bioRxiv preprint doi: https://doi.org/10.1101/417782; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Identification of potential microRNAs associated with Herpesvirus family based on bioinformatic analysis

Kevin Lamkiewicz[1,2], Emanuel Barth[2,3], Manja Marz[1,2,3,4,*] Bashar Ibrahim[1,2,*]

[1]European Virus Bioinformatics Center, Jena, Germany
[2]RNA Bioinformatics and High-Throughput Analysis, Friedrich Schiller University Jena, Jena, Germany
[3]FLI Leibniz Institute for Age Research, Jena, Germany
[4]German Center for Integrative Biodiversity Research, Halle-Jena-Leipzig, Germany
[*]Coresponding author.

ABSTRACT

MicroRNAs (miRNAs) are known key regulators of gene expression on posttranscriptional level in many organisms encoded in mammals, plants and also several viral families. To date, no homologous gene of a virus-originated miRNA is known in other organisms. To date, only a few homologous miRNA between two different viruses are known, however, no gene of a virus-originated miRNA is known in any organism of other kingdoms. This can be attributed to the fact that classical miRNA detection approaches such as homology-based predictions fail at viruses due to their highly diverse genomes and their high mutation rate.
Here, we applied the virus-derived precursor miRNA (pre-miRNA) prediction pipeline `ViMiFi`, which combines information about sequence conservation and machine learning-based approaches, on Human Herpesvirus 7 (HHV7) and Epstein-Barr virus (EBV). `ViMiFi` was able to predict 61 candidates in EBV, which has 25 known pre-miRNAs. From these 25, `ViMiFi` identified 20. It was further able to predict 18 candidates in the HHV7 genome, in which no miRNA had been described yet. We also studied the undescribed candidates of both viruses for potential functions and found similarities with human snRNAs and miRNAs from mammals and plants.

Key words: HHV7, EBV, miRNA, machine learning, Herpesvirus

1 Introduction

MicroRNAs (or miRNAs) are small RNA molecules (∼20–24 nt) involved in the regulation of gene expression by targeting messenger RNAs (mRNA) for cleavage or translational repression. Usually, two processing steps are required to generate miRNAs. First, the nuclear RNase III Drosha cuts the primary miRNA which leads to the precursor miRNA (pre-miRNA)[1–3]. The pre-miRNA is subsequently transported to the cytoplasm and further processed by Dicer, another RNase III protein, to the double-stranded miRNA/miRNA*-complex[4–6]. After separating the mature miRNA from the complementary miRNA* sequence, it is loaded into the RNA-induced silencing complex (or RISC). The loaded RISC complex targets messenger RNA (mRNA) very specifically, depending on the sequence of the miRNA. An interaction between RISC and mRNA inhibits the translation of the targeted mRNA[7].
Aside from eukaryotes, viruses can also encode miRNAs in their genome[8,9]. Over the last decade,

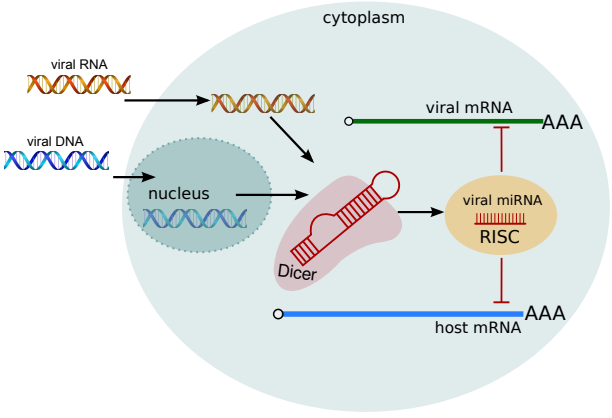

Figure 1: Once the viral genome entered the host cell, transcription, translation and processing mechanisms of the host are exploited. Viral precursor miRNAs are processed by Dicer and loaded into the RISC complex. Interaction with both the viral and the host mRNA is possible.

bioRxiv preprint doi: https://doi.org/10.1101/417782; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

several hundred viral miRNAs have been discovered, most of them in DNA viruses[10–14]. Viral genomes tend to be highly diverse (especially in RNA virus) and/or adapted to the host organism (e.g. in terms of codon usage)[15–17]. As depicted in Figure 1, viral miRNAs are shown to not only regulate viral mRNAs but also cellular host mRNAs[18–20].

Identifying miRNAs accurately is a challenging task that requires the integration of experimental approaches with computational methods. Several miRNA prediction algorithms have been developed, but unfortunately none is able to provide a list of miRNA candidates for viruses. These prediction algorithms can be categorized into two groups. The first group deals with homology information of closely related species and compares the query sequence with annotations of known miRNAs in these species. Examples for homology-based tools are `miRscan`[21], `miRSeeker`[22] and `miRAlign`[23]. The second group, including tools like `TripletSVM`[24], `NOVOMIR`[25] and `HHMMiR`[26], uses machine learning where known miRNAs are used to train a classification model and eventually to predict new candidates in query sequences.

However, as shown in Figure 2, both approaches fail for viruses (especially RNA viruses), due to the high diversity of their genomes and the lack of sufficient data.

Recently, we developed a pipeline for prediction of viral precursor miRNAs *de novo*, called `ViMiFi`, (*vi*ral *mi*RNA *fi*nder) [submitted elsewhere]. The pipeline combines homology-based and machine learning-based approaches and has been tested with six different classifier. So far, other machine learning approaches predict pre-miRNA in single sequences[24–26], whereas our approach processes multiple sequence alignments of viral genomes, enabling a larger feature space for classification.

In this study, we present the results of the first application of `ViMiFi`, applied on Human Herpesvirus 7 (HHV7). This virus has a double-stranded DNA (dsDNA) genome with 145 kb in size and belongs to the genus of *Roseoloviruses* and the (sub-)family of *Herpesviridae*. There are nine different herpesviruses known that infect human. To date, most of these nine viruses encode for viral miRNAs[8,9,11,27–30]. The Varicella-Zoster virus encodes several small ncRNAs[29], however, no miRNA was cloned yet. For the HHV7 no miRNAs have been reported either.

HHV7 infections are associated with a number of symptoms such as acute febrile respiratory disease, fever, diarrhea and low lymphocyte counts[31]. Furthermore, there are indications that HHV7 could contribute to the development of drug-induced hypersensitivity syndrome[32], encephalopathy[33], hemiconvulsion-hemiplegia-epilepsy syndrome[34] and hepatitis infection[35]. Thus, HHV7 has a high relevance and identifying viral miRNAs within the HHV7 genome may lead to new therapies against it. For other human herpesviruses it has been shown that viral miRNAs may play an important role in maintaining the latency phase during virus infection[36].

Here, we identified 18 new regions of the HHV7 genome containing potential pre-miRNAs and searched for homologous sequences in other species from all kingdoms. In order to measure the quality of our predictions, we scanned the Epstein-Barr virus (EBV). This virus is known to encode for at least 25 pre-miRNAs. From these, `ViMiFi` identified 20 pre-miRNAs and further proposes 41 novel candidates within the EBV genome. Figure 6 gives a schematic overview of the genome organization of EBV with the positions of our predicted candidates.

2 Methods

Virus Sequence Data We used eleven different HHV7 genome sequences for our analyses. All sequences were downloaded from the ViPR database[37] (see Table 1). The multiple sequence alignment (MSA) was created with `mafft`[38], containing 163.407 columns and an average number of 259 gaps per sequence.

Table 1: Overview of HHV7 genomes used in this study. Sequences were downloaded from the ViPR database[37].

