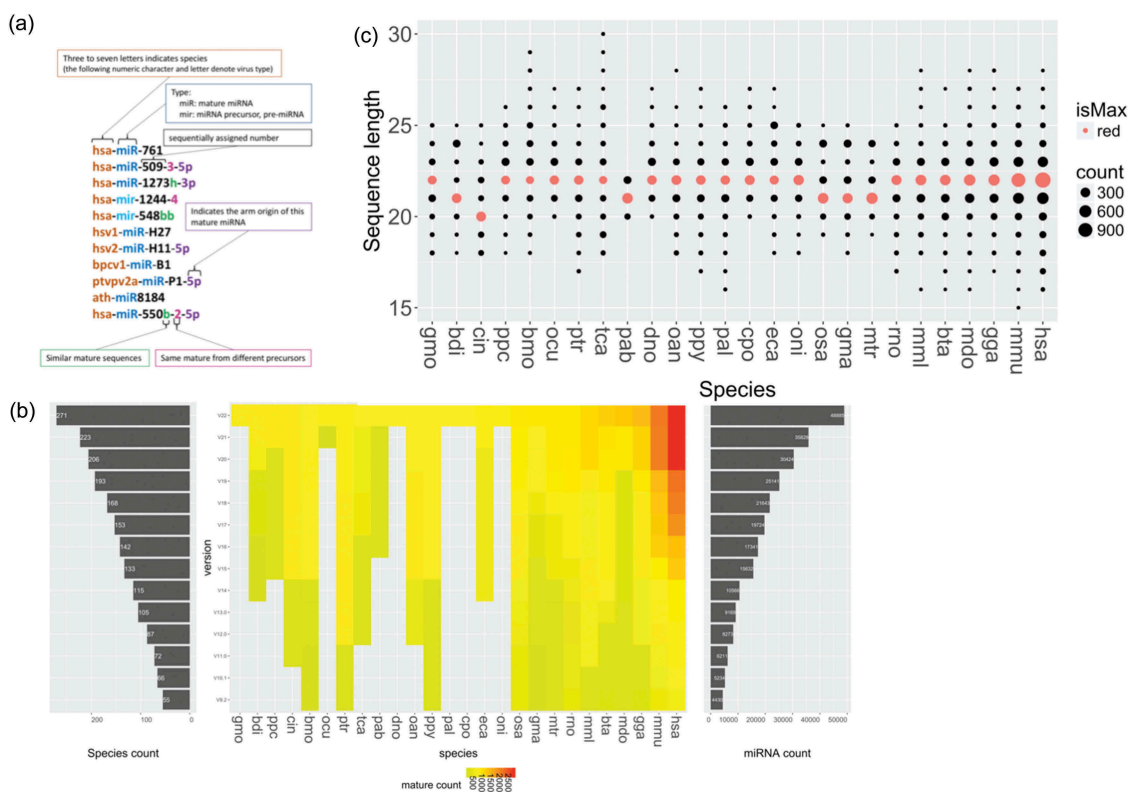


Figure 1. (a) Schematic of miRNA nomenclature used in miRBase release 22. Each entry contains the following fields delimited with a hyphen: (1) three to seven letters indicating species; (2) miR/mir indicating miRNA and miRNA hairpin precursor, respectively; (3) numeric suffix that is assigned sequentially to new entries (in plant miRNAs there is no hyphen delimiter between this field and the species field). Additional letters indicate mature miRNA sequence is shared by entries with the same numeric suffix, and additional numbers indicate miRNA is generated from different hairpin precursors; (4) 3p/5p indicates from which arm of the hairpin precursor the miRNA was generated. (b) Overview of content in successive releases of miRBase (Y-axis corresponds to miRBase releases from 9.2 to 22). Left: Bar plot showing number of annotated species (X-axis) in each miRBase release. Middle: Heat map of number of miRNA entries for the 26 species with more than 500 entries in miRBase v22. X-axis corresponds to the 26 species, ordered by total number of miRNAs for each species. Only a few species contain a large number of miRNAs (in red); Right: Bar plot showing total number of miRNA entries summed over all species (X-axis). (c) Sequence length distribution of miRNAs from the 26 species in (b). The average miRNA length is 21 ~ 22 nucleotides but many entries are shorter or longer than this.

change history across releases for that entry. However, this feature is not yet available for miRNAs and the information is only available via the ftp site and is not amenable to visual inspection. Third party tools are also available to help users investigate changes. miRBase Tracker [27], miRsystem [28], miRiadne [29], miRBaseConverter [30], miRNANameConverter [31] and miRBase Tracker [27] investigate changes between releases to provide identify nomenclature inconsistencies and sequence changes in updated releases. However, they do not provide more comprehensive details for each change.

Motivation

miRBase is the standard reference for miRNAs. However, many users underappreciate the richness and complexity of the annotation or consider how it may impact their data analysis or experimental design. While there are several publications that report specific annotation characteristics, there is no single publication that provides an integrated overview. Additionally, each new release contains many additions, as well as removal or revision of existing entries. As different technologies and analyses may use different versions it is important to readily aware of the changes between versions. However, as the resource has grown, it is becoming more difficult to track these changes by examining annotation files, for example, to determine the possible impact of using two different miRBase versions. Currently, there is no way to explore the resource in an automated way to examine specific features or subsets of the annotation. To this end, we developed miRBaseMiner.

Results

miRBaseMiner, a python package for investigating miRBase annotation content

miRBaseMiner is a Python pipeline, with accompanying R scripts for data visualization. Using miRBaseMiner, a user can specify a range of miRBase versions to be investigated and the pipeline will download the associated annotation files from the miRBase website. These data can then be used to characterize and compare versions from the following perspectives: (1) basic features of miRNAs and hairpin precursors, such as sequence length and number of annotated entries; (2) estimated minimum free energy (MFE) distribution; (3) nucleotide composition pattern at the 5' and 3' ends of miRNA sequences; (4) sequence similarity; (5) reverse complementary; (6) annotated genome coordinates. To demonstrate usage and to gain a more comprehensive understanding of miRBase content, we used miRBaseMiner to investigate miRBase annotation in successive releases from v9.2 to 22, the current release.

Overview of miRBase

The evolution of miRBase releases in terms of species content and number of entries is summarized in Figure 1(b). Species coverage has increased considerably, from 55 species in version 9.2 to 271 in version 22 (Figure 1(b) left panel) with 48 new species added in the most recent release. Similarly, the

Table 1. Example of differences in miRNA length and read length filtering used in the miRNA studies.

miRNA description	Read length filtering
~22 nt ^{1, 2}	19-20 nt ³
~23 nt ⁴	≤ 25 nt ⁵
21-23 nt ⁶	19-26 nt ⁷
~19- 24 nt ⁸	16-30 nt ¹⁰
21 -24 nt ⁹ (plant)	16-27 nt ¹¹
~20-22 nt ¹²	> 17 nt ¹³
17-23 nt ¹⁴	
18-24 nt ¹⁵	
20-24 nt ¹⁶	
19-22 nt ¹⁷	

total number of miRNA entries in miRBase has increased dramatically; the first miRBase release contained 218 mature miRNAs and 218 hairpin precursor sequences from five species, version 22 contains 48,885 mature miRNAs and 38,589 precursors from 271 species (Figure 1(b) right panel). However, there is only a small number of species annotated with large numbers of miRNAs (Figure 1(b) middle panel, areas in red; the plot for all species is shown in Supplementary Figure 1). Supplementary Figure 2 shows a word cloud representation of species abundance in terms of miRNA entries. *Homo sapiens*, *Mus musculus*, *Gallus gallus*, *Monodelphis domestica* and *Bos taurus* are the five most annotated species with more than 1000 miRNAs. Only 45% of species are annotated with more than 100 miRNAs.

