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 Sequencing Jena
[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 – only recently recognized – also several viral families. To date, no homologous gene of a virus-originated miRNA is known in other organisms. 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 homology- 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

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, several hundred viral miRNAs have been discovered, most of them in DNA viruses[10–14]. Viral

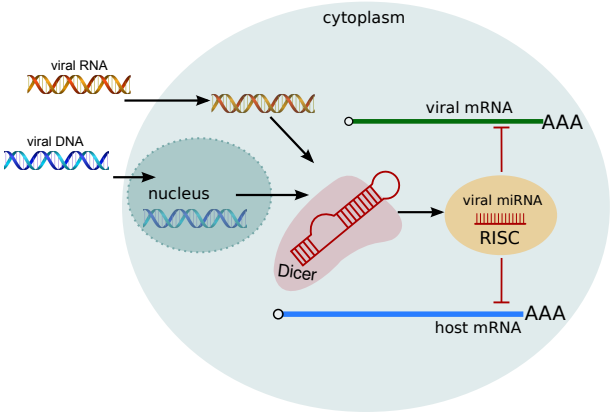

Figure 1: Once the viral genome entered the host cell, transcription, translation and processing mechanisms of the host are exploited. Viral precursor miRNAs are processed by Dicer and loaded into the RISC complex. Interaction with both the viral and the host mRNA is possible.

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 host mRNAs but also viral mRNAs.

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`[18], `miRSeeker`[19] and `miRAlign`[20]. The second group, including tools like `TripletSVM`[21], `NOVOMIR`[22] and `HHMMiR`[23], 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 classifiers.

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 our best knowledge, all of these nine viruses but the HHV7 encode for viral miRNAs[8,9,11,24–27].

HHV7 infections are associated with a number of symptoms such as acute febrile respiratory disease, fever, diarrhea and low lymphocyte counts[28]. Furthermore, there are indications that HHV7 could contribute to the development of drug-induced hypersensitivity syndrome[29], encephalopathy[30], hemiconvulsion-hemiplegia-epilepsy syndrome[31] and hepatitis infection[32]. 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[33].

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 4 gives a schematic overview of the genome organization of EBV with the positions of our predicted candidates.

Methods

Virus Sequence Data

We used eleven different HHV7 genome sequences for our analyses. All sequences were downloaded from the ViPR database[34] (see Table 1). The multiple sequence alignment (MSA) was created with `mafft`[35], 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[34].

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[36]. For each viral precursor structure of the positive set we sampled three sequences from the same virus. 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. 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[37].

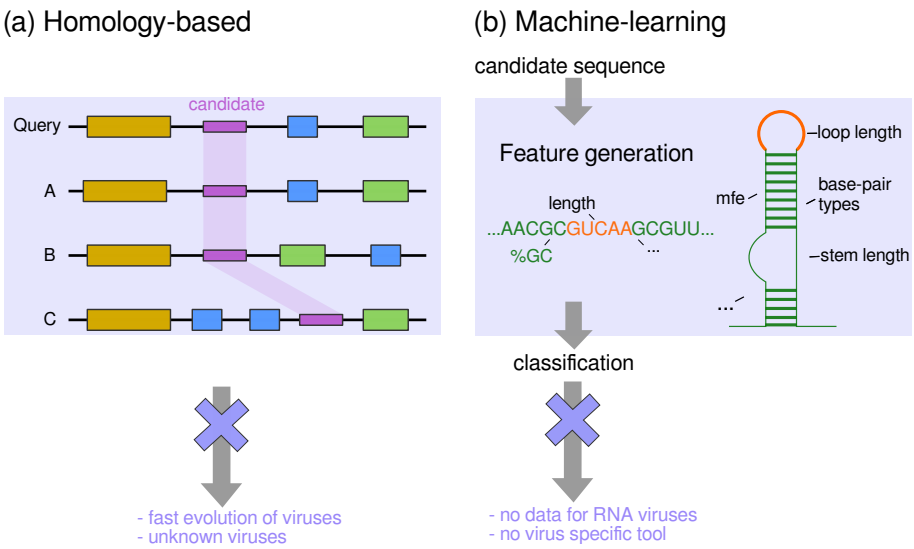

Figure 2: Common approaches for the prediction of miRNAs and their drawbacks on viral data. For both, the homology-based and the machine learning-based approach, the lack of data and the high diversity of viral genomes make these analyses difficult. In our comparison, homology-based approaches without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

Feature discrepancy

Properties derived from the sequence and the structure of a query RNA are called features. Here, we used triplet frequencies (introduced in[21]), 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-Score 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`. The MSA is processed with a sliding window and each window is folded with `RNAalifold 2.4.9`[38] from the ViennaRNA package to determine the consensus secondary structure, the MFE and the z-score derived p-value.