Accession ID	Sequence length
NC_001716	153.080
AF037218	153.080
DD138892	153.080
DJ052727	153.080
DL242081	153.080
FV537034	153.080
FV537035	153.080
U43400	144.861
DD250331	144.861
DI115158	144.861
HV987120	144.861

Training data Our positive training data consists of all viral precursor miRNAs (320) that are publicly available in `miRBase` 22.0[39]. For each viral precursor structure of the positive set we randomly sampled three sequences from the whole genome of the same virus. The length of such a sequence was randomly set between 70

bioRxiv preprint doi: https://doi.org/10.1101/417782; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

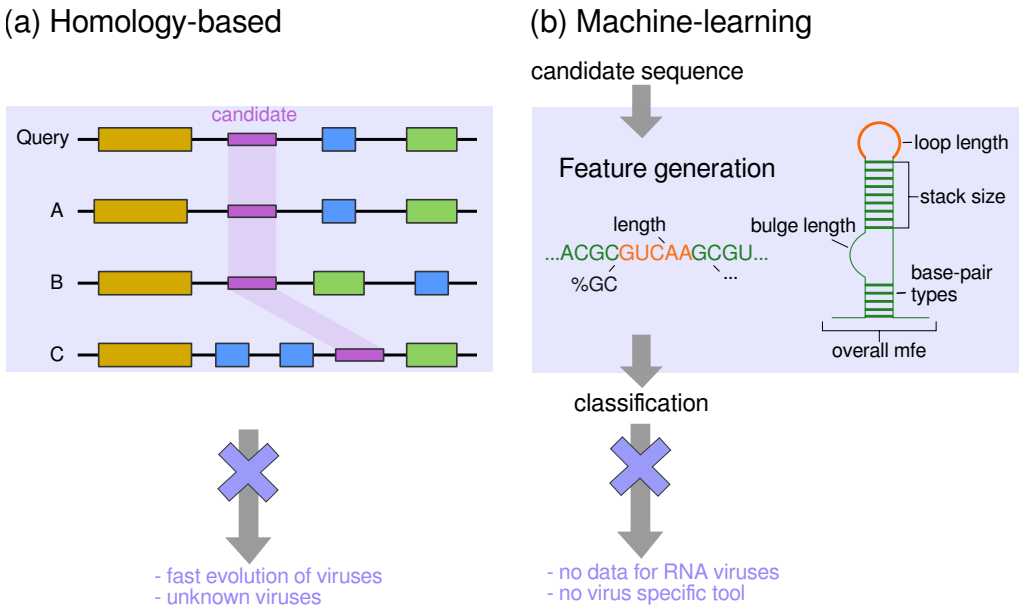

Figure 2: Common approaches for the prediction of miRNAs and their drawbacks on viral data. For both, the homology-based and the machine learning-based approach, the lack of data and the high diversity of viral genomes make these analyses difficult. In our comparison, homology-based approaches without the usage of synteny were studied.

and 120 nucleotides. The sampling procedure was repeated until three negative instances per precursor were found. Each sampled sequence had to fold into a characteristic stem-loop structure of minimum size of 50 nucleotides and at least 14 base pairings. Furthermore, the minimum free energy (MFE) of the sampled sequences had to be lower than -18 kcal/mol. These parameters were chosen as they represent a canonical pre-miRNA. Performing an `RNAclust`[40] analysis on clustered viral pre-miRNAs indicates that the consensus structure of 157 pre-miRNA sequences fulfill the described characteristics. The structure is shown in Figure 3. As a final filter, we checked whether the sampled sequence had a sequence similarity over 90 % with one of the known precursor sequences. Sequence similarity was calculated with the Levenshtein distance[41].

Feature discrepancy Properties derived from the sequence and the structure of a query RNA are called features. Here, we used triplet frequencies (introduced in[24]), relative GC-content of the sequence, the MFE of the structure, number and length of bulges, the number of pairings in the largest stacking and the loop length of the structure. In order to analyze which feature is important for the classification we performed a F-value calculation (see Equation 1), where $\overline{x}_i^{(+)}$, $\overline{x}_i^{(-)}$ and $\overline{x}_i$ are the averages of the ith feature of the positive, negative and the whole dataset, respectively. Furthermore, n_+ and n_- denote the number of instances in the positive and negative set.

Classification by ViMiFi Here, we used the `alignment mode` of `ViMiFi v.0.1`. MSAs have to be provided to `ViMiFi`. As `ViMiFi` is not published yet, we briefly describe the workflow and default parameters used in this study. The MSA is processed with a sliding window of size 120 and stepsize 20. Each window is folded with `RNAalifold 2.4.9`[42] from the ViennaRNA package.
We use the combination of `mafft` and `RNAalifold` since we are looking in complete viral genome alignments for pre-miRNAs. Algorithms like `LocARNA`[43] do consider sequence and structural information for the creation of an alignment simultanously, however, it is not feasible to create such an alignment for several genomes of over 150.000 nucleotides in size.
Each window of our MSA is then shuffled column-wise 1.000 times. The shuffled window is folded in the same way as the original MSA fragment. From the resulting structures a p-value is calculated based on the z-score of the MFE structures.
Among the features mentioned above, `ViMiFi` relies on so-called triplet features, proposed in `TripletSVM`[24]. Triplets are suitable to model the sequence and local structure of a query sequence. For each triplet of nucleotides, the second

bioRxiv preprint doi: https://doi.org/10.1101/417782; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

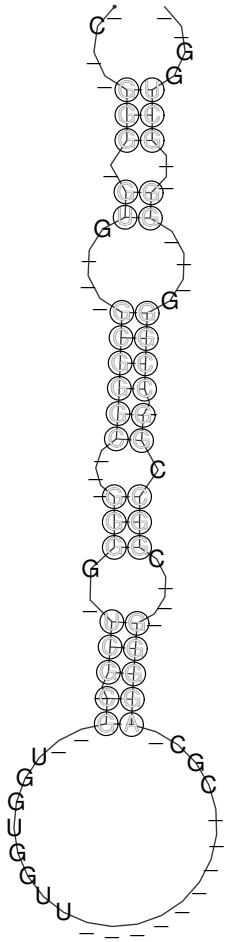

Figure 3: Consensus structure of 157 viral pre-miRNAs clustered with `RNAclust` and folded with `RNAalifold`. Due to the features of this structure we set the sampling parameters for negative instances as described.

nucleotide itself and the information whether a nucleotide of a triplet is paired or unpaired in the predicted secondary structure is stored. Thus for example, the sequence `GGGGUCCCC` with the structure `((((...))))` in dot-bracket format would lead to the following triplets: $\overset{.G.}{(((}$, $\overset{.G.}{((.}$, $\overset{.G.}{(..}$, $\overset{.U.}{...}$, $\overset{.C.}{..)}$, $\overset{.C.}{.))}$, $\overset{.C.}{)))}$.
Parameters of `ViMiFi` were set to their respective default values (windowsize: 120, stepsize: 20, number of shuffled sequences: 1.000), with the exception of the p-value cutoff, which was set to $p < 0.01$.

Secondary Structure analysis of single candidates Potential candidates for viral pre-miRNAs were analyzed regarding their secondary structure. For this, we applied `RNAfold 2.4.9`[42] with parameters `-noLP -p` on each candidate sequence and colored the resulting MFE structure based on the base pairing probabilities observed in the Boltzmann distribution of secondary structures.

Homolog Search & Function An exhaustive search on the `Rfam` database version 13.0[44,45] was performed using `Infernal 1.1.2`[46] to annotate some of the predicted candidates functionally. Using `cmscan` from the `Infernal` package with default parameters, all potential pre-miRNA of HHV7 and EBV were compared with the covariance models of the `Rfam` release. Furthermore, for each pre-miRNA candidate a `BLAST 2.7.1+`[47] search was performed against all `miRBase` precursor sequences. For this, we reduced the word size to 12 for the blast search, thus resulting in the command `blastn -word_size 12`.

Validation of predictions using small RNA-Seq data In order to see whether candidates predicted in the HHV7 by `ViMiFi` are already supported by publicly available RNA-Seq data, we used the RNA-Seq libraries of the recent publication by Lewandowska *et al.* (2017)[48]. The libraries are accessible via `http://doi.org/10.5281/zenodo.400950`. Viral reads were mapped with `TopHat2 v2.1.1`[49] with default parameters against the HHV7 reference genome (NCBI accession NC_001716) and data was displayed using the Integrative Genome Browser (IGV)[50].

