Bioinformatics Advance Access published June 10, 2016

Prediction of conserved long-range RNA-RNA interactions in full viral genomes

Markus Fricke [1], Manja Marz [1,2*]

[1]Faculty of Mathematics and Computer Science, Friedrich Schiller University Jena, Jena, Germany;
[2]FLI Leibniz Institute for Age Research, Jena, Germany;

Associate Editor: Prof. Cenk Sahinalp

ABSTRACT

Motivation: Long-range RNA-RNA interactions (LRIs) play an important role in viral replication, however, only a few of these interactions are known and only for a small number of viral species. Up to now, it has been impossible to screen a full viral genome for LRIs experimentally or *in silico*. Most known LRIs are cross-reacting structures (pseudoknots) undetectable by most bioinformatical tools.

Results: We present `LRIscan`, a tool for the LRI prediction in full viral genomes based on a multiple genome alignment. We confirmed 14 out of 16 experimentally known and evolutionary conserved LRIs in genome alignments of HCV, Tombusviruses, Flaviviruses and HIV-1. We provide several promising new interactions, which include compensatory mutations and are highly conserved in all considered viral sequences. Furthermore, we provide reactivity plots highlighting the hot spots of predicted LRIs.

Availability: Source code and binaries of `LRIscan` freely available for download at `http://www.rna.uni-jena.de/en/supplements/lriscan/`, implemented in Ruby/C++ and supported on Linux and Windows.

Supplementary information Supporting data is available at `http://www.rna.uni-jena.de/en/supplements/lriscan/`.

Contact: manja@uni-jena.de

1 INTRODUCTION

Long-range RNA-RNA interactions (LRIs) have been marginally reported in various positive strand RNA viruses like Tombusvirus (8 in CIRV and TBSV), Hepacivirus (5 in HCV), Coronavirus (1 in TGEV), Flavivirus (3 in DENV and WNV), Luteovirus (2 in BYDV), Apthovirus (2 in FMDV), Pestivirus (1 in CSFV) and Human immunodeficiency virus (5 in HIV-1) (Huthoff and Berkhout, 2001; Abbink and Berkhout, 2003; Andersen *et al.*, 2004; Ooms *et al.*, 2007; Beerens and Kjems, 2010). According to their definition, a long-range interaction spans distances between a few hundred and several thousands of nucleotides (>26 kb in TGEV). LRIs are often located in loop regions or internal bulges of local RNA structures (known as cis-acting regulatory elements) and therefore build pseudoknot-like structures. Various programs have been developed for general RNA-RNA interaction prediction, which can be classified into five groups (Kato *et al.*, 2010; Seemann *et al.*, 2011): The first group neglects intra-molecular base-pairs, based on the hybrids minimum free energy (MFE). Members of this group are `RNAduplex` and `RNAplex` (Lorenz *et al.*, 2011; Tafer and Hofacker, 2008) or `RNAhybrid` (Rehmsmeier *et al.*, 2004). The second category includes `RNAcofold` (Bernhart *et al.*, 2006) and `PairFold` (Andronescu *et al.*, 2005). These tools concatenate two interacting RNA sequences and calculate the MFE of the joint RNA sequences. The third group considers intra-molecular and inter-molecular RNA-RNA interactions in separated steps, however only one binding site is predicted. Members of this group are `IntaRNA` (Busch *et al.*, 2008) or `RNAup` (Mückstein *et al.*, 2006). The fourth group considers more complex RNA-RNA interactions and allows also more than one binding site. This group includes tools like `RactIP` (Kato *et al.*, 2010), `inteRNA` (Alkan *et al.*, 2006) or `inRNAs` (Salari *et al.*, 2010). The final group contains e.g. `PETcofold` (Seemann *et al.*, 2011), `RNAaliduplex` (Lorenz *et al.*, 2011), `IRBIS` (Pervouchine, 2014), `ripalign` (Li *et al.*, 2011) and `simulfold` (Meyer and Miklós, 2007). These tools consider not only a pair of single sequences, like the tools mentioned above, they use multiple sequence alignments as input. With this comparative method, it is possible to reduce the false-positive rate by incorporating evolutionary conserved information. All of these programs have different properties, unsuitable for viral genomes. For example, `RNAaliduplex` is unable to predict pseudoknots and neglects intra-molecular RNA foldings. `PETcofold` considers both, intra- and inter-molecular interactions as well as pseudoknots, but returns per default only a single secondary structure, which makes the detection of multiple functional binding sites impossible. `ripalign`'s running time makes the program not applicable to viral sequences and `IRBIS` is only applicable for predictions of RNA interactions related to RNA splice sites. Up to now, we are aware of a single program that is designed for LRI prediction, called `CovaRNA` (Bindewald and Shapiro, 2013). This tool detects long-range nucleotide covariation from multiple sequence alignments of eukaryotic genomes using an index-based algorithm to find clusters of covarying base-pairs. The extended function `CovStat` determines the statistical significance of observed covariation cluster. `CovaRNA` has very strict filter criteria and is therefore very conservative in predicting LRIs. For short genomes, such as viral genomes, this leads to almost no predicted interactions.