MiRNA length range

Figure 1(c) shows the miRNA length distribution for all species that have more than 500 annotated miRNAs in release 22. The peak of the length distribution is 21 or 22 nucleotides for most species, but there are some exceptions such as *cin*: (*Ciona intestinalis*, NCBI-taxid: 7719, total miRNA count:550) which has a peak at 20 nucleotides. Also, even when a species has a maximum at 21 or 22 nucleotides, there are many miRNA entries that are longer than 25 nucleotides, such as *hsa-miR-7161-3p/MIMAT0028233* (28 nucleotides) and *mmu-miR-7222-5p/mmu-miR-7222-5p* (27 nucleotides). The longest entry in miRBase 22 is *hco-miR-5971/MIMAT0023478*, with a length of 34nt, the shortest entry is 15 nucleotides (e.g. *mmu-miR-7238-3p/MIMAT0028445*). The shortest human entry is 16 nucleotides (eight entries, e.g. *hsa-miR-4279/MIMAT0016909*).

As miRNAs longer than 26nt [32] are dismissed as false positives by many studies [33], we also considered the high confidence set of human miRNAs released by miRBase and found this also contains a length range of 18 to 28nt. Moreover, MirGeneDB, which performs the most conservative filtering of miRBase entries by only retaining miRNAs that are conserved across species, contains miRNAs with a length range of 20 to 27. Table 1 shows the different windows used for miRNA length in various studies. The lack of consensus among the studies highlights the challenge associated with defining miRNAs based on length range.

Characteristics of precursor hairpin entries in miRBase

Although the hairpin precursor entries in miRBase are generally based on secondary structure predictions and not

necessary representative of the true pre-miRNAs, they are commonly used as references for applications such as knock-out studies or positive data for pre-miRNA prediction tools. We therefore examined the properties of these precursor hairpin entries.

In the miRNA biogenesis model, a single hairpin will generate one mature miRNA from both the 3p and 5p arm of the pre-miRNA. However, there are cases where the same miRNA is generated from more than one pre-miRNA, either from an identical pre-miRNA sequence found in two genome locations, or from two distinct pre-miRNAs [48]. In this case, multiple pre-miRNAs generate a single identical miRNA. This can be relevant in a biological context; for example, when trying to knock down a pre-miRNA, as it is important to ensure the correct pre-miRNA is being targeted.

A stable hairpin structure for the pre-miRNA is a prerequisite for miRNA maturation [49,50]. The estimated minimum free energy (MFE) is a widely used measure of the stability of RNA secondary structure [51] and miRNA precursors have a considerably lower MFE compared to predicted secondary structures of transfer RNAs and ribosomal RNAs [52].

Since *Homo sapiens* (*hsa*) and *Mus musculus* (*mmu*) are the two most abundant species in miRBase, we investigated the MFE distributions of their hairpin precursors. While there is no obvious trend between average MFE across all entries and miRBase release, there is a gradual decrease in average stability of entries from version 9.2 to 22 for both of these species (Supplementary Figure 4), together with a reduction in mean sequence length and increase in mean GC content. Finally, the MFE index [53] has increased slightly both in human and mouse precursors in recent releases (Supplementary Figure 4). These variations may be a consequence of relaxation of what represents a stable structure [54,55]. The variation is not an indication of any annotation error, but any analysis based on these entries should correct for these variations.

A set of miRNAs contain poly(A) at their 3' end

MirBaseMiner can also investigate sequence composition in miRBase entries. This can be useful for experimental design and data processing. For example, Poly(A) tailing is a method used in library preparation for small RNA sequencing [56] and the presence of 3' adenosines in the miRNA sequence can confound the read trimming step in Next Generation Sequencing (NGS) studies. In miRBase v22, 11,213 out of 48,885 miRNAs terminate with an A in their sequence, and 109 of these have a 3' AAAA tail, including four human miRNAs (*hsa-miR-1468-3p/MIMAT0026638*, *hsa-miR-5009-3p/MIMAT0021042*, *hsa-miR-559/MIMAT0003223* and *hsa-miR-6128/MIMAT0024611*) and one mouse miRNA (*mmu-miR-691/MIMAT0003470*). Moreover, two mouse miRNAs, *mmu-miR-706/MIMAT0003496* and *mmu-miR-7116-5p/MIMAT0028129*, have AAAAA 3'tails. The presence of these poly(A) miRNAs should be kept in mind when performing adapter trimming on sequencing reads generated from poly(A) based library kits enrichment method and when interpreting downstream analysis results.

How entries change between successive miRBase releases

In addition to adding newly identified miRNAs, each new miRBase release also contains updates to existing entries based on new evidence (commonly deep sequencing datasets). Updated entries in a new release are placed within one or more of four categories: (1) NEW: newly added entries; (2) DELETE: entries in previous release removed from current release; (3) SEQUENCE: sequence of entry updated; (4) NAME: entry name affiliated with unique accession ID is updated. These changes are specified in the '*miRNA.diff*' file (available from the miRBase ftp site) that accompanies each miRBase release (after miRBase v3.1). Using miRBaseMiner to parse the update files for releases 9.2 to 22, we obtained an overview of the type of updates that defined each new release in terms of these four categories. The frequencies of these four update categories are summarized in Figure 2(a), for annotation in human and mouse from miRBase v9.2 to 22.

Sequence changes between successive releases

Every miRBase release has been accompanied by sequence modifications in existing entries across multiple species. We used miRBaseMiner to obtain a summary of the sequence changes that occurred in human and mouse from miRBase release 9.2 to 22. The results are shown in Figure 2(a). The highest number of sequence changes (103) occurred between release 21 and 22. The sequence changes can be grouped into five types: (1) 3' change; (2) 5' change; (3) 5' and 3' change; (4) internal nucleotide substitution; (5) no clear similarity (Figure 2b).

5' changes (nine human miRNAs and eight mouse miRNAs in miRBase 22, Figure 2 (b)) are more functionally significant as they change the miRNA seed sequence, which plays a critical role in targeting [57,58]. miRNA targets are commonly based on computational predictions and these in turn depend on the miRBase release on which the predictions are based. While the numbers of changes are relatively few compared to the total number of miRBase entries in our chosen species, they become relevant when occurring in an entry that has been previously identified as being of significance in a study. For example, *hsa-miR-146b-5p/MIMAT0002809* gained a 3' nucleotide extension between version 21 and 22, despite being associated with regulatory roles in leukemogenesis [59] and atherosclerosis [60,61] and has been patented for treatment and as a diagnostic biomarker [62,63]. These findings were based on microarray and RT-PCR experiments, which in turn are dependent on miRNA sequence. There are also many cases of deleted miRNAs being reported in studies, for example, *hsa-miR-4520b-5p/MIMAT0020299*, *hsa-miR-3669/MIMAT0018092* and *hsa-miR-3673/MIMAT0018096* were removed in version 21 (released June 2014) but have been subsequently reported in miRNA studies published after this date [64,65]. Finally, *hsa-mir-941-1/MI0005763*, *hsa-mir-941-2/MI0005764*, *hsa-mir-941-3/MI0005765* and *hsa-mir-941-4/MI0005766* had sequence changes in miRBase version 16 and again in version 20. A full list of changes in human and mouse miRNAs and hairpin precursor sequence annotations from version 9.2 to 22 is given in Supplementary Table 11.