3 Results

Known EBV pre-miRNAs are predicted by `ViMiFi` We examined several different Epstein-Barr virus genomes to identify pre-miRNAs. For this, all EBV-derived pre-miRNAs were excluded

bioRxiv preprint doi: https://doi.org/10.1101/417782; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$F(i) = \frac{(\overline{x}_i^{(+)} - \overline{x}_i)^2 + (\overline{x}_i^{(-)} - \overline{x}_i)^2}{\frac{1}{n_+ - 1}} \sum_{k=1}^{n_+} (x_{k,i}^{(+)} - \overline{x}_i^{(+)})^2 + \frac{1}{n_- - 1} \sum_{k=1}^{n_-} (x_{k,j}^{(-)} - \overline{x}_i^{(-)})^2 \,, \quad (1)$$

from our training set to avoid an overfitting effect with regard to EBV. Figure 6 gives an overview of the coding sequences (CDS) of both strands within the EBV genome, the 25 known pre-miRNAs annotated in `miRBase` and all 61 candidates predicted by `ViMiFi`.

We have successfully predicted 20 of 25 known pre-miRNAs curated in `miRBase`. The five missing pre-miRNAs contain three features being different from the other 315 of the positive set pre-miRNAs: the triplet $\begin{smallmatrix} \cdot C \cdot \\ \cdot\cdot(\end{smallmatrix}$, the triplet $\begin{smallmatrix} \cdot A \cdot \\ \cdot((\end{smallmatrix}$ and the loop length. Figure 4 visualizes these differences based on our F-value analysis. On average, the loop length of missed pre-miRNAs is about 6 nucleotides smaller, whereas the frequencies of the two triplets are higher compared to the identified pre-miRNAs. We hypothesize that `ViMiFi` cannot identify these five missing pre-miRNAs, because their features are not similar enough to the ones learned from our positive training set. Reevaluating the feature design and feature selection of `ViMiFi`, as well as obtaining more validated viral pre-miRNAs for the training set, may lead to even better predictions, and thus increases the precision.

We compared the start and stop positions of our predicted EBV candidates with the genomic coordinates of annotated pre-miRNAs of `miRBase`. This comparison revealed that 20 known pre-miRNAs were identified by `ViMiFi`.

ViMiFi predicted 41 novel pre-miRNA candidates We predicted with `ViMiFi` 41 novel pre-miRNA candidates for the EBV genome and investigated them further with `Infernal` and `blastn`. As shown in Figure 6, the majority of our predictions cluster in three regions. These regions do not have a CDS on either strand in the NCBI annotation and two of these regions are already known to be miRNA clusters – namely the BHRF1 and the BART cluster. However, one cluster of our predictions (around position 7.800 – 8.100) neither aligns with known pre-miRNAs nor with CDS regions. The candidates miRNA_ebv_03 to miRNA_ebv_07 belong to this cluster. None of these candidates have a homolog pre-miRNA in the `miRBase`. Thus, we extracted the region from position 7.800 – 8.100 of the EBV genome and made a structure analysis using `RNAfold`[42]. The MFE structure is shown in Figure 5 and resembles the shape of a potential primary miRNA (pri-miRNA)[51–53]. For all results of the `blastn` search we refer to Supplement Table S1.

Applying a covariance model search from the `Infernal` toolkit against all known RNA families of the `Rfam` database revealed 37 hits with a significant (< 0.01) e-value. The results of this search are shown in Table 2. Interestingly, the candidates miRNA_ebv_03 to miRNA_ebv_07 are not associated with EBV. However, all of them have similarities with the RNA family miR-563. This family is conserved among mammals, however, the precise function has not been identified, yet. Since the e-values of these hits are close to 1, the predicted EBV pre-miRNA cluster does probably not belong to this RNA family, but might resemble its own RNA family.

Furthermore, the candidates miRNA_ebv_09 to miRNA_ebv_15 have hits on the RNA family ebv-sisRNA-2, which is not a miRNA, but a long non-coding RNA (lncRNA) of EBV. `ViMiFi` detected these six candidates in the region of ebv-sisRNA-2, because this lncRNA has many stable local stem-loops in its secondary structure which are similar to pre-miRNAs. The ebv-sisRNA-2 lncRNA may play an important role in the maintenance of virus latency[54,55].

18 novel pre-miRNAs in HHV7 `ViMiFi` identified 18 different regions in the HHV7 alignment, consisting of the genomes as displayed in Table 1, to be potential pre-miRNAs. Table 3 shows all candidates sorted by their genomic coordinates relative to the reference genome and Figure 7 shows the positions of the candidates in relation to the CDS annotation of NCBI.

Each candidate was analyzed with `blastn` and `Infernal`. Performing the `blastn` search against the `miRBase` database led to some hits in several plants, *Homo sapiens* and *Macaca mulatta* (see Supplement Table S3), however the e-values indicate that those hits are probably not real homologs, but at least share some similarities in terms of sequence.

Seven HHV7 candidates are not overlapping with coding regions From our 18 predicted candidates within the HHV7 genome, 7 do not overlap with annotated CDS on any strain. These 7 candidates are miRNA_HHV7_01, miRNA_HHV7_02, miRNA_HHV7_12, miRNA_HHV7_14, miRNA_HHV7_15, miRNA_HHV7_16 and miRNA_HHV7_18. In Figure 7 the corresponding green tiles are drawn

bioRxiv preprint doi: https://doi.org/10.1101/417782; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

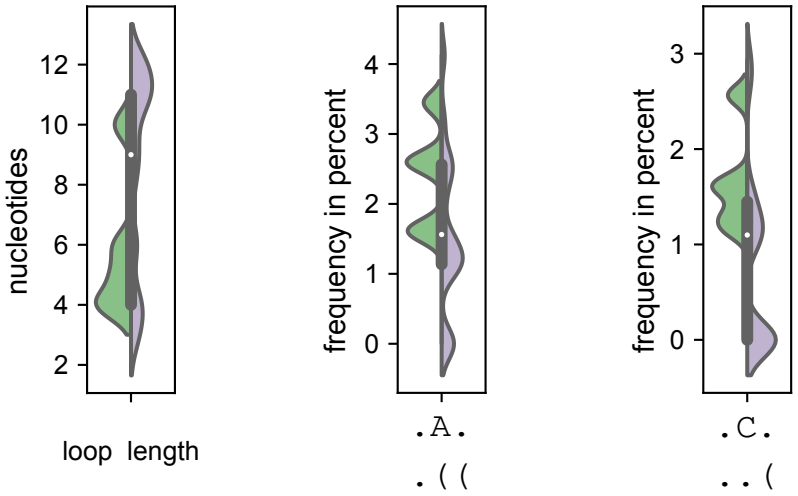

Figure 4: Violinplots of the three features with the highest F-value between missing (green) and identified (purple) EBV pre-miRNAs. Missed pre-miRNAs tend to have a smaller loop region as well as higher frequencies of the two triplets $\begin{smallmatrix}\cdot A\cdot\\ \cdot((\end{smallmatrix}$ and $\begin{smallmatrix}\cdot C\cdot\\ \cdot\cdot(\end{smallmatrix}$.

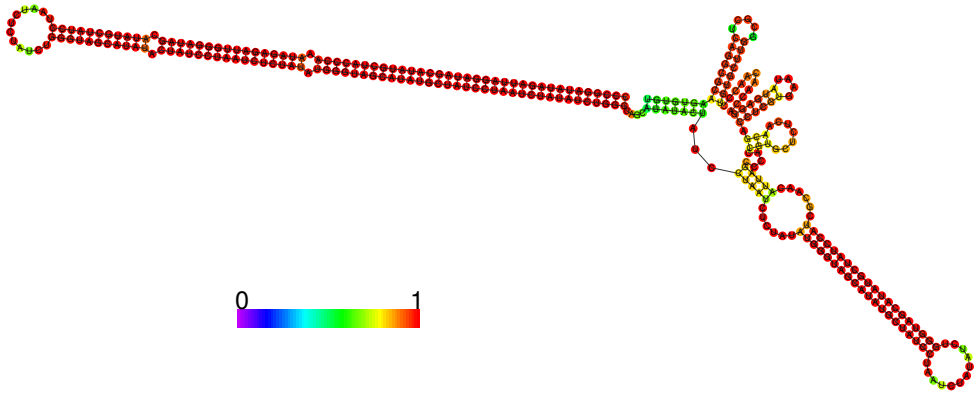

Figure 5: MFE structure of the region 7.800 – 8.100 of the EBV genome, folded with RNAfold and colored based on the base pairing probabilities observed in the set of sampled structures in the Boltzmann ensemble. The structure resembles a plausible pri-miRNA.

bioRxiv preprint doi: https://doi.org/10.1101/417782; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2: Results of the covariance model search with `Infernal` applied to predicted EBV candidates. We searched against the Rfam database. For the complete list of all `Infernal` results we refer to Supplement Table S2.