Here, we present `LRIscan` for detecting long-range RNA-RNA interactions in complete viral genome alignments without prior knowledge. Sparse alignment interaction dotplots in combination

*to whom correspondence should be addressed

with RNAalifold (Bernhart *et al.*, 2008) secondary structure foldings based on minimum free energy calculations are used to predict possible LRIs. We add several filter steps and scoring functions to reduce false positive candidates. We confirm 14 out of 16 experimentally known and evolutionary conserved LRIs in HCV, Tombusviruses, Flaviviruses and HIV-1. We predict several promising new interactions, being highly conserved in all considered viral sequences with multiple compensatory mutations and highly conserved in all considered viral sequences.

2 METHODS

With LRIscan we propose for the first time a method for conserved genome-wide long-range RNA-RNA interaction (LRI) prediction in viral genomes based on a multiple sequence alignment. LRIscan is based on the C-library of the ViennaRNA Package 2.0 (Lorenz *et al.*, 2011). The pipeline consists of four basic steps (see workflow Fig. 1):

(1) Calculate alignment coverage and complexity.

(2) Find LRI seeds with a sparse dotplot method.

(3) Filter LRI candidates based on MFE; calculate z-score/p-value and compensatory score.

(4) Extend seed interaction.

The input of LRIscan is a nucleotide alignment A of length n with m sequences. By A_i we denote the i-th column of the alignment. Entry a_i^k is the k-th row of column i. We define the alignment coverage for each column A_i as percentage of nucleotides over all sequences m, without gaps. We introduce the pairing matrix Π with entries $\Pi_{ij} = 1$ if at least t percent (default $t = 0.95$ for more than 100 sequences, otherwise default $t = 0.80$) of the corresponding sequences can form a base-pair $(a_i^k, a_j^k) \in \{\text{AU, UA, UG, GU, GC, CG}\}$, otherwise $\Pi_{ij} = 0$.

2.1 Alignment cleaning, coverage, complexity

To improve the alignment quality for the nucleotide folding, rarely occurring IUPAC nucleotide ambiguities are replaced by the most occurring valid nucleotides of the same alignment column.

We introduce the coverage matrix Φ with entries $\Phi_{ij} = 1$ if the coverage of columns A_i and A_j is greater than the minimum coverage defined by the user (default 0.5), otherwise $\Phi_{ij} = 0$.

Let δ be the compression function replacing stretches of identical nucleotides by a single nucleotide, e.g. $\delta(AAGUUUCC) = AGUC$.

The complexity for each alignment column A_i is stored by the complexity matrix C computed as

$$C_i = \frac{1}{m} \sum_{k=1}^{m} \frac{|\delta(a^k_{i\ldots i+s-1})|}{|a^k_{i\ldots i+s-1}|}, \qquad (1)$$

where s is the minimum seed interaction length (default: 5 bp). With the complexity score we avoid calculations of regions with gaps or low complexity, such as poly-A/U stretches.