Among the features mentioned above, `ViMiFi` relies on so-called triplet features, proposed in `TripletSVM`[21]. Triplets are suitable to model the sequence and local structure of a query sequence. For each triplet of nucleotides, the second 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: $\frac{.G.}{(((}$, $\frac{.G.}{((.}$, $\frac{.G.}{(..}$, $\frac{.U.}{...}$, $\frac{.C.}{..)}$, $\frac{.C.}{.))}$, $\frac{.C.}{)))}$.

Parameters of `ViMiFi` were set to their respective default values, 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`[38] with parameters `-noLP -p` on each candidate sequence and colored the resulting MFE structure based on the base pairing probabilities derived from the partition function. Thus, we obtain a MFE structure that does not allow for lonely basepairs.

Homolog Search & Function

An exhaustive search on the `Rfam` database version 13.0[39,40] was performed using `Infernal 1.1.2`[41] 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+`[42] 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`.

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

Validation of predictions using small RNA-Seq data

In order to see whether candidates predicted 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)[43]. The libraries are accessible via `http://doi.org/10.5281/zenodo.400950`. Viral reads were mapped with `TopHat2 v2.1.1`[44] with default parameters against the HHV7 reference genome (NCBI accession NC_001716) and data was displayed using the Integrative Genome Browser (IGV)[45].

Results

Known EBV pre-miRNAs are predicted by `ViMiFi`

We examined several different Epstein-Barr virus genomes to identify pre-miRNAs. Figure 4 gives an overview of the coding sequences (CDS) 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 $\begin{smallmatrix} .C. \\ ..(\end{smallmatrix}$, the triplet $\begin{smallmatrix} .A. \\ .((\end{smallmatrix}$ and the loop length. Figure 3 visualizes these differences based on our F-Score 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`. Performing a `blastn` search against `miRBase` only results in 16 different EBV pre-miRNA hits (see Supplement Table S1). This is due to the stringent sequence similarity filter in the `blastn` search. Reducing the `word_size` or increasing the E-Value cutoff gives all 20 pre-miRNAs, however, this introduced too many false positive hits and is thus not further investigated.

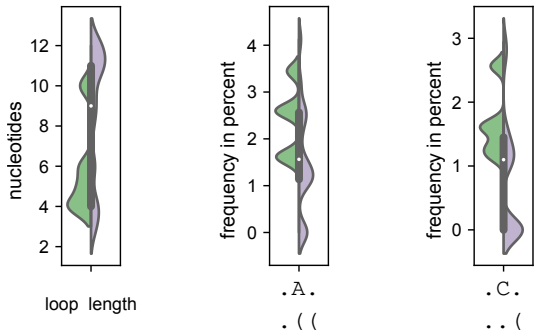

Figure 3: Violinplots of the three features with the highest F-Score between missing (green) and identified (purple) EBV pre-miRNAs. Missed pre-miRNAs tend to have a smaller loop region as well as higher frequencies of the two triplets $\begin{smallmatrix} .A. \\ .((\end{smallmatrix}$ and $\begin{smallmatrix} .C. \\ ..(\end{smallmatrix}$.

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 4, the majority of our predictions cluster in three regions. These regions do not have a CDS 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 7800 – 8000) 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`. We therefore hypothesize that the region between position 7800 and 8000 in the EBV genome may be a potential unknown pre-miRNA cluster. For all results of the `blastn` search we refer to Supplement Table S2.

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

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 S3.

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

family miR-563. This family is conserved among mammals, however, the precise function has not been identified yet. The e-value of our Rfam hits with miR-563 are relatively high, therefore this 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[46,47].