Candidate	Family Name	Accession ID	E-Value	Bitscore
miRNA_ebv_03	miR-563	RF01003	0.35	16.1
miRNA_ebv_04	miR-563	RF01003	0.0091	22.0
miRNA_ebv_05	miR-563	RF01003	0.065	18.8
miRNA_ebv_06	miR-563	RF01003	0.016	21.3
miRNA_ebv_07	miR-563	RF01003	0.24	16.6
miRNA_ebv_09	ebv-sisRNA-2	RF02598	2.3e-20	105.5
miRNA_ebv_10	ebv-sisRNA-2	RF02598	8e-16	85.1
miRNA_ebv_11	ebv-sisRNA-2	RF02598	2.3e-20	105.5
miRNA_ebv_12	ebv-sisRNA-2	RF02598	2.3e-20	105.5
miRNA_ebv_13	ebv-sisRNA-2	RF02598	8e-16	85.1
miRNA_ebv_14	ebv-sisRNA-2	RF02598	2.4e-20	105.5
miRNA_ebv_15	ebv-sisRNA-2	RF02598	2.5e-20	105.5
miRNA_ebv_20	miR-BHRF1-3	RF00367	1.7e-26	100.1
miRNA_ebv_35	miR-BART3	RF00866	4.8e-22	95.7
miRNA_ebv_36	miR-BART3	RF00866	3e-22	96.2
miRNA_ebv_38	miR-BART17	RF00863	1.2e-26	107.8
miRNA_ebv_39	miR-BART5	RF00867	0.0021	25.2
miRNA_ebv_47	miR-BART7	RF00869	2.1e-24	105.1
miRNA_ebv_50	miR-BART12	RF00874	0.0021	24.7
miRNA_ebv_52	miR-BART12	RF00874	6e-19	84.0
miRNA_ebv_54	miR-BART20	RF00864	6.5e-25	102.1

Table 3: Overview of predicted pre-miRNAs in HHV7 using `ViMiFi`. Each candidate is assigned an unique ID. The start and stop positions are relative to the reference genome accessible via the NCBI accession NC_001716.

Candidate	Start Position	Stop Position
miRNA-HHV7_01	9.663	9.761
miRNA-HHV7_02	28.215	28.316
miRNA-HHV7_03	34.374	34.467
miRNA-HHV7_04	51.590	51.668
miRNA-HHV7_05	61.730	61.807
miRNA-HHV7_06	67.481	67.585
miRNA-HHV7_07	75.954	76.024
miRNA-HHV7_08	87.892	87.995
miRNA-HHV7_09	94.797	94.881
miRNA-HHV7_10	104.503	104.588
miRNA-HHV7_11	110.869	110.943
miRNA-HHV7_12	116.193	116.306
miRNA-HHV7_13	122.944	123.026
miRNA-HHV7_14	127.429	127.537
miRNA-HHV7_15	127.487	127.595
miRNA-HHV7_16	127.723	127.813
miRNA-HHV7_17	148.530	148.625
miRNA-HHV7_18	152.705	152.811

bioRxiv preprint doi: https://doi.org/10.1101/417782; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

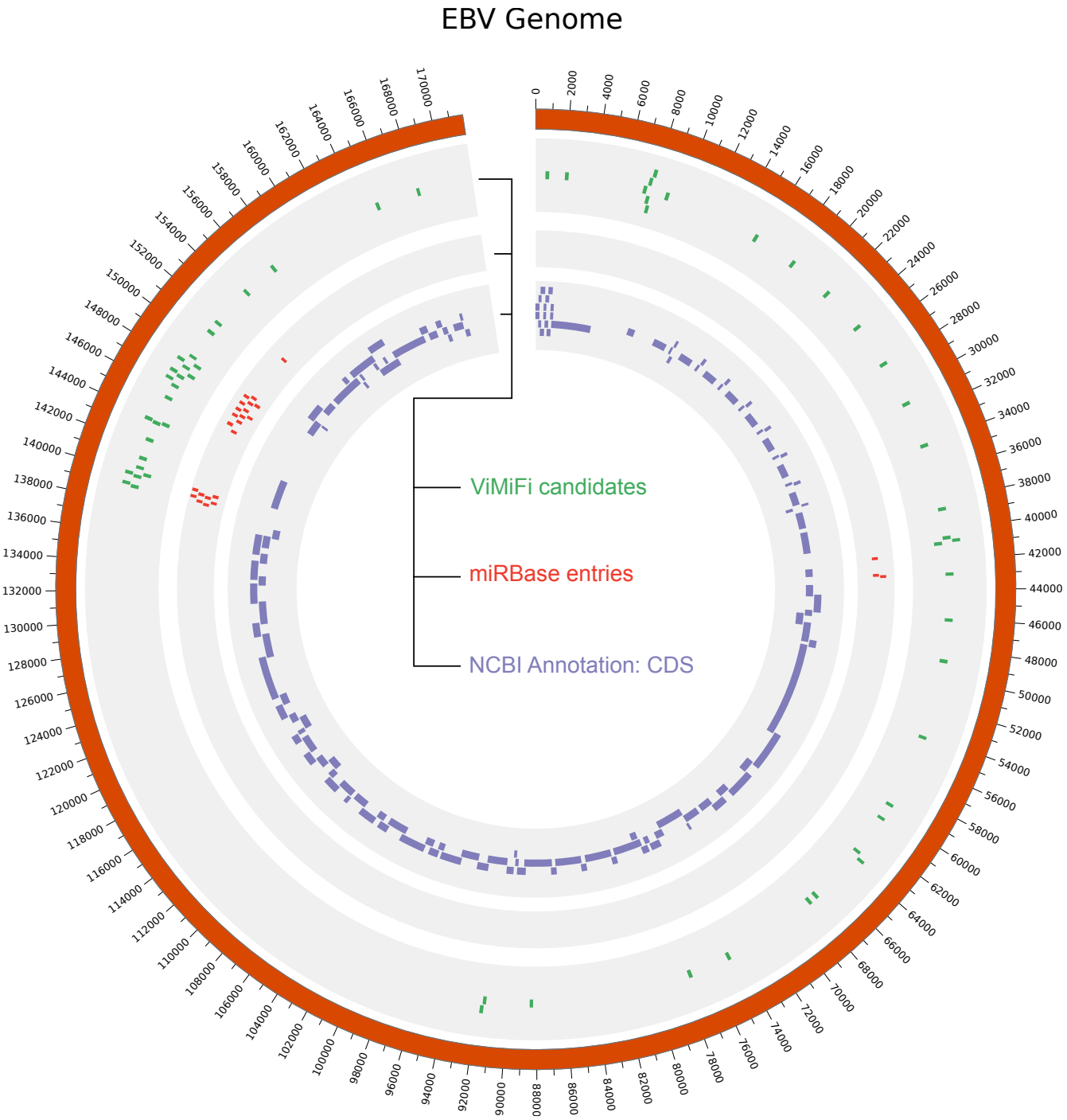

Figure 6: Circular representation of the Epstein-Barr virus genome and annotation of known and predicted pre-miRNAs. The inner track in purple represent annotated coding regions from NCBI. The middle track shows all pre-miRNA entries of the `miRBase` in red. All candidates predicted by `ViMiFi` are represented in green in the outer track.