A minimum coverage and complexity threshold can be defined by the user (default: $coverage = 0.5$, $complexity = 0.5$), with a direct effect on run time.

2.2 LRI seed detection

To find LRIs we use a dotplot calculation combined with several base-pairing criteria computed by RNAalifold (Bernhart *et al.*, 2008). To efficiently identify interacting seed regions, we initialize the dotplot seed matrix S with $S_{i,j} = 0$. Each entry $S_{i,j}$ with minimum distance w between column i and j (default $w = 100$ nt) is calculated, following the recursion:

$$S_{i,j} = (S_{i-1,j+1} + \Pi_{ij}) \cdot \Phi_{ij}, \quad 0 < i < n - w, i + w < j < n \qquad (2)$$

$S_{i,j}$ will be considered as seed candidate, if (i) $S_{i+1,j-1} = 0$ and $S_{i,j} \geq s$, (ii) the minimum free energy of the seed alignment, calculated by RNAalifold, is smaller as the maximum MFE defined by the user (default $-10\,kcal/mol$), and (iii) the mean sequence complexity of the seed is greater than the user defined threshold (see red dots in Fig. 1B).

To speed up the seed finding, which needs quadratic time $\mathcal{O}(n^2)$, multiple CPU's can be used. Each CPU calculates a row of 100 nt (see Fig. 1B) overlapping by the minimum seed length s. To save memory, we store only the last valid entry for each seed in a hash.

2.3 LRI seed scoring

For each LRI, we calculate the MFE with RNAalifold. Sequences including only gaps are neglected. Based on the MFE we calculate a z-score to determine the reliability of each LRI, compared to a randomly sampled alignment. The z-score, can be calculated for each predicted LRI as

$$Z = \frac{X - \mu}{\sigma} \qquad (3)$$

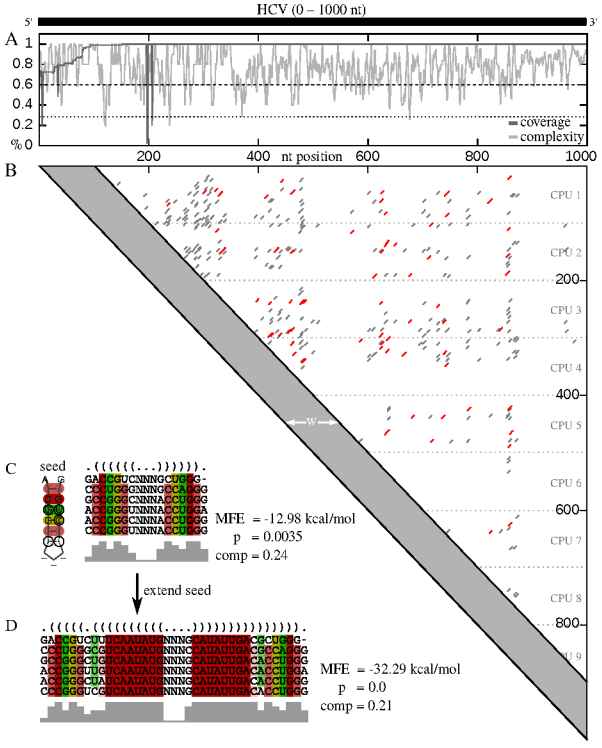

Fig. 1. LRIscan workflow. **(A)** Coverage (dark gray) and complexity (light gray) of the entire alignment. Only regions which pass both, the coverage-threshold (dotted line) and the complexity-threshold (dashed line) are considered for further calculations. **(B)** Dotplot containing all possible seed interactions (gray and red lines) without gaps and a given minimum interaction length. We calculate only interactions with a distance $> w$. To decrease the run time multiple CPUs are used, calculating only a specific range of the dotplot matrix. All ranges overlap by the minimum seed length. **(C)** For all seeds we calculate the minimum free energy (MFE) of the alignment with RNAalifold. For each seed passing the MFE threshold (red dots in B) we calculate a p-value based on the z-score and the compensatory score as defined in the methods. **(D)** Seeds are extended towards both sides. An extended z-score/p-value and compensatory score is calculated.