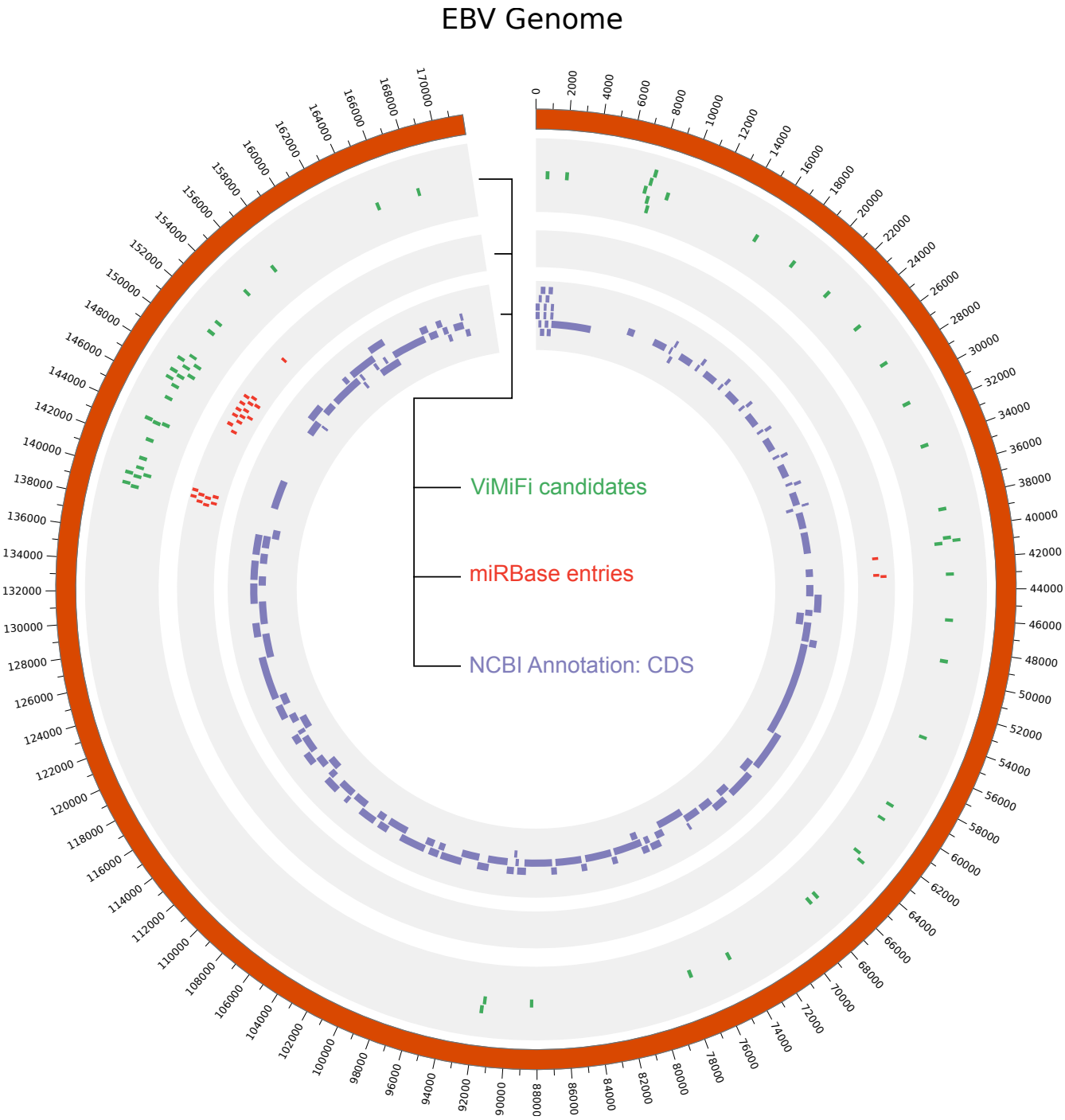

Figure 4: Circular representation of the Epstein-Barr virus genome and annotation of known and predicted pre-miRNAs. The inner track in purple represent annotated coding regions from NCBI. The middle track shows all pre-miRNA entries of the `miRBase` in red. All candidates predicted by `ViMiFi` are represented in green in the outer track.

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 5 shows the positions of the candidates in relation to the CDS annotation of NCBI.

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

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 S4), 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. 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 5 the corresponding green tiles are drawn with a line into the CDS track. Even more intriguing, for the candidates miRNA_HHV7_01 and miRNA_HHV7_18, no `blastn` hits 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 6). 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[48]. As first step of validation, we examined the recently published RNA-Seq data of Lewandowska *et al.*[43]. 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[49]. Nevertheless, this candidate is also worth investigating in our eyes since it shows minor similarities to one of its host's pre-miRNA sequence.

Two HHV7 candidates can be linked to known snoRNAs

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 S5. 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[50], 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[51,52]. The `Infernal` search also found