bioRxiv preprint doi: https://doi.org/10.1101/417782; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

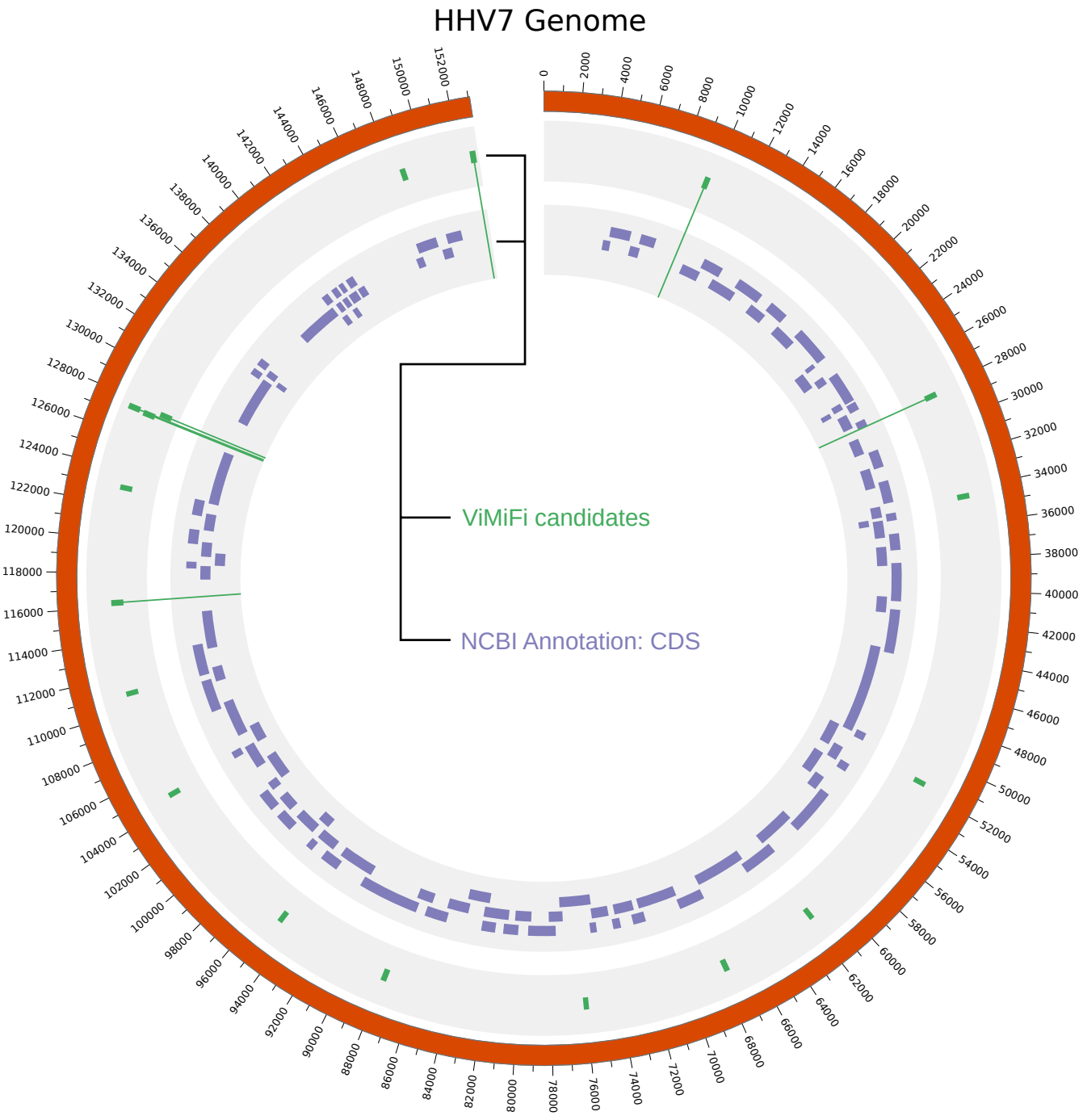

Figure 7: Circular representation of the Human Herpesvirus 7 genome and annotation of predicted pre-miRNAs. The inner track in purple represent annotated coding regions from NCBI. All candidates predicted by `ViMiFi` are represented in green in the outer track. Lines from candidates going into the inner track indicate potential pre-miRNAs that do not overlap with annotated CDS regions.

bioRxiv preprint doi: https://doi.org/10.1101/417782; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

with a line into the CDS track. Even more intriguing, for the candidates miRNA_HHV7_01 and miRNA_HHV7_18, no `blastn` hits against the `miRBase` were obtained at all. Thus, no other known pre-miRNA derived from all organisms in the `miRBase` are similar to these new candidates. Analyzing the secondary structures of the 7 candidates shows that all of them are able to fold into a structure with at least one stable stem-loop (see Figure 8). Even structures that do not form the canonical single hairpin structure of pre-miRNAs, may exploit another pathway of pre-miRNA processing, for example the tRNase Z-dependent pathway[56]. As first step of validation, we examined the recently published RNA-Seq data of Lewandowska *et al.*[48]. We observed weak expression for the candidates miRNA_HHV7_02, miRNA_HHV7_12, miRNA_HHV7_14, miRNA_HHV7_15 and miRNA_HHV7_17 (see Supplement Figures S1 – S4). These observations indicate transcription activity at the predicted pre-miRNA regions, however, they are no clear evidence for the presence of pre-miRNAs. We therefore argue that these novel candidates are worth investigating by experiments in order to validate their presence and potential targets during infection of human cells.
Interestingly, one of the candidates, that overlaps with CDS, miRNA_hhv7_06 has a significant `blastn` hit on the human miRNA hsa-miR-4432. However, this human miRNA is not confidently reported yet and has around 100 of predicted potential targets within the human transcriptome[57].

Two HHV7 candidates can be linked to known snRNAs Further, we searched for similarities of the predicted HHV7 candidates with other RNA families in the `Rfam` database. A covariance model search for each candidate was performed. The most significant results are shown in Table 4, whereas all results are shown in Supplement Table S4. Only one hit achieved an e-value < 0.001, namely the non-overlapping miRNA candidate miRNA_hhv7_15 on the RNA family C**y**anobacterial **f**unctional **R**NA 19 (Yfr19). Yfrs are known to regulate gene expression in cyanobacteria. In particular, Yfr19 is controlled by phage infection[58], linking this bacterial ncRNA to virus infections.
Furthermore, miRNA_HHV7_04 has similarities to the RNA family SNORD111. SNORD RNAs are small non-coding RNAs which are involved in the modification of small nuclear RNAs (snRNAs). The SNORD111 (also known as HBII-82) is being predicted to guide the 2'O-ribose methylation of 28S rRNA in mouse, human, other mammals and aves[59,60]. The `Infernal` search also found similarities between the RNA family snoU18, which is mainly observed in insects, and the candidate miRNA_HHV7_10. The snoU18 is known to be a C/D-box snRNA and is associated with methylation as well. The predicted candidates may compete with these cellular snRNAs to prevent methylation. Moreover, there exists a non-canonical pathway for pre-miRNA processing that is derived from C/D-box snRNAs[56].

4 Conclusion

We applied the virus specific pre-miRNA prediction pipeline `ViMiFi` on the human herpesvirus 7, the solely human-infecting herpesvirus that has no miRNAs described, and on Epstein-Barr virus, a human-infecting herpesvirus with 25 described pre-miRNAs. We were able to identify 20 out of these 25 precursor structures, and further, predict 41 more regions to be potential pre-miRNAs candidates. Out of these 41 candidates, 5 do not overlap with any known annotation, neither the NCBI coding region nor the `miRBase` entries.