Pseudocode 1. Algorithm to find seed interactions in a multiple genome alignment.

```
S[i,j] = 0
for i in 1...n-w
  if PHI(i) > 0
    for j in i+w...n
      if PHI(j) > 0 and PI(i,j) >= 0
        S[i,j] = S[i-1,j+1] + 1
      else
        if S[i-1,j+1] >= MIN_SEED_LENGTH
        and C(S[i-1,j+1]) > MIN_COMPLEX
        and mfe(S[i-1,j+1]) < MAX_MFE
          return S[i-1,j+1]
        end
      end
    end
  end
end
```

where X is the MFE of the corresponding interaction, μ the mean MFE and σ the corresponding standard deviation of the z-times randomly swap-shuffled alignments.

For each consensus base-pair $b(A_i, A_j)$ of a given seed (of length $|b|$), we determine the compensatory score τ to find LRIs with a high amount of compensatory mutations coincident with a high amount of compatible base-pairs. We consider 1 nt changes (e.g. AU to GU), which preserve a base-pairing, as well as 2 nt changes (e.g. AU to GC) as compensatory mutations.

For each consensus base-pair, u is the number of different base-pair types ($u \leq 6$) and h is the number of compatible base-pairs ($h \leq m$). We normalize by the maximum number of different base-pairs over all consensus base-pairs ($|b|$) for all sequences k included in the LRI ($k \leq m$).

$$\tau = \frac{\sum_b (u \cdot h)}{6 \cdot |b| \cdot k} \tag{4}$$

2.4 LRI seed extension

For each LRI seed, we attempt to extend the alignment by 10 bp at the 5' and 3' end. For the extended alignment, the MFE is calculated by `RNAalifold`, with a hard constraint to build given seed base-pairs and a soft constraint which forces an inter-molecular interaction of the surrounding base-pairs. For the extended alignment a separate z-score and p-value calculation is possible (default off). These scores are independent from seed scores.

2.5 Output results

All resulting LRIs are presented in a tab-separated file and additionally as HTML table, linking all corresponding figures to allow the user to browse through the output.

The interacting alignment positions are calculated back to the original position of each virus isolate to easily assess the interactions.

2.6 CovaRNA

To compare `LRIscan` to `CovaRNA` we converted all alignments to the UCSC MAF format and used `CovaRNA` with a minimum number of two input sequences, reading out of the MAF input file (`-s 2`).

2.7 Dataset

For HCV, we downloaded 950 genomes from NCBI and the `HCV database`[1] (v. 2008 (Kuiken *et al.*, 2005)). Sequences which occur in both data sets were considered only once. For Tombusvirus and Flavivirus, we downloaded all genomes listed as complete genomes at NCBI-Genome, resulting in 13 sequences for Tombusvirus and six sequences for Flavivirus (mosquito/vertebrate Flaviviruses).

For the 950 HCV sequences, an alignment was generated with `MAFFT --auto`, v.6.8 (Katoh *et al.*, 2002) and a phylogenetic tree was built with `Geneious`, v.6.1 (Kearse *et al.*, 2012) Neighbor-Joining method `Tamura-Nei` (Tamura and Nei, 1993). Based on this tree and the annotations from `NCBI` and the `HCV database` the dataset was reduced to two of the longest genomes from each subtype if available, resulting in 106 sequences from 65 subtypes.

For Tombusvirus (13 sequences), Flavivirus (6 sequences) and the reduced HCV set (106 sequences), we generated `MAFFT --maxiterate 1000 --localpair` alignments as input for `LRIscan`.