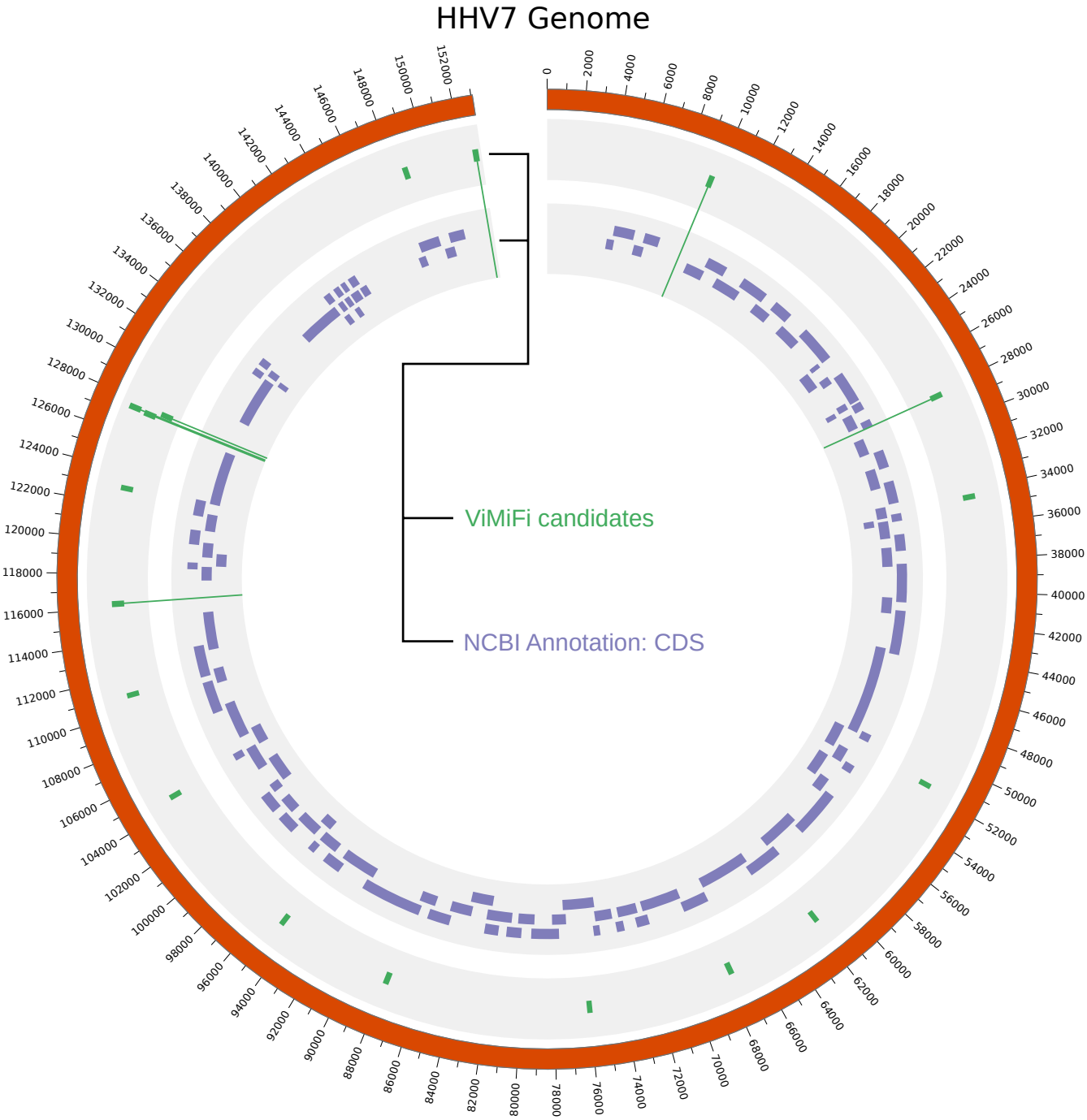

Figure 5: Circular representation of the Human Herpesvirus 7 genome and annotation of predicted pre-miRNAs. The inner track in purple represent annotated coding regions from NCBI. All candidates predicted by `ViMiFi` are represented in green in the outer track. Lines from candidates going into the inner track indicate potential pre-miRNAs that do not overlap with annotated CDS regions.

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 S5.

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

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[48].