Moreover, we proposed 18 pre-miRNAs candidates in HHV7. Out of these, 16 candidates show similarities to known pre-miRNAs in plants, *Homo sapiens* and *Macaca mulatta*, whereas the candidates miRNA_HHV7_01 and miRNA_HHV7_18 did not yield any result in the applied homology search. Further, seven candidates do not overlap with known annotated protein coding genes. The predicted secondary structures and the genomic context of these seven candidates indicate that they are undescribed pre-miRNAs encoded by Human Herpesvirus 7. Even though, some structures do not resemble the canonical pre-miRNA, they could be processed via non-canonical pathways like the tRNase-Z or snRNA-derived pathway.

Comparisons with RNA families stored at the Rfam database show minor similarities with other miRNA and snoRNA families. These findings suggest that the candidates predicted by `ViMiFi` may have similar functions. Intriguingly, two HHV7 candidates have similarities to C/D-box snRNAs, which are known to guide the methylation of rRNA. These methylation may lead to "specialized ribosomes"[61]. We hypothesize that viral ncRNAs with similarities to such cellular snRNAs may compete with these snRNAs and prevent the creation of specialized ribosomes, and thus, ensure viral protein translation.

The next obvious step is the validation of our

bioRxiv preprint doi: https://doi.org/10.1101/417782; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

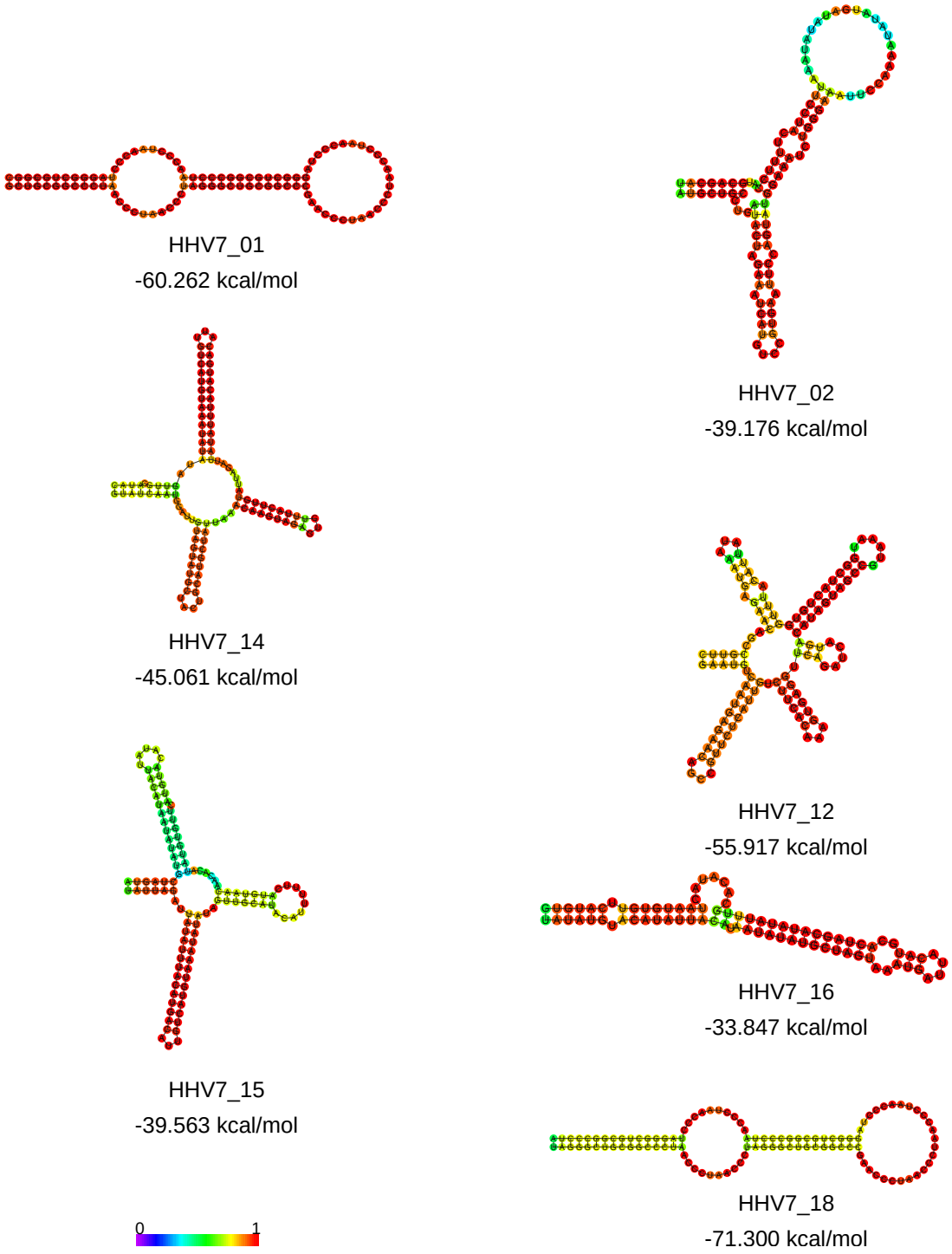

Figure 8: Predicted secondary structures of the potential novel pre-miRNAs encoded by HHV7. Displayed are the minimum free energy structures of all candidates that do not overlap with annotated CDS. Each structure is colored based on the base pairing probabilities derived from the partition function. The more a nucleotide is colored red, the higher the likelihood of the predicted structure. Since the candidates 02, 12, 14 and 15 do not fold like known pre-miRNAs they arguably could be false positives. However, Supplement Figures S1 – S4 show that at least some reads can be aligned to the positions of those candidates, indicating active transcription of these regions. It might be possible that those four predicted candidates are processed in alternative miRNA pathways in contrast to the canonical Drosha and Dicer pathway.

bioRxiv preprint doi: https://doi.org/10.1101/417782; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 4: Results of the covariance model search with `Infernal` applied to predicted HHV7 candidates. We searched against the Rfam database. For each candidate the most significant hit is reported. For all hits we refer to Supplement Table S4.

Candidate	Family Name	Accession ID	E-Value	Bitscore
miRNA_HHV7_01	miR-497	RF00793	0.29	16.7
miRNA_HHV7_03	miR-1829	RF00955	1.1	14.1
miRNA_HHV7_04	SNORD111	RF00611	0.046	20.9
miRNA_HHV7_05	SNORD43	RF00221	1.5	12.6
miRNA_HHV7_06	ES003	RF02750	4.3	10.4
miRNA_HHV7_07	sn2991	RF01202	4.9	12.1
miRNA_HHV7_08	miR-278	RF00729	2.7	10.6
miRNA_HHV7_09	Alpha_RBS	RF00140	6.4	7.0
miRNA_HHV7_10	snoU18	RF01159	0.089	11
miRNA_HHV7_12	SNORD91	RF00580	2.3	14.6
miRNA_HHV7_13	CRISPR-DR5	RF01318	2	15.6
miRNA_HHV7_14	Yfr19	RF02366	0.12	21.6
miRNA_HHV7_15	Yfr19	RF02366	0.0026	29.2
miRNA_HHV7_16	miR-374	RF00840	0.45	15.8
miRNA_HHV7_17	miR-563	RF01003	4.3	11.6
miRNA_HHV7_18	Telomerase-vert	RF00024	0.24	10.9

candidates *in vitro* and/or *in vivo*. Since the data of viral pre-miRNAs is limited, and thus the performance of `ViMiFi` is limited as well, we hope that virologists are intrigued by these findings and are open for collaborations – each validated pre-miRNA can refine the model used by `ViMiFi` and thus lead to more accurate results in the future.

Author Contributions KL, BI and MM conceived of the presented idea. KL performed the computations. KL, EB and BI analyzed the data and discussed the results. EB performed the computational RNA-Seq validations. KL and BI wrote the manuscript with critical input from EB and MM.

Acknowledgments The work of KL is funded by the German Federal Ministry for Higher Education and Research (BMBF): STIKO Serologie (InfectControl 2020), Project Number 03ZZ0820A. We thank Anna Strototte and Katja Meyer for proofreading the manuscript.

Competing Interests The authors declare no competing interests.

References

[1] J. C. Carrington, V. Ambros, Role of microRNAs in plant and animal development, Science 301 (5631) (2003) 336–338.