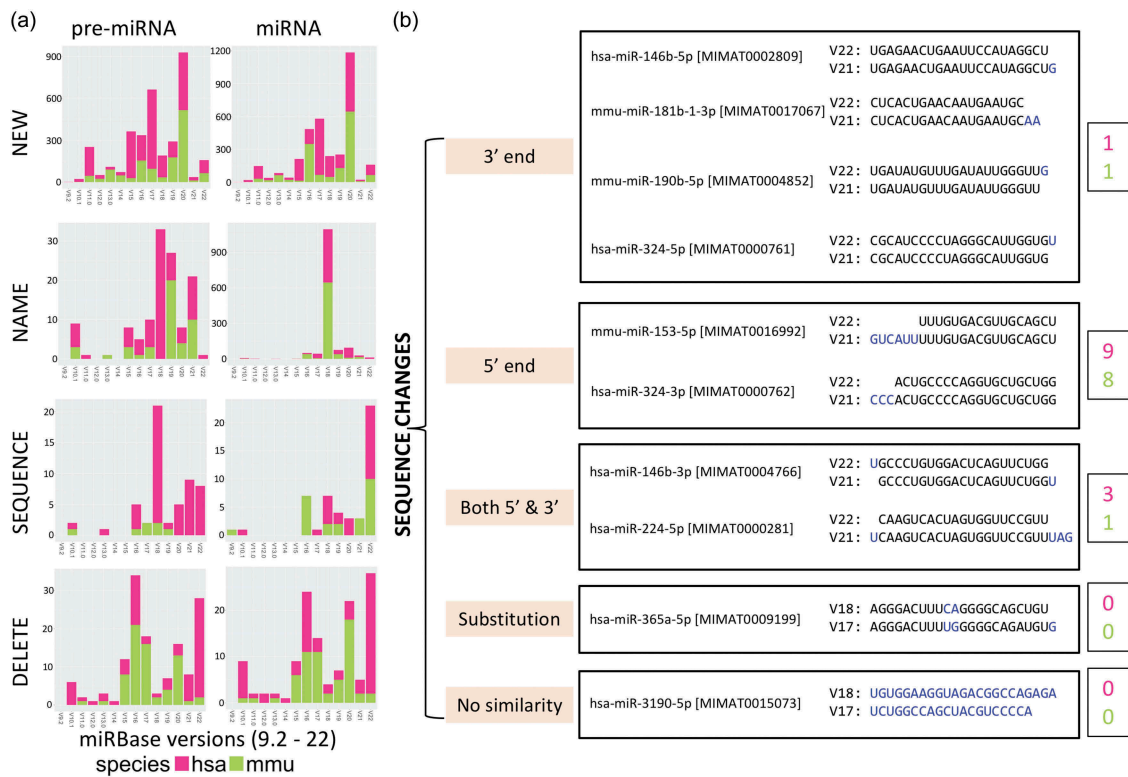


Figure 2. (a) Bar plot showing the number of updated miRNAs (right column) and pre-miRNA (left column) entries that were updated between subsequent versions of miRBase from release 9.2 to 22. The rows correspond to four categories: NEW, NAME, SEQUENCE and DELETE. In each plot, X-axis: miRBase versions in chronological order. Y-axis: the number of updated miRNAs/hairpin precursors. Red corresponds to human data and green for mouse. (b) Five types of sequence changes that occur between successive miRBase releases. First three examples are changes between release 21 and 22. Last two examples are changes between miRBase version 17 and 18. Nucleotides changes in sequences between two releases are marked in blue. The bottom sequence is the original miRNA sequence in the previous version of miRBase, the top sequence is in the newer release. The number after sequence box represents the frequency of corresponding miRNA sequence changes that occurred in miRBase 22.

A subset of entries can add additional complexity to miRNA studies

One of the most common applications of the miRBase annotation is miRNA expression studies. These studies typically apply microarray or NGS technology to identify differentially expressed miRNAs (or additionally, in the case of NGS studies, seek identification of novel miRNAs). Individual findings are then verified by techniques such as qPCR. However, certain annotation features should be considered in these analyses and subsequent experimental studies. In particular, users need to be aware of entries that share common or highly similar sequences, as mapping settings in the mapping software (such as allowed number of mismatches and how to share reads across equivalent locations) can affect how reads are counted across features sharing identical or highly similar sequences. From a biological perspective, downstream studies such as sequencing of flanking regions can be problematic as identical miRNAs may differ in these regions.

A set of miRNAs share full-length identical sequences

The philosophy behind miRBase is that each miRNA is associated with a sequence, and the genomic location and sequence together represent a unique entry. However, when two or more than two miRNA entries have distinct names, accession numbers, and genome locations but identical

sequence, they will be indistinguishable both from a functional (i.e. targeting) and discovery perspective and users should be aware of this when mapping reads to reference sequences. Zhao *et al.* [66] identified 38 distinct human miRNA entries annotated with 15 unique sequences in miRBase 21. We therefore used miRBaseMiner to search for the presence of similar entries in versions 9.2 to 22 for human and mouse annotation. The results are summarized in Figure 3(a,b).

Sequence duplication is present in both human and mouse annotation and the current release contains 16 unique sequences shared by 40 human miRNAs, and 12 unique sequences shared by 27 mouse miRNAs. Sequence duplication also occurs in hairpin precursor entries in both human and mouse annotations. A full list of sequence degenerate miRNAs and hairpin precursor is given in *Supplementary Table 6*. This is not an annotation problem, but users should keep these entries in mind when analysing and interpreting miRNA data, particularly in the context of NGS experiments.

Overlapping sequences among miRNA entries

An additional potential challenge is that longer miRNA entries can overlap shorter miRNAs. This is significant for the same reasons as above – sequencing mapping tools will not be able to uniquely place a read that is a perfect match to

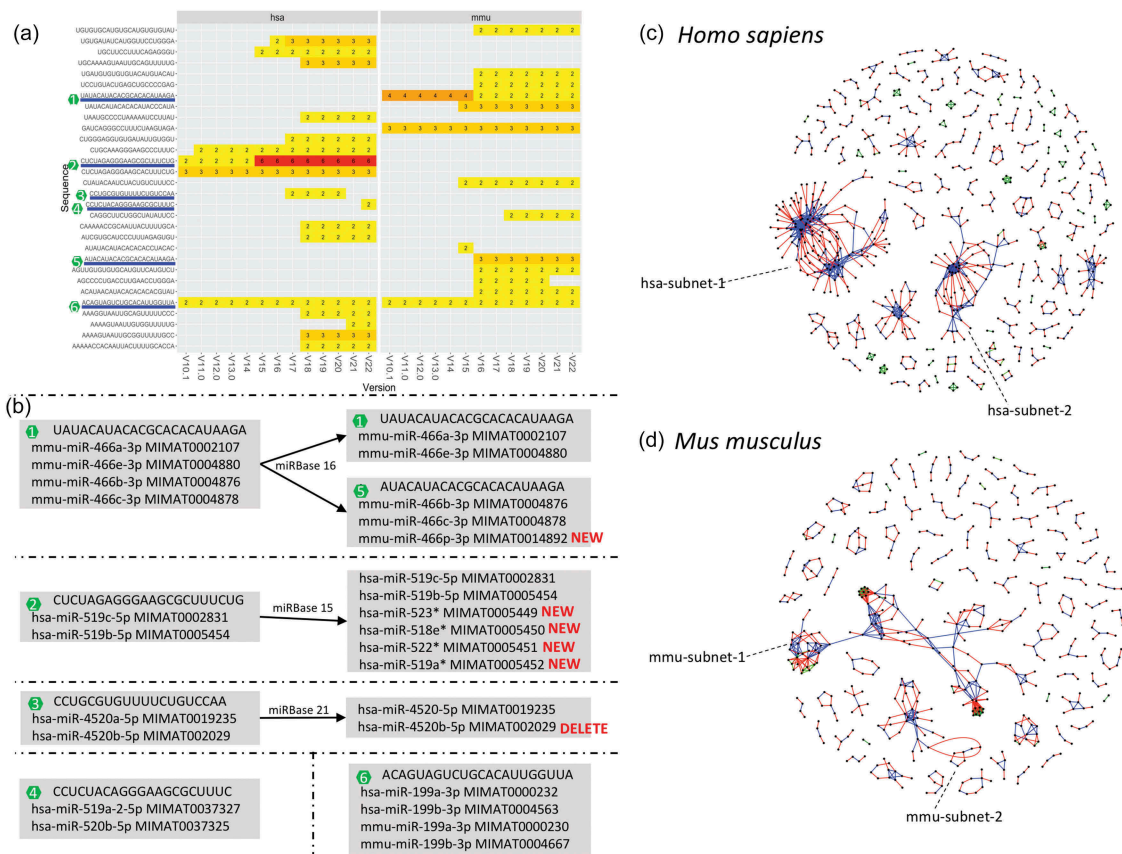


Figure 3. The presence of identical sequences in human and mouse miRNA entries from miRBase version 9.2 to 22 (A, B), and sequence similarity in miRBase 22 (C, D). (a) Left: human miRNAs; Right: mouse miRNAs. In miRBase 9.2, there are no miRNAs in human or mouse sharing identical sequence with other entries. The colour denotes the number of miRNAs annotated with that sequence, from yellow to red indicating increasing number. The number in each cell represents the number of miRNA entries with the sequence in that row and the miRBase entry (corresponding to that column). X-axis indicates the miRBase version; Y-axis indicates the duplicated miRNA sequence. (B) Examples of miRNAs sharing identical sequence. The number in green hexagon refers to the corresponding row in (A). The text in red upper case indicates the type of annotation change; NEW: newly added miRNA; DELETE: miRNA entry deleted from miRBase. The text above arrows indicates miRBase versions in which that change occurred. Right-hand plots. The similarity network of human and mouse miRNAs and hairpin precursors in miRBase version 22 based on the pairwise Levenshtein distance matrix for Levenshtein distances less than three nucleotides. (C) human miRNA and hairpin precursor network; (D) mouse miRNA and pre-miRNA network. Each dot represents a miRNA or hairpin precursor. Blue edge: two similar miRNAs; red edge: pre-miRNA and its respective miRNA; green edge: two similar hairpin precursors. The darker color (blue/green) corresponds to a Levenshtein distance equal to 0, the lighter colour corresponds to larger Levenshtein distances.