For HIV we downloaded the hand curated compendium alignment from the `HIV sequence database`[2] choosing all subtypes of the HIV-1 alignment of 2014, resulting in an alignment of 200 sequences.

All input alignments can be found at the supplemental page[3].

2.8 Sensitivity and specificity

To calculate the sensitivity and specificity of `LRIscan` we performed di-nucleotide shuffling of each sample alignment with `multiperm` (Anandam *et al.*, 2009) and hid true positive LRIs. We defined all experimentally detected LRIs as true positives, as well as the top LRIs of the original alignment. We applied `LRIscan` with the same parameters as with the original alignments.

3 RESULTS AND DISCUSSION

To validate our tool, we chose the four positive strand RNA viruses with known LRIs: Hepatitis C virus (HCV), Flavivirus, Tombusvirus and Human immunodeficiency virus (HIV). The best studied viruses, with highest number of known LRIs, are HCV and Tombusviruses. In HCV, five LRIs are experimentally verified (Fig. 2B) and another twelve LRIs have been predicted semi-manually in Fricke *et al.* (2015). Currently, eight LRIs in Tombusvirus (Fig. 5B), three LRIs in Flaviviruses (Fig. 4B) and two LRIs in HIV-1 (Fig. 6B) have been experimentally verified. As input for `LRIscan` we used alignments of 106 HCV genomes, 13 Tombusvirus genomes, 6 Flavivirus genomes and 200 HIV-1 genomes.

3.1 HCV

To determine the LRIs of the HCV alignment, we used `LRIscan` with default parameters. Due to general assembly problems of the 5'/3' UTR of viral sequences, the HCV alignment consists of only 19 complete sequences. Therefore, we set the minimum number of involved sequences to 17% (18 sequences), resulting in 311 predicted LRIs (74 LRIs with $p < 0.05$).

We plotted all detected LRIs of HCV, passing the p-value threshold, to the corresponding genome alignment position