[2] V. Ambros, R. C. Lee, A. Lavanway, P. T. Williams, D. Jewell, MicroRNAs and other tiny endogenous RNAs in *C. elegans*, Curr Biol 13 (10) (2003) 807–18.

[3] B. Bartel, D. P. Bartel, MicroRNAs: at the root of plant development?, Plant Physiol 132 (2) (2003) 709–717.

[4] G. Hutvágner, J. McLachlan, A. E. Pasquinelli, É. Bálint, T. Tuschl, P. D. Zamore, A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA, Science 293 (5531) (2001) 834–838.

[5] A. Grishok, A. E. Pasquinelli, D. Conte, N. Li, S. Parrish, I. Ha, D. L. Baillie, A. Fire, G. Ruvkun, C. C. Mello, Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing, Cell 106 (1) (2001) 23–34.

[6] Y. Lee, K. Jeon, J.-T. Lee, S. Kim, V. N. Kim, MicroRNA maturation: stepwise processing and subcellular localization, EMBO J 21 (17) (2002) 4663–4670.

[7] J. Krol, I. Loedige, W. Filipowicz, The widespread regulation of microRNA biogenesis, function and decay, Nat Rev Genet 11 (9) (2010) 597–610.

[8] S. Pfeffer, M. Zavolan, F. A. Grässer, M. Chien, J. J. Russo, J. Ju, B. John, A. J. Enright, D. Marks, C. Sander,

bioRxiv preprint doi: https://doi.org/10.1101/417782; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

T. Tuschl, Identification of virus-encoded microRNAs., Science 304 (2004) 734–736. doi:10.1126/science.1096781.

[9] S. Pfeffer, A. Sewer, M. Lagos Quintana, R. Sheridan, C. Sander, F. A. Grässer, L. F. van Dyk, C. K. Ho, S. Shuman, M. Chien, et al., Identification of microRNAs of the herpesvirus family, Nat Methods 2 (4) (2005) 269–276.

[10] W. Dunn, P. Trang, Q. Zhong, E. Yang, C. Van Belle, F. Liu, Human cytomegalovirus expresses novel microRNAs during productive viral infection, Cell Microbiol 7 (11) (2005) 1684–1695.

[11] F. Grey, A. Antoniewicz, E. Allen, J. Saugstad, A. McShea, J. C. Carrington, J. Nelson, Identification and characterization of human cytomegalovirus-encoded microRNAs, J Virol 79 (18) (2005) 12095–12099.

[12] B. R. Cullen, Viruses and microRNAs, Nat Genet 38 (2006) S25–S30.

[13] R. L. Skalsky, B. R. Cullen, Viruses, microRNAs, and host interactions, Annu Rev Microbiol 64 (2010) 123–141.

[14] K. Plaisance-Bonstaff, R. Renne, Antiviral RNAi: Concepts, Methods, and Applications, Vol. 721 of Methods in Molecular Biology, Springer, 2011, Ch. Viral miRNAs, pp. 43–66. doi:10.1007/978-1-61779-037-9.

[15] P. Agudelo-Romero, P. Carbonell, M. A. Perez-Amador, S. F. Elena, Virus adaptation by manipulation of host's gene expression, PLOS One 3 (6) (2008) e2397.

[16] J. K. Taubenberger, J. C. Kash, Influenza virus evolution, host adaptation, and pandemic formation, Cell Host Microbe 7 (6) (2010) 440–451.

[17] N. D. Grubaugh, J. Weger-Lucarelli, R. A. Murrieta, J. R. Fauver, S. M. Garcia-Luna, A. N. Prasad, W. C. Black, G. D. Ebel, Genetic drift during systemic arbovirus infection of mosquito vectors leads to decreased relative fitness during host switching, Cell Host Microbe 19 (4) (2016) 481–492.

[18] R. W.-M. Lung, J. H.-M. Tong, Y.-M. Sung, P.-S. Leung, D. C.-H. Ng, S.-L. Chau, A. W.-H. Chan, E. K.-O. Ng, K.-W. Lo, K.-F. To, Modulation of LMP2A expression by a newly identified Epstein-Barr virus-encoded microRNA miR-BART22., Neoplasia (New York, N.Y.) 11 (2009) 1174–1184.

[19] E. Gottwein, D. L. Corcoran, N. Mukherjee, R. L. Skalsky, M. Hafner, J. D. Nusbaum, P. Shamulailatpam, C. L. Love, S. S. Dave, T. Tuschl, U. Ohler, B. R. Cullen, Viral microRNA targetome of KSHV-infected primary effusion lymphoma cell lines., Cell Host Microbe 10 (2011) 515–526. doi:10.1016/j.chom.2011.09.012.

[20] R. L. Skalsky, D. L. Corcoran, E. Gottwein, C. L. Frank, D. Kang, M. Hafner, J. D. Nusbaum, R. Feederle, H.-J. Delecluse, M. A. Luftig, T. Tuschl, U. Ohler, B. R. Cullen, The viral and cellular microRNA targetome in lymphoblastoid cell lines., PLoS Pathog 8 (2012) e1002484. doi:10.1371/journal.ppat.1002484.

[21] L. P. Lim, N. C. Lau, E. G. Weinstein, A. Abdelhakim, S. Yekta, M. W. Rhoades, C. B. Burge, D. P. Bartel, The microRNAs of *Caenorhabditis elegans*., Genes Dev 17 (2003) 991–1008. doi:10.1101/gad.1074403.

[22] E. C. Lai, P. Tomancak, R. W. Williams, G. M. Rubin, Computational identification of *Drosophila* microRNA genes., Genome Biol 4 (2003) R42. doi:10.1186/gb-2003-4-7-r42.

[23] X. Wang, J. Zhang, F. Li, J. Gu, T. He, X. Zhang, Y. Li, MicroRNA identification based on sequence and structure alignment., Bioinformatics 21 (2005) 3610–3614. doi:10.1093/bioinformatics/bti562.

[24] C. Xue, F. Li, T. He, G.-P. Liu, Y. Li, X. Zhang, Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine., BMC Bioinf 6 (2005) 310. doi:10.1186/1471-2105-6-310.

[25] J.-H. Teune, G. Steger, NOVOMIR: de novo prediction of microRNA-coding regions in a single plant-genome, J Nucleic Acids 2010.

[26] S. Kadri, V. Hinman, P. V. Benos, HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models, BMC Bioinf 10 (S1) (2009) S35.

[27] I. Jurak, M. F. Kramer, J. C. Mellor, A. L. van Lint, F. P. Roth, D. M. Knipe, D. M. Coen, Numerous conserved and divergent microRNAs expressed by herpes simplex viruses 1 and 2., J Virol 84 (2010) 4659–4672. doi:10.1128/JVI.02725-09.

[28] M. Nukui, Y. Mori, E. A. Murphy, A human herpesvirus 6A-encoded microRNA: role in viral lytic replication., J Virol 89 (2015) 2615–2627. doi:10.1128/JVI.02007-14.

bioRxiv preprint doi: https://doi.org/10.1101/417782; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[29] A. Markus, L. Golani, N. K. Ojha, T. Borodiansky-Shteinberg, P. R. Kinchington, R. S. Goldstein, Varicella-Zoster Virus Expresses Multiple Small Noncoding RNAs., J Virol 91. doi:10.1128/JVI.01710-17.

[30] X. Liu, C. Happel, J. M. Ziegelbauer, Kaposi's Sarcoma-Associated Herpesvirus MicroRNAs Target GADD45B To Protect Infected Cells from Cell Cycle Arrest and Apoptosis., J Virol 91. doi:10.1128/JVI.02045-16.

[31] S. Suga, T. Yoshikawa, T. Nagai, Y. Asano, Clinical features and virological findings in children with primary human herpesvirus 7 infection., Pediatrics 99 (1997) E4.

[32] H. Hara, M. Kobayashi, A. Yokoyama, M. Tochigi, A. Matsunaga, H. Shimizu, J. Goshima, H. Suzuki, Drug-induced hypersensitivity syndrome due to carbamazepine associated with reactivation of human herpesvirus 7., Dermatology (Basel) 211 (2005) 159–161. doi:10.1159/000086449.