a miRNA that is a subsequence of a longer miRNA. For example, the sequence read ‘AAAAGUAAUUGCGGUCUUU’ is a perfect match to *hsa-miR-548ap-5p/MIMAT0021037* (AAAAGUAAUUGCGGUCUUU – 19 nucleotides) but is also a perfect match to *hsa-miR-548j-5p/MIMAT0005875* (AAAAGUAAUUGCGGUCUUUGGU – 22 nucleotides). This problem is further exacerbated by the presence of isomiRs [67]; although miRBase does not currently consider isomiRs, from a sequence perspective, *hsa-miR-548ap-5p/MIMAT0021037* and *hsa-miR-548j-5p/MIMAT0005875* can also be considered as isomiR variants rather than distinct miRNAs. This problem arises for the same reasons as outlined above for the degenerate families and, while renaming conflicting entries would solve the problem, this would make it difficult to link results using old and new names. Thus, any read mapping step needs to consider the possibility of overlapping entries. The exact way in which this issue is handled depends both on the mapping tool and the selected parameters, and will accordingly lead to different read counts, possibly impacting the subsequent interpretation of results.

A list of overlapping human miRNAs is summarized in Supplementary Table 7 and 8.

The similarity of mature miRNAs in miRBase

We also used miRBaseMiner to calculate and visualize the similarity of miRNA/hairpin precursor sequences in miRBase for human and mouse, respectively, by calculating the pairwise Levenshtein distance [68] between all miRNA sequences and hairpin precursor sequences in release 22. The graphs for human and mouse are shown in Figure 3 C&D for Levenshtein distances up to three nucleotides (the maximum number of mismatches allowed by the bowtie mapping tool [69]).

Each node in the network graph represents a miRNA or pre-miRNA, with the colour of each edge indicating the relationship between two nodes and a darker edge (in blue or green) indicating higher sequence similarity. The nodes that are highly connected by blue edges represent highly similar miRNAs/hairpin precursors. Nodes connected by green edges correspond to highly similar hairpin precursors and form more isolated networks. An interesting feature of

both graphs is the red edges that connect miRNAs with their pre-miRNAs. Rather than forming isolated networks centred on a specific hairpin precursor, there are hairpin precursors/miRNAs that form interconnected networks. The graph also reveals that miRNAs can originate from more than one hairpin precursor.

The Levenshtein graphs for human and mouse are quite distinct. In human, similar hairpin precursors form multiple local networks, whereas the mouse network contains two large clusters of similar precursors networked with mature miRNAs. In both graphs, the clustered miRNAs originate from host-specific miRNA families. For human, hsa-subNet-1 is associated with the mir-548 family. For mouse, mmu-subNet-1 is a mixed network primarily containing two sub-clusters of mir-466 and mir-467 family members that are connected by precursors mmu-mir-467a and mmu-mir-669a, respectively. The overlapping miRNAs described in the previous section will not be captured completely by the Levenshtein graph if there is a total of more than three additional nucleotides beyond the overlapping region. While the origins of these differences in the primary differences between the two graphs are understood, if not taken into account then they have the potential to produce distinct effects during the mapping step.

Some miRNA pairs share reverse complementarity

As miRNAs entries for human and mouse are almost equally distributed on the forward and reverse strands, a mapping tool must search both strands of a reference genome. However, Zhao *et al.* [66] reported nine pairs of human miRNAs in miRBase 21 that share reverse complementarity (RC). For example, *hsa-miR-4433b-5p/MIMAT0030413* and *hsa-miR-4433a-3p/MIMAT0018949* are annotated on opposite strands at chr2:64,340,809–64,340,829. A mapping tool cannot assign a definite location in this situation as a read will be randomly assigned to either strand, leading to potential read counting problems for the RC pair. RC miRNAs were identified in miRBase as early as version 5.0 and 6.0 when two entries, *hsa-mir-104/MI0000110* and *hsa-mir-108/MI0001432*, were removed by miRBase due to reverse complementarity to other annotated entries. Based on this and Zhao's findings, we analysed miRBase releases from v9.2 to v22 and identified an additional RC miRNA pair in release 22 for *hsa-miR-7-5p/MIMAT0000252* (see Supplementary Table 3) as a consequence of a sequence change. Detailed information for these miRNAs in miRBase 22 is given in Supplementary Table 3, and a full list of RC pairs in all species and versions are given in Supplementary Table 9 and Supplementary Table 10. Palindrome could not be fully explained by the reverse complementary entries listed in Supplementary Table 3.

We extended the analysis to all species in release 22 and found a total of 108 RC miRNAs in 17 species (including human) but, surprisingly, did not detect any RC miRNAs in mouse – see Supplementary Tables 9 and 10. Moreover, we found reverse complementarity in hairpin precursors in miRBase 22 (12 RC human hairpin precursors and 3 RC mouse hairpin precursors). It has been shown that two distinct miRNAs can be driven from both forward and reverse

strands of same genome location [1], which is the cases for the majority of these 108 miRNAs.

A group of miRNAs are missing or have multiple genome coordinates

Although a miRBase miRNA entry should contain sequence, genome coordinates and supporting evidence as minimal information, there are some entries lacking coordinate information and these miRNAs will not be profiled in approaches that map to a reference genome. In some cases, e.g. *hsa-mir-378g/MI0016761*, the entry previously had coordinates, but they were removed in the most recent releases (see Supplementary Table 4). This situation usually arises as the result of accumulation of deep sequencing data and an updated genome assembly. If a hairpin no longer maps to the new assembly, miRBase attempts to locate the new locus and update annotation accordingly. In the event, the new locus cannot be identified, and it appears that the region is missing in the new assembly, the coordinates will be removed. However, since the entry was present in the original assembly, it is unlikely it really is absent and, in general, it will be retained. However, there are some entries that had genome coordinates removed between versions despite using the same release. Additionally, there is a set of miRNAs and hairpin precursor entries that have multiple genome locations. Although this is biologically feasible, as these are present as paralogs that are processed to produce the same mature sequence, this can also be a problem when mapping to a reference genome, depending on the mapping tool [70]. There are 169 human miRNAs in this category, with the most extreme case occurring for *hsa-miR-1302/MIMAT0005890* with 11 annotated locations (i.e., identical accession number and name, but unique id). Both these effects occur for many different species. Release 22 also contains four hairpin precursors (three in human, one in mouse) with multiple coordinates. A full list of entries in these categories is given in Supplementary Table 12. Biologically, the presence of these multiple loci not surprising, but could represent a potential issue for both biologists in the wet lab and bioinformaticians working on sequencing data, and additional checks should be made to ensure these expression levels for these miRNAs are being correctly counted.

High confidence miRNA sets

miRBase provides a high confidence subset of hairpin precursors and miRNAs to represent entries that are most strongly supported by experimental evidence from NGS studies [14] (i.e., the datasets considered by miRBase) and which meet certain other criteria. These criteria include a requirement that reads must be present on both arms of a hairpin, and the hairpin has an MFE < -0.2 kcal/mol/nt.