[1] `http://hcv.lanl.gov/content/sequence/NEWALIGN/align.html`

[2] `http://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html`

[3] `http://www.rna.uni-jena.de/supplements/LRIscan/`

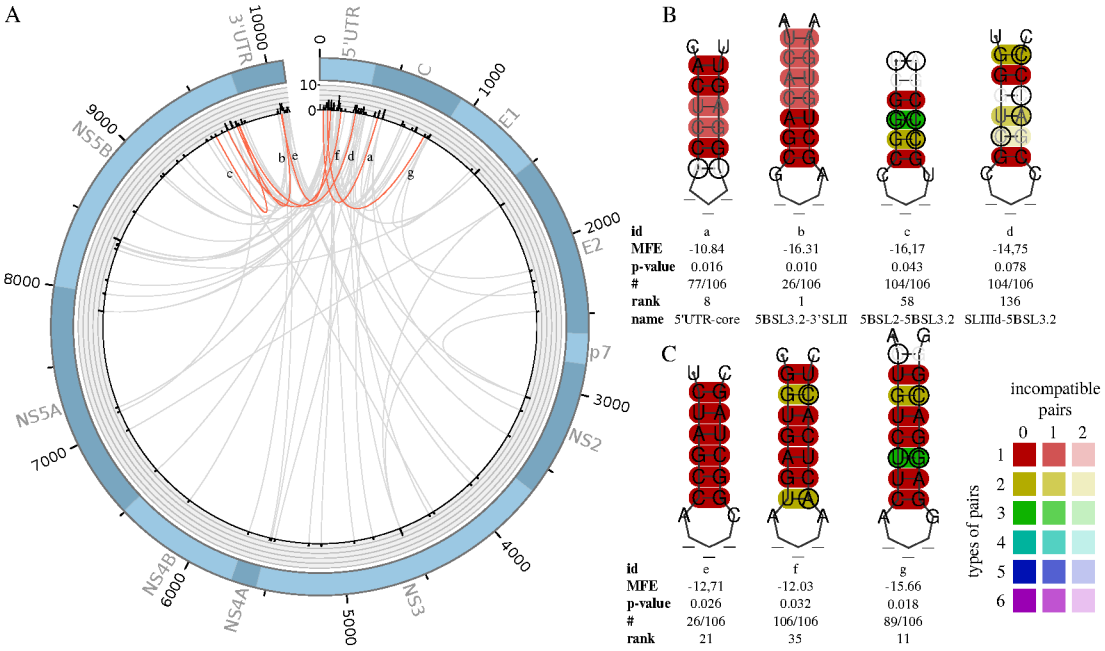

Fig. 2. **(A)** Plot of all predicted LRIs with $p < 0.05$ (74) found in the HCV alignment of 106 sequences. The outer circle represents the genome. The histogram represents the number of LRIs per alignment position. High reactive genome positions can be found in the 5'/3' UTR and the coding region of the core gene *C*. The inner circle shows all predicted interactions between all genome positions. gray – all new LRIs; colored – LRIs corresponding to B and C. The plot was created with `Circos` (Krzywinski *et al.*, 2009). **(B)** Experimentally verified LRIs, which can be predicted by `LRIscan`, named SLIIId-5BSL3.2 (Romero-López and Berzal-Herranz, 2012, 2009), 5BSL2-5BSL3.2 (Romero-López *et al.*, 2014; Tuplin *et al.*, 2012), 5BSL3.2-3'SLII (Friebe *et al.*, 2005) and 5'UTR-core (Beguiristain *et al.*, 2005; Honda *et al.*, 1999). **(C)** Highly interesting new LRIs predicted by `LRIscan`. In a former study, we suggested that LRI *2e* could be a seed interaction for a HCV genome circularization (Fricke *et al.*, 2015). A complete list including all predicted LRIs can be found at the supplemental page. Colors are used to indicate conserved base-pairs: from red (no variation of a base-pair within the alignment) to purple (all six base-pair types are found); from dark (all sequences contain this base-pair) to light colors (1 or 2 sequences are unable to form this base-pair). Compensatory mutations are marked by a circle around the variable base(s).

(Fig. 2A) to detect regions with a high number of interactions. Interestingly, we detected some highly reactive regions in the 5' UTR and 3' UTR, but also high reactivity sites in the CDS of the core and NS5B protein coding region. The UTRs and the core region show the highest amount of possible LRIs, in agreement with the known highly structured regions for all HCV subtypes (Fricke *et al.*, 2015). Consistent with regions of high reactivity (superior interacting regions), it has been shown that the 5' UTR includes three LRIs (Filbin and Kieft, 2011; Romero-López and Berzal-Herranz, 2012, 2009; Beguiristain *et al.*, 2005; Honda *et al.*, 1999).

For HCV, five LRIs have been experimentally verified: SLII-SLIV (Filbin and Kieft, 2011), SLIIId-5BSL3.2 (Romero-López and Berzal-Herranz, 2012, 2009), 5BSL2-5BSL3.2 (Romero-López *et al.*, 2014; Tuplin *et al.*, 2012), 5BSL3.2-3'SLII (Friebe *et al.*, 2005) and 5'UTR-core (Beguiristain *et al.*, 2005; Honda *et al.*, 1999). We identified three known LRIs within the first 58 hits ranked by p-value (Fig. 2B). We missed SLII-SLIV because of the unfavourable seed MFE of only -7 kcal/mol (default threshold -10 kcal/mol). The LRI SLIIId-5BSL3.2 did not pass our conservative p-value threshold, but is part of the `LRIscan` output (LRI 2d). Increasing the MFE or p-value threshold would increase the amount of LRIs dramatically, very likely resulting in a high false positive rate (Fig. 3).

We present three highly interesting new LRIs based on p-value, compensatory score and location in the HCV genome (Fig. 2C).

The LRI *g* of Fig. 2 (LRI *2g*) is one of the best ranked LRI with a very long seed interaction of 8 nt and three compensatory base-pairs, whereof one base-pair consists of three different types of base-pairs changed in both sites. This LRI is conserved in 89/106 isolates and spans a distance of 754 nt between 5' UTR and the coding region of the core protein. An experimental verification of the interactions would be highly recommended.