[33] J. S. van den Berg, J. H. van Zeijl, J. J. Rotteveel, W. J. Melchers, F. J. Gabreëls, J. M. Galama, Neuroinvasion by human herpesvirus type 7 in a case of exanthem subitum with severe neurologic manifestations., Neurology 52 (1999) 1077–1079.

[34] J.-I. Kawada, H. Kimura, T. Yoshikawa, M. Ihira, A. Okumura, T. Morishima, F. Hayakawa, Hemiconvulsion-hemiplegia syndrome and primary human herpesvirus 7 infection., Brain Dev 26 (2004) 412–414. doi:10.1016/j.braindev.2003.12.003.

[35] T. Hashida, E. Komura, M. Yoshida, T. Otsuka, S. Hibi, S. Imashuku, S. Imashuku, T. Ishizaki, A. Yamada, S. Suga, Hepatitis in association with human herpesvirus-7 infection., Pediatrics 96 (1995) 783–785.

[36] I. W. Boss, K. B. Plaisance, R. Renne, Role of virus-encoded microRNAs in herpesvirus biology., Trends Microbiol 17 (2009) 544–553. doi:10.1016/j.tim.2009.09.002.

[37] B. E. Pickett, E. L. Sadat, Y. Zhang, J. M. Noronha, R. B. Squires, V. Hunt, M. Liu, S. Kumar, S. Zaremba, Z. Gu, L. Zhou, C. N. Larson, J. Dietrich, E. B. Klem, R. H. Scheuermann, ViPR: an open bioinformatics database and analysis resource for virology research., Nucleic Acids Res 40 (Database issue) (2012) D593–D598. doi:10.1093/nar/gkr859.

[38] K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, Mol Biol Evol 30 (4) (2013) 772–780.

[39] A. Kozomara, S. Griffiths-Jones, miRBase: integrating microRNA annotation and deep-sequencing data., Nucleic Acids Res 39 (2011) D152–D157. doi:10.1093/nar/gkq1027.

[40] Rnaclust: A tool for clustering of rnas based on their secondary structures using locarna, http://www.bioinf.uni-leipzig.de/ kristin/Software/RNAclust/, accessed: 2018-10-31.

[41] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, in: Dokl Phys, Vol. 10, 1966, pp. 707–710.

[42] R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, I. L. Hofacker, ViennaRNA package 2.0., Algorithms Mol Biol 6 (2011) 26. doi:10.1186/1748-7188-6-26.

[43] S. Will, K. Reiche, I. L. Hofacker, P. F. Stadler, R. Backofen, Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering., PLOS Comput Biol 3 (4) (2007) e65.

[44] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, S. R. Eddy, Rfam: an RNA family database., Nucleic Acids Res 31 (2003) 439–441.

[45] I. Kalvari, J. Argasinska, N. Quinones-Olvera, E. P. Nawrocki, E. Rivas, S. R. Eddy, A. Bateman, R. D. Finn, A. I. Petrov, Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families., Nucleic Acids Res 46 (2018) D335–D342. doi:10.1093/nar/gkx1038.

[46] E. P. Nawrocki, S. R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches., Bioinformatics 29 (2013) 2933–2935. doi:10.1093/bioinformatics/btt509.

[47] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, BLAST+: architecture and applications., BMC Bioinf 10 (2009) 421. doi:10.1186/1471-2105-10-421.

[48] D. W. Lewandowska, P. W. Schreiber, M. M. Schuurmans, B. Ruehe, O. Zagordi, C. Bayard, M. Greiner, F. D. Geissberger, R. Capaul, A. Zbinden, J. Böni,

bioRxiv preprint doi: https://doi.org/10.1101/417782; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

C. Benden, N. J. Mueller, A. Trkola, M. Huber, Metagenomic sequencing complements routine diagnostics in identifying viral pathogens in lung transplant recipients with unknown etiology of respiratory infection., PLoS One 12 (2017) e0177340. doi:10.1371/journal.pone.0177340.

[49] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S. L. Salzberg, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions., Genome Biol 14 (4) (2013) R36. doi:10.1186/gb-2013-14-4-r36.
URL http://dx.doi.org/10.1186/gb-2013-14-4-r36

[50] H. Thorvaldsdóttir, J. T. Robinson, J. P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration., Briefings Bioinf 14 (2013) 178–192. doi:10.1093/bib/bbs017.

[51] S. Diederichs, D. A. Haber, Sequence variations of microRNAs in human cancer: alterations in predicted secondary structure do not affect processing., Cancer Res 66 (2006) 6097–6104. doi:10.1158/0008-5472.CAN-06-0537.

[52] D. Liu, J. Fan, M. Mei, S. Ingvarsson, H. Chen, Identification of miRNAs in a liver of a human fetus by a modified method., PLoS One 4 (2009) e7594. doi:10.1371/journal.pone.0007594.

[53] Y.-K. Kim, B. Kim, V. N. Kim, Re-evaluation of the roles of DROSHA, Exportin 5, and DICER in microRNA biogenesis., Proc Natl Acad Sci U S A 113 (2016) E1881–E1889. doi:10.1073/pnas.1602532113.

[54] M. J. Farrell, A. T. Dobson, L. T. Feldman, Herpes simplex virus latency-associated transcript is a stable intron., Proc Natl Acad Sci USA 88 (1991) 790–794.

[55] W. N. Moss, N. Lee, G. Pimienta, J. A. Steitz, RNA families in Epstein-Barr virus., RNA Biol 11 (2014) 10–17. doi:10.4161/rna.27488.

[56] A. M. Abdelfattah, C. Park, M. Y. Choi, Update on non-canonical microRNAs., Biomol Concepts 5 (2014) 275–287. doi:10.1515/bmc-2014-0012.

[57] D. D. Jima, J. Zhang, C. Jacobs, K. L. Richards, C. H. Dunphy, W. W. L. Choi, W. Y. Au, G. Srivastava, M. B. Czader, D. A. Rizzieri, A. S. Lagoo, P. L. Lugar, K. P. Mann, C. R. Flowers, L. Bernal-Mizrachi, K. N. Naresh, A. M. Evens, L. I. Gordon, M. Luftig, D. R. Friedman, J. B. Weinberg, M. A. Thompson, J. I. Gill, Q. Liu, T. How, V. Grubor, Y. Gao, A. Patel, H. Wu, J. Zhu, G. C. Blobe, P. E. Lipsky, A. Chadburn, S. S. Dave, H. M. R. Consortium, Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs., Blood 116 (2010) e118–e127. doi:10.1182/blood-2010-05-285403.

[58] A. W. Thompson, K. Huang, M. A. Saito, S. W. Chisholm, Transcriptome response of high- and low-light-adapted Prochlorococcus strains to changing iron availability., ISME J 5 (2011) 1580–1594. doi:10.1038/ismej.2011.49.

[59] A. Hüttenhofer, M. Kiefmann, S. Meier-Ewert, J. O'Brien, H. Lehrach, J. P. Bachellerie, J. Brosius, RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse., EMBO J 20 (2001) 2943–2953. doi:10.1093/emboj/20.11.2943.

[60] L. Lestrade, M. J. Weber, snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs., Nucleic Acids Res 34 (2006) D158–D162. doi:10.1093/nar/gkj002.

[61] J. Erales, V. Marchand, B. Panthu, S. Gillot, S. Belin, S. E. Ghayad, M. Garcia, F. Laforêts, V. Marcel, A. Baudin-Baillieu, P. Bertin, Y. Couté, A. Adrait, M. Meyer, G. Therizols, M. Yusupov, O. Namy, T. Ohlmann, Y. Motorin, F. Catez, J.-J. Diaz, Evidence for rRNA 2'-O-methylation plasticity: Control of intrinsic translational capabilities of human ribosomes., Proc Natl Acad Sci USA 114 (2017) 12934–12939. doi:10.1073/pnas.1707674114.