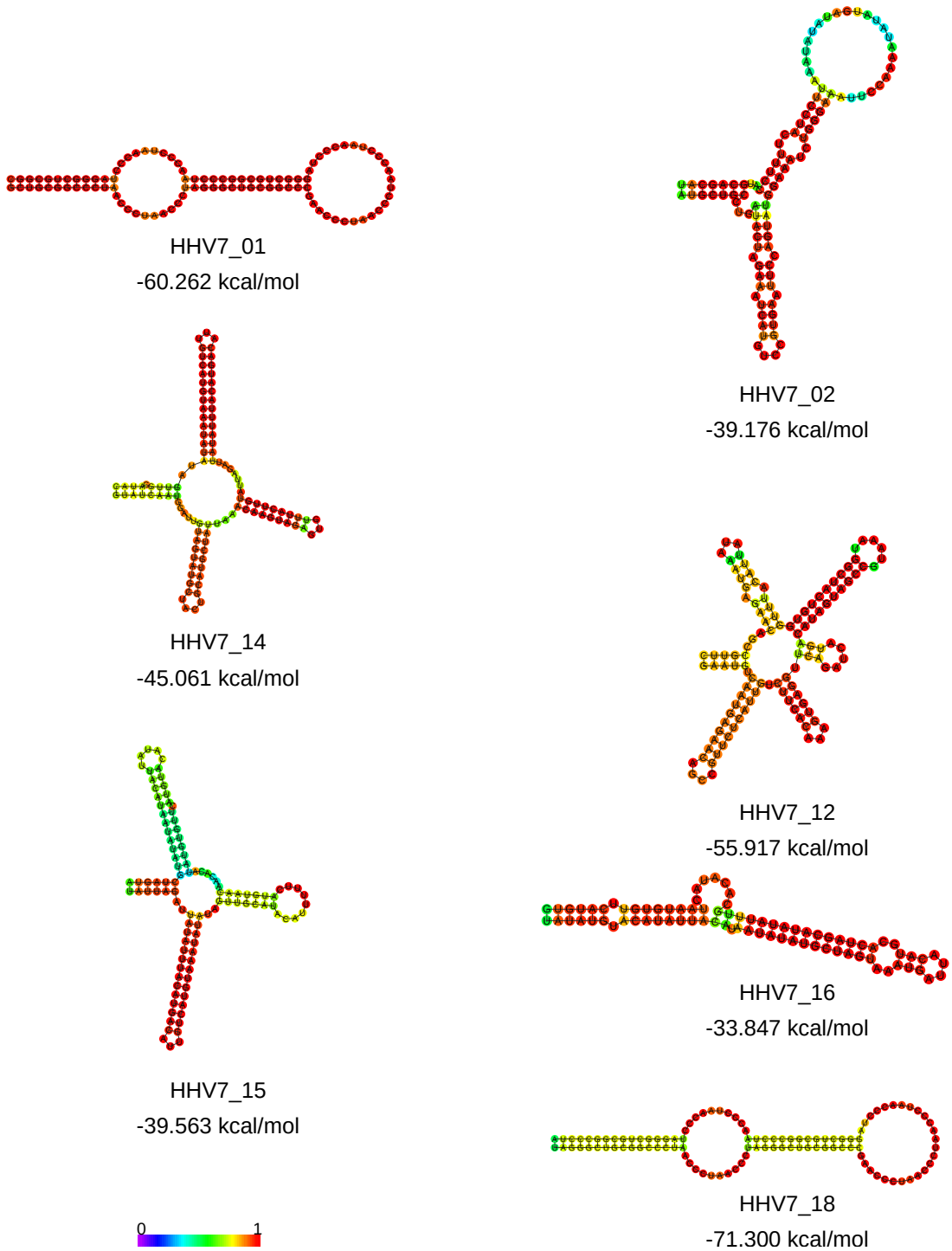

Figure 6: Predicted secondary structures of the potential novel pre-miRNAs encoded by HHV7. Displayed are the minimum free energy structures of all candidates that do not overlap with annotated CDS. Each structure is colored based on the base pairing probabilities derived from the partition function. The more a nucleotide is colored red, the higher the likelihood of the predicted structure.

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"[53]. 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 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. We believe that this tool is a first step towards achieving one of the long-term goal of the new era of virus bioinformatics, which is – among others – the rapid virus-specific bioinformatical software development[54–56].

Acknowledgements 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.

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.

Competing Interests The authors declare no competing interests.

References

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

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

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

[4] György Hutvágner, Juanita McLachlan, Amy E Pasquinelli, Éva Bálint, Thomas Tuschl, and Phillip D Zamore. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*, 293(5531):834–838, 2001.

[5] Alla Grishok, Amy E Pasquinelli, Darryl Conte, Na Li, Susan Parrish, Ilho Ha, David L Baillie, Andrew Fire, Gary Ruvkun, and Craig 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):23–34, 2001.

[6] Yoontae Lee, Kipyoung Jeon, Jun-Tae Lee, Sunyoung Kim, and V Narry Kim. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J*, 21(17): 4663–4670, 2002.

[7] Jacek Krol, Inga Loedige, and Witold Filipowicz. The widespread regulation of microRNA biogenesis, function and decay. *Nat Rev Genet*, 11(9):597–610, 2010.

[8] Sébastien Pfeffer, Mihaela Zavolan, Friedrich A Grässer, Minchen Chien, James J Russo, Jingyue Ju, Bino John, Anton J Enright, Debora Marks, Chris Sander, and Thomas Tuschl. Identification of virus-encoded microRNAs. *Science*, 304:734–736, 2004. ISSN 1095-9203. doi: 10.1126/science.1096781.