The set of high confident entries is also dynamic, with additions and deletions within each new release as new evidence is accumulated. We therefore used miRBaseMiner to inspect and visualize these changes. The results are summarized in Supplementary Figure 8. In particular, there is a set of entries that were annotated as high confident in version 20, removed in version 21, and then added back in version 22. For example, *hsa-miR-3157-3p/MIMAT0019210* was

annotated as high confidence in miRBase 20 and 22, but not in release 21. This high confidence entry ‘hopping’ between versions is likely to be a consequence of (1) updates to high confidence annotation criteria and (2) an increased number of datasets [15]. Thus, while there are a total of 3298 distinct hairpin precursors across the three releases, only 925 (231 human hairpins) are present across all three releases. In addition, the high confident annotation also contains entries with characteristics such as reverse complementary and missing genome coordinate (see human annotation in Table 2, and mouse in Supplementary Table 5). A complete list of the history of high confident entries across all available releases is given in Supplementary Table 13. Once again, these are not annotation issues, but factors that users should be aware of when selecting an annotation set and performing their analyses.

Impact of annotation set on differential expression analysis

logFC (log fold change) and FDR (false discovery rate, adjusted p-value) or p-value are commonly used for selecting significant differentially expressed miRNAs between two conditions [37–74]. While multiple studies have examined the effect of various differential analysis tools on outcome, there has been no investigation of the effect of different miRNA annotation sets. We therefore selected and analysed a small RNA sequencing dataset using a standard analysis pipeline (miRDeep2 [75] and edgeR 3.18) with four different annotation reference sets (miRBase 21, miRBase 21 high confidence, miRBase 22 and miRBase 22 high confidence). These results are summarized in Supplementary Figure 11. While a difference in the total number of miRNAs in the reference set has a minor impact on logCPM and logFC values, a large difference impacts both the p-value and FDR. Thus, significant differences were observed in the sets of identified differentially expressed miRNAs between healthy and disease state in all pairwise comparisons of annotations, with the exception of miRBase 21 versus miRBase 22. Hence, caution is required in interpreting DE analysis results and comparing results among different studies.

Discussion

Introduction

miRNAs are one of the most widely studied non-coding RNA features, with more than 80 000 related entries in PubMed as of April 2019. The majority of studies focus on identifying miRNAs that are dysregulated between conditions and determining the impact of this dysregulation by prediction of their targets. A key part of these studies is access to reliable information about sequence and genome location and miRBase has evolved to become the standard reference source. However, many users fail to appreciate the complexity and richness of the annotation and how it may impact experimental design or data analysis. In particular, miRBase is a dynamic resource with addition, removal and revision of entries – for example, in release 22, the annotation for miRNAs from *mir-566*, *mir-1273*, *mir-4419* and *mir-6723* families are removed [15]. Additionally, as the resource has grown, more effort is required to understand how specific features or annotation change between releases.

There have been various reports that have investigated miRBase annotation and which have provided updated annotation sets [48,76–78]. However, these have focused on annotation within a specific release and have not considered how the resource has changed across multiple updates. In this report, we introduce a tool, miRBaseMiner for exploring miRBase annotation. We demonstrate the value of the tool by performing a comprehensive investigation of miRBase annotation. By considering changes in annotation from version 9.2 to the latest version 22 we are able to provide an overview of how the resource has evolved and how specific annotation characteristics should be considered when applying the information to experimental or analytical studies. We started with v9.2 as this annotation is the basis of the commercially available Miltenyi miRXPlore Universal Reference set commonly used for spike-in studies. We also considered the representation of all species in miRBase but we focused on human and mouse entries for some aspects of this study as they have significantly more annotation compared to any other species. Our results indicate that the annotation should

Table 2. Comparison between the content of the full set and the high confidence set of human entries for miRBase release 22.

Issues	Full miRBase entries	High_confidence miRBase entries
species representation in miRNAs	hsa: 2656/48,885 (5.43%) mmu: 1978/48,885 (4.05%)	hsa: 1198/3982 (30.09%) mmu: 1109/3982 (27.85%)
species representation in precursors	hsa: 1917/38,589 (4.97%) mmu: 1234/38,589 (3.20%)	hsa: 658/2162 (30.43%) mmu: 614/2162 (28.40%)
miRNAs with identical sequence	hsa: 40/2656 (1.51%) mmu: 27/1978 (1.37%)	hsa: 0/1198 (0%) mmu: 0/1109 (0%)
Reverse complementary miRNAs	hsa: 20/2656 (0.75%) mmu: 0/1978 (0%)	hsa: 12/1198 (1.00%) mmu: 0/1109 (0%)
Reverse complementary pre-miRNAs	hsa: 24/1917 (1.25%) mmu: 6/1234 (0.49%)	hsa: 15/658 (2.28%) mmu: 2/614 (0.33%)
miRNAs overlapping (including identical)	hsa: 71/2656 (2.67%) mmu: 43/1978 (2.17%)	hsa: 17/1198 (1.42%) mmu: 15/1109 (1.35%)
miRNAs with poly(A) tail (AAAA)	hsa: 4/2656 (0.15%) mmu: 3/1978 (0.15%)	hsa: 1/1198 (0.08%) mmu: 0/1109 (0%)
miRNAs (Levenshtein distance ≤ 3)	hsa: 348/2656 (13.10%) mmu: 278/1978 (14.05%)	hsa: 218/1198 (18.20%) mmu: 226/1109 (20.38%)
Pre-miRNAs (Levenshtein distance ≤ 3)	hsa: 111/1917 (5.79%) mmu: 60/1234 (4.86%)	hsa: 39/658 (5.93%) mmu: 46/614 (7.49%)
Hairpins are missing genome coordinates	hsa: 4/1917 (0.21%) mmu: 8/1234 (0.65%)	hsa: 0/1198 (0%) mmu: 1/1109 (0.09%)

be reviewed before applying miRNA profiling studies and a curated annotation set could often be appropriate.

In particular, users should be aware that some entries do not have coordinates, or share the same sequence with other entries. These are not annotation errors, but failing to take them into account can confound analysis of expression data in NGS studies. For example, entries missing coordinates would not be considered in approaches that map reads to a reference genome; for entries with identical sequence, the user should consider how reads will be mapped and counted by specific mapping tools and counting methods [70,79].

As miRBase has evolved, as well as the addition of newly identified miRNAs and hairpin precursors, some entries were removed, or the sequence coordinates or miRBase name was modified. These changes can make it difficult to compare results using different miRBase releases. For example, the Applied Biosystems miRNA GeneChip uses releases 15, 17 and 20 in versions 2.0, 3.0 and 4.0 of their chips, respectively. Also, as many NGS studies use the miRBase release that was most current at the time, care must be taken when comparing across studies and platforms.

In addition to investigating the evolution of miRBase, miRBaseMiner is also able to characterize entry features, which can assist the user when select analysis parameters. For example, miRNA lengths are often characterized as being from 19 to 26 nucleotides, and this information is sometimes used in a pre-filtering step either in library preparation to maximize reads or to avoid unnecessary read mapping [37,35,80]. However, given that (for example) human miRNA entries range from 16 nucleotides to 28 nucleotides, care should be taken using this approach. While miRNAs longer than 26 nucleotides are commonly dismissed as false positives, both high confidence annotation sets from miRBase as well as the highly conservative annotation generated by MirGeneDB support the presence of longer miRNAs. This is also highlighted in Table 1, which shows variation in miRNA length range used by in different analyses. An additional observation was the presence of many entries with multiple adenosines at the 3' end; for library preparation using poly(A) tailing, this may complicate the adapter trimming step. In such cases, detailed information from the manufacturers of regarding adapter trimming is needed to determine the impact on these entries. Again, none of these characteristics represent annotation errors, but they represent factors that should be considered in any analysis or experimental design.

True-positives and false-positives in miRBase

There have been many studies that have attempted to identify issues in miRBase annotation, one study estimated that as many as two-thirds of miRBase annotation are false-positives [33]. However, there is no consensus among these studies regarding the definition of a false positive, leading to wildly differing estimates of their presence. The purpose of miRBaseMiner is not to predict incorrect entries, but rather to provide the miRBase user with the tools to perform their own analysis of the resource to determine the impact of changes in the annotation.