[9] Sébastien Pfeffer, Alain Sewer, Mariana Lagos Quintana, Robert Sheridan, Chris Sander, Friedrich A Grässer, Linda F van Dyk, C Kiong Ho, Stewart Shuman, Minchen Chien, et al. Identification of microRNAs of the herpesvirus family. *Nat Methods*, 2(4):269–276, 2005.

[10] Walter Dunn, Phong Trang, Qiu Zhong, Edward Yang, Christopher Van Belle, and Fenyong Liu. Human cytomegalovirus expresses novel microRNAs during productive viral infection. *Cell Microbiol*, 7(11):1684–1695, 2005.

[11] Finn Grey, Andy Antoniewicz, Edwards Allen, Julie Saugstad, Andy McShea, James C Carrington, and Jay Nelson. Identification and characterization of human cytomegalovirus-encoded microRNAs. *J Virol*, 79(18): 12095–12099, 2005.

[12] Bryan R Cullen. Viruses and microRNAs. *Nat Genet*, 38: S25–S30, 2006.

[13] Rebecca L Skalsky and Bryan R Cullen. Viruses, microRNAs, and host interactions. *Annu Rev Microbiol*, 64:123–141, 2010.

[14] Karlie Plaisance-Bonstaff and Rolf Renne. *Antiviral RNAi: Concepts, Methods, and Applications*, volume 721 of *Methods in Molecular Biology*, chapter Viral miRNAs, pages 43–66. Springer, 2011. doi: 10.1007/978-1-61779-037-9.

[15] Patricia Agudelo-Romero, Pablo Carbonell, Miguel A Perez-Amador, and Santiago F Elena. Virus adaptation by manipulation of host's gene expression. *PLOS One*, 3 (6):e2397, 2008.

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

[17] Nathan D Grubaugh, James Weger-Lucarelli, Reyes A Murrieta, Joseph R Fauver, Selene M Garcia-Luna, Abhishek N Prasad, William C Black, and Gregory D Ebel. Genetic drift during systemic arbovirus infection of mosquito vectors leads to decreased relative fitness during host switching. *Cell Host Microbe*, 19(4):481–492, 2016.

[18] Lee P Lim, Nelson C Lau, Earl G Weinstein, Aliaa Abdelhakim, Soraya Yekta, Matthew W Rhoades, Christopher B Burge, and David P Bartel. The microRNAs of *Caenorhabditis elegans*. *Genes Dev*, 17:991–1008, April 2003. ISSN 0890-9369. doi: 10.1101/gad.1074403.

[19] Eric C Lai, Pavel Tomancak, Robert W Williams, and Gerald M Rubin. Computational identification of *Drosophila* microRNA genes. *Genome Biol*, 4:R42, 2003. ISSN 1474-760X. doi: 10.1186/gb-2003-4-7-r42.

[20] Xiaowo Wang, Jing Zhang, Fei Li, Jin Gu, Tao He, Xuegong Zhang, and Yanda Li. MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, 21: 3610–3614, September 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti562.

[21] Chenghai Xue, Fei Li, Tao He, Guo-Ping Liu, Yanda Li, and Xuegong Zhang. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinf*, 6:310, 2005. ISSN 1471-2105. doi: 10.1186/1471-2105-6-310.

[22] Jan-Hendrik Teune and Gerhard Steger. NOVOMIR: de novo prediction of microRNA-coding regions in a single plant-genome. *J Nucleic Acids*, 2010, 2010.

[23] Sabah Kadri, Veronica Hinman, and Panayiotis V Benos. HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. *BMC Bioinf*, 10(S1): S35, 2009.

[24] Igor Jurak, Martha F. Kramer, Joseph C. Mellor, Alison L van Lint, Frederick P Roth, David M Knipe, and Donald M Coen. Numerous conserved and divergent microRNAs expressed by herpes simplex viruses 1 and 2. *J Virol*, 84: 4659–4672, 2010. ISSN 1098-5514. doi: 10.1128/JVI.02725-09.

[25] Masatoshi Nukui, Yasuko Mori, and Eain A Murphy. A human herpesvirus 6A-encoded microRNA: role in viral lytic replication. *J Virol*, 89:2615–2627, 2015. ISSN 1098-5514. doi: 10.1128/JVI.02007-14.

[26] Amos Markus, Linoy Golani, Nishant Kumar Ojha, Tatiana Borodiansky-Shteinberg, Paul R Kinchington, and Ronald S Goldstein. Varicella-Zoster Virus Expresses Multiple Small Noncoding RNAs. *J Virol*, 91, 2017. ISSN 1098-5514. doi: 10.1128/JVI.01710-17.

[27] Xiaoyan Liu, Christine Happel, and Joseph M Ziegelbauer. Kaposi's Sarcoma-Associated Herpesvirus MicroRNAs Target GADD45B To Protect Infected Cells from Cell Cycle Arrest and Apoptosis. *J Virol*, 91, 2017. ISSN 1098-5514. doi: 10.1128/JVI.02045-16.

[28] S Suga, T Yoshikawa, T Nagai, and Y Asano. Clinical features and virological findings in children with primary human herpesvirus 7 infection. *Pediatrics*, 99:E4, 1997. ISSN 1098-4275.

[29] Hiroyuki Hara, Maiko Kobayashi, Ai Yokoyama, Mizuki Tochigi, Akie Matsunaga, Hidenao Shimizu, Junko Goshima, and Hiroyuki Suzuki. Drug-induced hypersensitivity syndrome due to carbamazepine associated with reactivation of human herpesvirus 7. *Dermatology (Basel)*, 211:159–161, 2005. ISSN 1018-8665. doi: 10.1159/000086449.

[30] J S van den Berg, J H van Zeijl, J J Rotteveel, W J Melchers, F J Gabreëls, and J M Galama. Neuroinvasion by human herpesvirus type 7 in a case of exanthem subitum with severe neurologic manifestations. *Neurology*, 52: 1077–1079, 1999. ISSN 0028-3878.

[31] Jun-Ichi Kawada, Hiroshi Kimura, Tetsushi Yoshikawa, Masaru Ihira, Akihisa Okumura, Tsuneo Morishima, and Fumio Hayakawa. Hemiconvulsion-hemiplegia syndrome and primary human herpesvirus 7 infection. *Brain Dev*, 26:412–414, 2004. ISSN 0387-7604. doi: 10.1016/j.braindev.2003.12.003.

[32] T Hashida, E Komura, M Yoshida, T Otsuka, S Hibi, S Imashuku, S Imashuku, T Ishizaki, A Yamada, and S Suga. Hepatitis in association with human herpesvirus-7 infection. *Pediatrics*, 96:783–785, 1995. ISSN 0031-4005.

[33] Isaac W Boss, Karlie B Plaisance, and Rolf Renne. Role of virus-encoded microRNAs in herpesvirus biology. *Trends Microbiol*, 17:544–553, December 2009. ISSN 1878-4380. doi: 10.1016/j.tim.2009.09.002.

[34] Brett E Pickett, Eva L Sadat, Yun Zhang, Jyothi M Noronha, R Burke Squires, Victoria Hunt, Mengya Liu, Sanjeev Kumar, Sam Zaremba, Zhiping Gu, Liwei Zhou, Christopher N Larson, Jonathan Dietrich, Edward B Klem, and Richard H Scheuermann. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res*, 40(Database issue):D593–D598, 2012. doi: 10.1093/nar/gkr859.

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

[36] Ana Kozomara and Sam Griffiths-Jones. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*, 39:D152–D157, 2011. ISSN 1362-4962. doi: 10.1093/nar/gkq1027.

[37] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Dokl Phys*, volume 10, pages 707–710, 1966.

[38] Ronny Lorenz, Stephan H Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA package 2.0. *Algorithms Mol Biol*, 6:26, 2011. ISSN 1748-7188. doi: 10.1186/1748-7188-6-26.

[39] Sam Griffiths-Jones, Alex Bateman, Mhairi Marshall, Ajay Khanna, and Sean R Eddy. Rfam: an RNA family database. *Nucleic Acids Res*, 31:439–441, January 2003. ISSN 1362-4962.

[40] Ioanna Kalvari, Joanna Argasinska, Natalia Quinones-Olvera, Eric P Nawrocki, Elena Rivas, Sean R Eddy, Alex Bateman, Robert D Finn, and Anton I Petrov. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res*, 46:D335–D342, January 2018. ISSN 1362-4962. doi: 10.1093/nar/gkx1038.

[41] Eric P Nawrocki and Sean R Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29:2933–2935, November 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt509.

[42] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. BLAST+: architecture and applications. *BMC Bioinf*, 10:421, December 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-421.

[43] Dagmara W Lewandowska, Peter W Schreiber, Macé M Schuurmans, Bettina Ruehe, Osvaldo Zagordi, Cornelia Bayard, Michael Greiner, Fabienne D Geissberger, Riccarda Capaul, Andrea Zbinden, Jürg Böni, Christian Benden, Nicolas J Mueller, Alexandra Trkola, and Michael Huber. Metagenomic sequencing complements routine diagnostics in identifying viral pathogens in lung transplant recipients with unknown etiology of respiratory infection. *PLoS One*, 12:e0177340, 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0177340.