Nomenclature

As outlined in the introduction, miRBase handles most of these characteristics by using a clearly defined nomenclature. For example, the nomenclature can distinguish miRNA entries with the same mature sequence but different hairpin precursors, from miRBase entries with the same mature and hairpin precursor sequences, but the precursor sequence originating from a different genome location. To aid miRBase users, miRBaseMiner can parse the miRNA and hairpin precursors to generate lists of entries falling within each category. Importantly, it can also identify annotation errors. For example, there are five hairpin precursors named *hsa-mir-3180-1* to *hsa-mir-3180-5*, which, according to the nomenclature will each generate a miRNA with identical sequence. However, miRBaseMiner reveals that only three of the hairpin precursors generate a mature product from the -5p arm (Supplementary Figure 10).

Mirna targeting

miRBase provides predicted targets for a miRNA entry via links to miRDB [81,82], TargetMiner [83] and TargetScan [84]. However, because these are static links into the respective databases, it is not apparent to which database release the link is pointing. For example, in miRBase 22, miRDB links currently point to entries based on miRBase v21 [81], TargetMiner links point to entries bases on the older (i.e. 2007) miRBase release [83], and TargetScan links point to entries bases on miRBase release 21 [84]. Thus, the miRBase entry for *hsa-mir-324/MI0000813* contains links to predicted targets for *hsa-miR-324-3p/MIMAT0000762* but the sequence was updated in release 22, adding three Cs at the 5 prime end, modifying the seed region and consequently the set of targets. None of the hard links in miRBase reflect this update. In addition to inconsistencies between miRNA annotation in the current miRBase and the link to predicted target profiles in targeting repositories, TargetScan links in miRBase v22 point to TargetScan release 7.1, which has been superseded by release 7.2. This can also lead to notable differences. For example, *hsa-miR-135a-5p/MIMAT0000428* has 843 transcripts containing 941 conserved sites and 321 poorly conserved sites in TargetScan release 7.1, but 715 transcripts containing 803 conserved sites and 293 poorly conserved sites in TargetScan release 7.2. Maintaining links to other databases is a common challenge that is faced by almost every database but we mention it here to remind users to check version information when using these target links.

Similar miRNAs, reverse complementary and mapping issues

miRBase allows the user to identify miRNAs sharing similar sequences via the nomenclature. However, while this identifies the similarity in miRNA families it cannot capture the relationship between entries outside these families sharing sequence similarity. Generating network graphs based on Levenshtein distance (Figure 3 (c,d)) provides a way to visualize these relationships and identifies clusters of highly similar miRNAs. In reality, the situation is even more complicated as the presence of isomiRs (not shown in the figure) will generate larger and more complex networks.

Understanding the topology of these networks is important as allowing mismatches during the mapping step will render miRNAs within the same Levenshtein network indistinguishable [85,86]. Thus, the choice reference source and mapping method need to be carefully considered; for example, when mapping to a reference genome, the impact of alignment settings, such as number of mismatches, can alter read counts on features. Entries sharing reverse complementary should also be considered when mapping to a reference genome as they represent duplicate mapping locations. One solution for miRNA or miRNA precursors is to perform the read mapping in two steps, in the first step mapping is to the forward strand, and in the second step, all remaining unmapped reads are mapped to the reverse strand (corresponding to the `- norc` and `- nowf` flags in the bowtie mapping tool). An example is given in the Supplementary Materials. The issue of overlapping entries also occurs in hairpin precursor sequences and is a further point that users need to consider. Some advantages and disadvantages of using different reference types (i.e. genome, precursor sequence or miRNA sequence) are summarized in Supplementary Table 1. While there is no standard recommendation for choosing a reference sequence, the table provides a guide based on the specific scientific question and highlights potential issues in mapping or result parsing that can affect the outcome.

Concluding comments

We conclude by emphasizing that the findings from our study are not intended in any way to be a criticism of miRBase or the approaches they use for annotation. Our goal is rather to provide a tool that can investigate miRBase annotation. In this way we can: (1) complement their efforts and attempt to make users aware that there may be consequences when a particular annotation is selected, i.e. the full entry set versus the high confidence set; (2) provide a pipeline that can be used to investigate the annotation and to characterize specific subsets (e.g. by species) in terms of different factors such as MFE, length distribution or degenerate (sequence) entries. (3) Allow the generation of custom GFF files based on filtering criteria using this pipeline.

Using miRBaseMiner, a user can compare annotation sets to investigate the stability of an analysis result (e.g. does a differentially expressed feature remain regardless of the annotation set, or is it sensitive to the choice of what is included in the annotation?). Thus, the user can achieve deeper insight into the consistency of an analysis result. Finally, to overcome the issues associated with name changes across different releases, we recommend that users provide miRNA name/accession number pair in their publication to avoid confusion. This could be in the form of a supplementary file listing all miRNA name/accession number pairs referenced in a manuscript and would be consistent with the MIBBI recommendations for data standards [87].

Material and methods

Data retrieval and parsing

The full data for each version (from 9.2 to 22) of miRBase were downloaded via its ftp site. Full details of data parsing and processing are provided in the Supplementary Materials.

Levenshtein distance

To estimate the similarity within the human and mouse sets of miRNAs and hairpin precursors, we calculated the pairwise Levenshtein distance (edit distance) [68] between each two miRNAs.

$$\begin{aligned} lev_{a,b}(i,j) &= f(x) \\ &= \begin{cases} \min = \begin{cases} \max(i,j), & \text{if } \min(i,j) = 0 \\ lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{a_i \neq b_j} \end{cases}, & \text{otherwise} \end{cases} \end{aligned}$$

where a and b are two strings (corresponding to miRNA_a and miRNA_b respectively) and $lev_{a,b}(i,j)$ is the distance between the first i characters (i.e. nucleotides) and $1_{a_i \neq b_j}$ is the Indicator function, equal to 0 when $a_i = b_j$ and equal to 1 otherwise.

Analysis pipeline

To systematically investigate miRNA and hairpin precursor annotation in miRBase releases, a pipeline (miRBaseMiner) was developed in Python. miRBaseMiner can analyse the full set of releases for all species, or the user can specify a subset of releases and species; if necessary, miRBaseMiner will download the required datasets from the miRBase FTP site. The data are then scanned to generate basic statistics for the specified releases and species. These include characteristics such as sequence length, GC content and MFE distributions. miRBaseMiner can also provide detailed analysis of specific characteristics of the miRNA and hairpin precursor entries content across releases. These comprise: (1) basic features within a release including sequence length and number of annotated entries; (2) estimated minimum free energy distribution; (3) tri-nucleotide composition pattern at both ends of the sequence and 3' Poly(A) tail analysis; (4) sequence similarity; (5) reverse complementary; (6) annotated genome coordinates. miRBaseMiner will use miRBase information whenever possible (e.g. GC content and MFE index). If this information is not available, then it will be calculated. A schematic of the workflow of miRBaseMiner is shown in Figure 4

Results for each characterization are written to a separate file and visualized using the R programming language (v3.5). miRBaseMiner can also generate a set of curated miRNA annotation entries based on user-defined filters obtain annotation that is best suited to the research question. For example, miRNA discovery and isomiR analysis face different challenges and will benefit from distinct annotation sets (Supplementary Table 1).

The software is available at: <https://github.com/joey0214/miRBaseMiner>

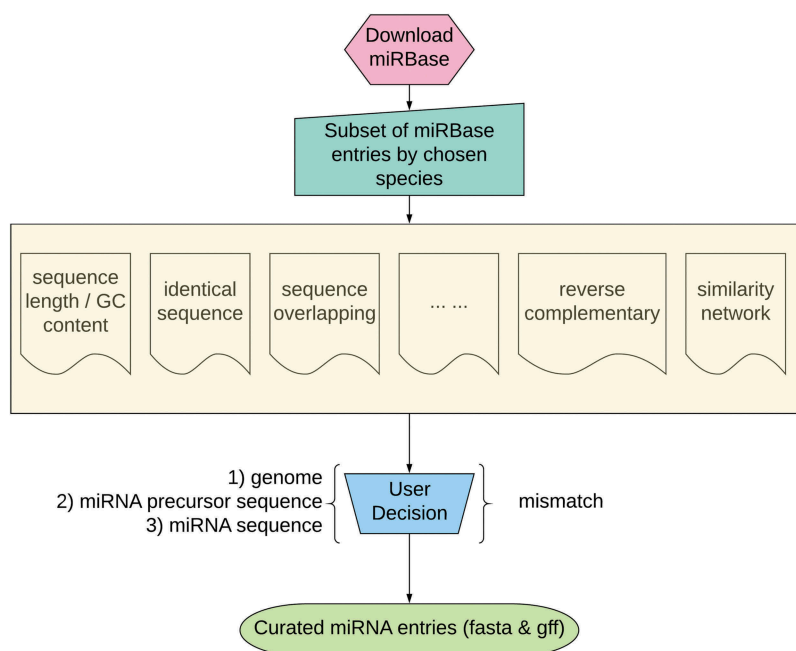


Figure 4. The workflow of miRBaseMiner.