[44] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, 14 (4):R36, 2013. doi: 10.1186/gb-2013-14-4-r36. URL http://dx.doi.org/10.1186/gb-2013-14-4-r36.

[45] Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings Bioinf*, 14:178–192, 2013. ISSN 1477-4054. doi: 10.1093/bib/bbs017.

[46] M J Farrell, A T Dobson, and L T Feldman. Herpes simplex virus latency-associated transcript is a stable intron. *Proc Natl Acad Sci USA*, 88:790–794, 1991. ISSN 0027-8424.

[47] Walter N Moss, Nara Lee, Genaro Pimienta, and Joan A Steitz. RNA families in Epstein-Barr virus. *RNA Biol*, 11: 10–17, 2014. ISSN 1555-8584. doi: 10.4161/rna.27488.

[48] Ahmed Maher Abdelfattah, Chanhyun Park, and Michael Y Choi. Update on non-canonical microRNAs. *Biomolecular concepts*, 5:275–287, August 2014. ISSN 1868-503X. doi: 10.1515/bmc-2014-0012.

[49] Dereje D Jima, Jenny Zhang, Cassandra Jacobs, Kristy L Richards, Cherie H Dunphy, William W L Choi, Wing Yan Au, Gopesh Srivastava, Magdalena B Czader, David A Rizzieri, Anand S Lagoo, Patricia L Lugar, Karen P Mann, Christopher R Flowers, Leon Bernal-Mizrachi, Kikkeri N Naresh, Andrew M Evens, Leo I Gordon, Micah Luftig, Daphne R Friedman, J Brice Weinberg, Michael A Thompson, Javed I Gill, Qingquan Liu, Tam How, Vladimir Grubor, Yuan Gao, Amee Patel, Han Wu, Jun Zhu, Gerard C Blobe, Peter E Lipsky, Amy Chadburn, Sandeep S Dave, and Hematologic Malignancies Research Consortium. Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs. *Blood*, 116: e118–e127, 2010. ISSN 1528-0020. doi: 10.1182/blood-2010-05-285403.

[50] Anne W Thompson, Katherine Huang, Mak A Saito, and Sallie W Chisholm. Transcriptome response of high- and low-light-adapted Prochlorococcus strains to changing iron availability. *ISME J*, 5:1580–1594, 2011. ISSN 1751-7370. doi: 10.1038/ismej.2011.49.

[51] A Hüttenhofer, M Kiefmann, S Meier-Ewert, J O'Brien, H Lehrach, J P Bachellerie, and J Brosius. RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J*, 20:2943–2953, 2001. ISSN 0261-4189. doi: 10.1093/emboj/20.11.2943.

[52] Laurent Lestrade and Michel J Weber. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res*, 34:D158–D162, 2006. ISSN 1362-4962. doi: 10.1093/nar/gkj002.

[53] Jenny Erales, Virginie Marchand, Baptiste Panthu, Sandra Gillot, Stéphane Belin, Sandra E Ghayad, Maxime Garcia, Florian Laforêts, Virginie Marcel, Agnès Baudin-Baillieu, Pierre Bertin, Yohann Couté, Annie Adrait, Mélanie Meyer, Gabriel Therizols, Marat Yusupov, Olivier Namy, Théophile Ohlmann, Yuri Motorin, Frédéric Catez, and Jean-Jacques Diaz. Evidence for rRNA 2'-O-methylation plasticity: Control of intrinsic translational capabilities of human ribosomes. *Proc Natl Acad Sci USA*, 114:12934–12939, 2017. ISSN 1091-6490. doi: 10.1073/pnas.1707674114.

[54] Franziska Hufsky, Bashar Ibrahim, Martin Beer, Li Deng, Philippe Le Mercier, Dino P McMahon, Massimo Palmarini, Volker Thiel, and Manja Marz. Virologists-heroes need weapons. *PLOS Pathog*, 14:e1006771, 2018. ISSN 1553-7374. doi: 10.1371/journal.ppat.1006771.

[55] Bashar Ibrahim, Dino P McMahon, Franziska Hufsky, Martin Beer, Li Deng, Philippe Le Mercier, Massimo Palmarini, Volker Thiel, and Manja Marz. A new era of virus bioinformatics. *Virus Res*, 251:86–90, 2018. ISSN 1872-7492. doi: 10.1016/j.virusres.2018.05.009.

[56] Bashar Ibrahim, Ksenia Arkhipova, Arno C Andeweg, Susana Posada-Céspedes, François Enault, Arthur Gruber, Eugene V Koonin, Anne Kupczok, Philippe Lemey, Alice C McHardy, et al. Bioinformatics meets virology: The european virus bioinformatics center's second annual meeting. *Viruses*, 10(5):256, 2018.