Acknowledgments

We would like to thank Sam Griffiths-Jones for valuable discussions and insight into the philosophy and considerations behind the process of handling miRBase entries.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Helse Sør-Øst Grants [2016122, 2015034] and Norwegian Research Council Grant [274715].

ORCID

Xiangfu Zhong  <http://orcid.org/0000-0002-1872-1186>
 Fatima Heinicke  <http://orcid.org/0000-0001-8253-6105>
 Simon Rayner  <http://orcid.org/0000-0001-8703-9140>

References

- [1] Bartel DP. Metazoan MicroRNAs. *Cell*. 2018;173:20–51.
- [2] Lee RC, Feinbaum RL, Ambros V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*. 1993;75:843–854.
- [3] Wightman B, Ha I, Ruvkun G. Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell*. 1993;75:855–862.
- [4] Truscott M, Islam AB, Frolov MV. Novel regulation and functional interaction of polycistronic miRNAs. *RNA*. 2016;22:129–138.
- [5] Zisoulis DG, Kai ZS, Chang RK, et al. Autoregulation of microRNA biogenesis by *let-7* and Argonaute. *Nature*. 2012;486:541–544.
- [6] Ruby JG, Jan CH, Bartel DP. Intronic microRNA precursors that bypass Droscha processing. *Nature*. 2007;448:83–86.
- [7] Yang JS, Phillips MD, Betel D, et al. Widespread regulatory activity of vertebrate microRNA* species. *RNA*. 2011;17:312–326.
- [8] Ambros V, Bartel B, Bartel DP, et al. A uniform system for microRNA annotation. *RNA*. 2003;9:277–279.
- [9] Griffiths-Jones S. The microRNA Registry. *Nucleic Acids Res*. 2004;32:D109–11.
- [10] Griffiths-Jones S, Grocock RJ, van Dongen S, et al. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*. 2006;34:D140–4.
- [11] Griffiths-Jones S, Saini HK, van Dongen S, et al. miRBase: tools for microRNA genomics. *Nucleic Acids Res*. 2008;36: D154–8.
- [12] Griffiths-Jones S. miRBase: microRNA sequences and annotation. *Curr Protoc Bioinformatics* 2010; Chapter 12:Unit 12 9 1-0.
- [13] Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*. 2011;39: D152–7.
- [14] Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*. 2014;42:D68–73.
- [15] Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res*. 2019;47: D155–D162.
- [16] The RNAcentral Consortium. RNAcentral: an international database of ncRNA sequences. *Nucleic Acids Res*. 2015;43:D123–9.
- [17] PThe RNAcentral Consortium. RNAcentral: a comprehensive database of non-coding RNA sequences. *Nucleic Acids Res*. 2017;45:D128–D134.
- [18] Backes C, Fehlmann T, Kern F, et al. miRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Res*. 2018;46: D160–D167.
- [19] Fromm B, Billipp T, Peck LE, et al. A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. *Annu Rev Genet*. 2015;49:213–242.
- [20] Mestdagh P, Hartmann N, Baeriswyl L, et al. Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. *Nat Methods*. 2014;11:809–815.
- [21] Crossland RE, Norden J, Pearce KF, et al. Serum and extracellular vesicle microRNAs MiR-423, MiR-199 and MiR-93* as biomarkers for acute graft versus host disease. *Front Immunol*. 2017;8:1446.
- [22] Fei Q, Yu Y, Liu L, et al. Biogenesis of a 22-nt microRNA in Phaseoleae species by precursor-programmed uridylation. *Proceedings of the National Academy of Sciences* 2018; 115:8037–8042.

- [23] Bang C, Batkai S, Dangwal S, et al. Cardiac fibroblast-derived microRNA passenger strand-enriched exosomes mediate cardiomyocyte hypertrophy. *J Clin Invest*. 2014;124:2136–2146.
- [24] Xia J, Zhang W. A meta-analysis revealed insights into the sources, conservation and impact of microRNA 5'-isoforms in four model species. *Nucleic Acids Res*. 2014;42:1427–1441.
- [25] Dong Y, Chang C, Liu J, et al. Targeting of GIT1 by miR-149* in breast cancer suppresses cell proliferation and metastasis in vitro and tumor growth in vivo. *Onco Targets Ther*. 2017;10:5873–5882.
- [26] Roberto GM, Engel EE, Scrideli CA, et al. Downregulation of miR-10B* is correlated with altered expression of mitotic kinases in osteosarcoma. *Pathol Res Pract*. 2018;214:213–216.
- [27] Van Peer G, Lefever S, Anckaert J, et al. miRBase Tracker: keeping track of microRNA annotation changes. *Database (Oxford)*. 2014;2014:1–8.
- [28] Lu T-P, Lee C-Y, Tsai M-H, et al. miRSystem: an integrated system for characterizing enriched functions and pathways of microRNA targets. *PLoS One*. 2012;7:e42390.
- [29] Bonnal RJ, Rossi RL, Carpi D, et al. miRiadne: a web tool for consistent integration of miRNA nomenclature. *Nucleic Acids Res*. 2015;43:W487–W92.
- [30] Xu T, Su N, Liu L, et al. miRBaseConverter: an R/Bioconductor package for converting and retrieving miRNA name, accession, sequence and family information in different versions of miRBase. *BMC Bioinformatics*. 2018;19:514.
- [31] Haunsberger SJ, Connolly NM, Prehn JH. miRNANameConverter: an R/bioconductor package for translating mature miRNA names to different miRBase versions. *Bioinformatics*. 2016;33:592–593.
- [32] Chiang HR, Schoenfeld LW, Ruby JG, et al. Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev*. 2010;24:992–1009.
- [33] Fromm B, Billipp T, Peck LE, et al. A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. *Annu Rev Genet*. 2015;49:213–242.
- [34] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*. 2004;116:281–297. DOI:10.1016/S0092-8674(04)00045-5
- [35] Brown M, Suryawanshi H, Hafner M, et al. Mammalian miRNA curation through next-generation sequencing. *Front Genet*. 2013;4:145.
- [36] Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell*. 2009;136:215–233.
- [37] Baglio SR, Rooijers K, Koppers-Lalic D, et al. Human bone marrow- and adipose-mesenchymal stem cells secrete exosomes enriched in distinctive miRNA and tRNA species. *Stem Cell Res Ther*. 2015;6:127.
- [38] Kuo W-T, Ho M-R, Wu C-W, et al. Interrogation of microRNAs involved in gastric cancer using 5p-arm and 3p-arm annotated microRNAs. *Anticancer Research*. 2015;35:1345–1352.
- [39] Szczesniak MW, Makalowska I. miRNEST 2.0: a database of plant and animal microRNAs. *Nucleic Acids Res* 2014; 42:D74–7
- [40] Humphreys DT, Hynes CJ, Patel HR, et al. Complexity of murine cardiomyocyte miRNA biogenesis, sequence variant expression and function. *PLoS One*. 2012;7:e30933.
- [41] Rissland OS, Hong S-J, Bartel DP. MicroRNA destabilization enables dynamic regulation of the mir-16 family in response to cell-cycle changes. *Molecular Cell*. 2011;43:993–1004. DOI: 10.1016/j.molcel.2011.08.021
- [42] Kulland JB, Pinto DLP, Bertolini E, et al. Mirvine: a microRNA expression atlas of grapevine based on small RNA sequencing. *BMC Genomics*. 2015;16:393.
- [43] Luciano DJ, Mirsky H, Vendetti NJ, et al. RNA editing of a miRNA precursor. *RNA*. 2004;10:1174–1177.
- [44] Leite DJ, Ninova M, Hilbrant M, et al. Pervasive microRNA duplication in chelicerates: insights from the embryonic microRNA repertoire of the spider parasteatoda tepidariorum. *Genome Biology and Evolution*. 2016;8:2133–2144.
- [45] McCall MN, Baras AS, Crits-Christoph A, et al. A benchmark for microRNA quantification algorithms using the openarray platform. *BMC Bioinformatics*. 2016;17:138.
- [46] Burroughs AM, Ando Y, de Hoon MJ, et al. A comprehensive survey of 3' animal miRNA modification events and a possible role for 3' adenylation in modulating miRNA targeting effectiveness. *Genome Res*. 2010;20:1398–410.
- [47] Thomou T, Mori MA, Dreyfuss JM, et al. Adipose-derived circulating miRNAs regulate gene expression in other tissues. *Nature*. 2017;542:450.
- [48] Hansen TB, Kjems J, Bramsen JB. Enhancing miRNA annotation confidence in miRBase by continuous cross dataset analysis. *RNA Biol*. 2011;8:378–383.
- [49] Ng Kwang Loong S, Mishra SK. Unique folding of precursor microRNAs: quantitative evidence and implications for de novo identification. *RNA*. 2007;13:170–187.
- [50] Prabu GR, Mandal AK. Computational identification of miRNAs and their target genes from expressed sequence tags of tea (*Camellia sinensis*). *Genomics Proteomics Bioinformatics*. 2010;8:113–121.
- [51] Schuster P, Fontana W, Stadler PF, et al. From sequences to shapes and back: a case study in RNA secondary structures. *Proc Biol Sci*. 1994;255:279–284.
- [52] Bonnet E, Wuys J, Rouze P, Van de Peer Y. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*. 2004;20:2911–2917.
- [53] Zhang BH, Pan XP, Cox SB, et al. Evidence that miRNAs are different from other RNAs. *Cell Mol Life Sci*. 2006;63:246–254.
- [54] Creighton CJ, Benham AL, Zhu H, et al. Discovery of novel microRNAs in female reproductive tract using next generation sequencing. *PLoS One*. 2010;5:e9637.
- [55] Hu J, Lin C, Liu M, et al. Analysis of the microRNA transcriptome of *Daphnia pulex* during aging. *Gene*. 2018;664:101–110.
- [56] Berezikov E, Thuemmler F, van Laake LW, et al. Diversity of microRNAs in human and chimpanzee brain. *Nat Genet*. 2006;38:1375–1377.
- [57] Ellwanger DC, Buttner FA, Mewes HW, et al. The sufficient minimal set of miRNA seed types. *Bioinformatics*. 2011;27:1346–1350.
- [58] Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*. 2005;120:15–20.
- [59] Zhang HM, Li Q, Zhu X, et al. miR-146b-5p within BCR-ABL1-Positive Microvesicles Promotes Leukemic Transformation of Hematopoietic Cells. *Cancer Res*. 2016;76:2901–2911.
- [60] Raitoharju E, Lyytikäinen LP, Levula M, et al. miR-21, miR-210, miR-34a, and miR-146a/b are up-regulated in human atherosclerotic plaques in the Tampere Vascular Study. *Atherosclerosis*. 2011;219:211–217.
- [61] Takahashi Y, Satoh M, Minami Y, et al. Expression of miR-146a/b is associated with the Toll-like receptor 4 signal in coronary artery disease: effect of renin-angiotensin system blockade and statins on miRNA-146a/b and Toll-like receptor 4 levels. *Clin Sci (Lond)*. 2010;119:395–405.
- [62] Xu S. miRNAs as Novel Therapeutic Targets and Diagnostic Biomarkers for Parkinson's Disease. United States Patent NO. US9540692B2. 2017.
- [63] Bentwich I, Avniel A, Karov Y, et al. HSA-MIR-146B-5P-related nucleic acids and uses thereof. United States Patent NO. US9650679B2 2017.
- [64] Li J, Xu X, Guan H, et al. Exosome-derived microRNAs contributes to prostate cancer chemoresistance. In: Proceedings of the 107th Annual Meeting of the American Association for Cancer Research; New Orleans (LA). Philadelphia (PA): AACR; Cancer Res 2016;76(14 Suppl):Abstract nr 965. 2016 Apr 16–20.
- [65] Goto A, Dobashi Y, Tsubochi H, et al. MicroRNAs associated with increased AKT gene number in human lung carcinoma. *Hum Pathol*. 2016;56:1–10.

- [66] Zhao S, Gordon W, Du S, et al. QuickMIRSeq: a pipeline for quick and accurate quantification of both known miRNAs and isomiRs by jointly processing multiple samples from microRNA sequencing. *BMC Bioinformatics*. 2017;18:180.
- [67] Morin RD, O'Connor MD, Griffith M, et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res*. 2008;18:610–621.
- [68] Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. *Sov Phys Dokl*. 1966;10:707–710.
- [69] Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10:R25.
- [70] Johnson NR, Yeoh JM, Coruh C, et al. Improved placement of multi-mapping small RNAs. *G3 (Bethesda)*. 2016;g3:116.030452.
- [71] Shi H, Chen J, Li Y, et al. Identification of a six microRNA signature as a novel potential prognostic biomarker in patients with head and neck squamous cell carcinoma. *Oncotarget*. 2016;7:21579.
- [72] Li X, Shi Y, Yin Z, et al. An eight-miRNA signature as a potential biomarker for predicting survival in lung adenocarcinoma. *J Transl Med*. 2014;12:159.
- [73] Cascione L, Gasparini P, Lovat F, et al. Integrated microRNA and mRNA signatures associated with survival in triple negative breast cancer. *PloS One*. 2013;8:e55910.
- [74] Baglio SR, Devescovi V, Granchi D, et al. MicroRNA expression profiling of human bone marrow mesenchymal stem cells during osteogenic differentiation reveals Osterix regulation by miR-31. *Gene*. 2013;527:321–331.
- [75] Friedländer MR, Mackowiak SD, Li N, et al. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res*. 2011;40:37–52.
- [76] Desvignes T, Batzel P, Berezikov E, et al. miRNA nomenclature: a view incorporating genetic origins, biosynthetic pathways, and sequence variants. *Trends Genet*. 2015;31:613–626.
- [77] Meyers BC, Axtell MJ, Bartel B, et al. Criteria for annotation of plant MicroRNAs. *Plant Cell*. 2008;20:3186–3190.
- [78] Budak H, Bulut R, Kantar M, et al. MicroRNA nomenclature and the need for a revised naming prescription. *Brief Funct Genomics*. 2016;15:65–71.
- [79] de Hoon MJ, Taft RJ, Hashimoto T, et al. Cross-mapping and the identification of editing sites in mature microRNAs in high-throughput sequencing libraries. *Genome Res*. 2010;20:257–264.
- [80] Toedling J, Servant N, Ciaudo C, et al. Deep-sequencing protocols influence the results obtained in small-RNA sequencing. *PLoS One*. 2012;7:e32724.
- [81] Wong N, Wang X. miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res*. 2015;43:D146–52.
- [82] Wang X. Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. *Bioinformatics*. 2016;32:1316–1322.
- [83] Bandyopadhyay S, Mitra R. TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics*. 2009;25:2625–2631.
- [84] Agarwal V, Bell GW, Nam JW, et al. Predicting effective microRNA target sites in mammalian mRNAs. *Elife*. 2015;4:e05005.
- [85] Camps C, Saini HK, Mole DR, et al. Integrated analysis of microRNA and mRNA expression and association with HIF binding reveals the complexity of microRNA expression regulation under hypoxia. *Mol Cancer*. 2014;13:28.
- [86] Campbell JD, Liu G, Luo L, et al. Assessment of microRNA differential expression and detection in multiplexed small RNA sequencing data. *RNA*. 2015;21:164–171.
- [87] Taylor CF, Field D, Sansone SA, et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotechnol*. 2008;26:889–896.