LRI *2f* is also highly conserved in all isolates and spans a distance of 9440 nt, connecting the 5'UTR with the NS5B coding regions (corresponding to LRI 4 in Fricke *et al.* (2015)). We identified also the possible genome circularization *2e*) between 5'SLII and 3'DLS (see Fricke *et al.* (2015)).

3.2 Flavivirus

For Flavivirus genomes we used `LRIscan` with default values. We predicted 113/157 LRIs with a p-value smaller 0.05. Although Flaviviruses and Hepacivirus are both assigned to the family *Flaviviridae*, the LRI distribution is very different. In Flaviviruses, we found some clearly separated peaks with an accumulation of

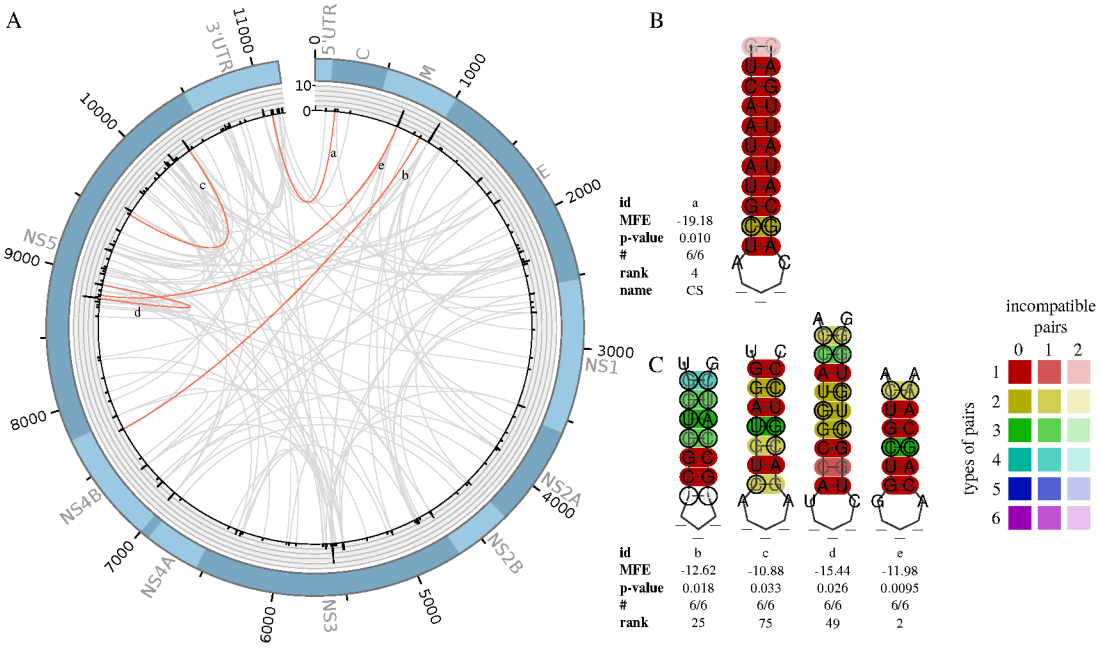

Fig. 4. **(A)** Plot of all predicted LRIs with $p < 0.05$ (113) found in the Flavivirus alignment of 6 sequences. The outer circle represents the genome. The histogram represents the number of LRIs per alignment position. The inner circle shows all predicted interactions between all genome positions. gray – all new LRIs; colored – LRIs corresponding to B and C. The plot was created with `Circos` (Krzywinski *et al.*, 2009). **(B)** Experimentally verified LRI, which can be predicted by `LRIscan`, named 5'-3' CS (Friebe and Harris, 2010). **(C)** Highly interesting new LRIs predicted by `LRIscan`. A complete list including all predicted LRIs can be found at the supplemental page.

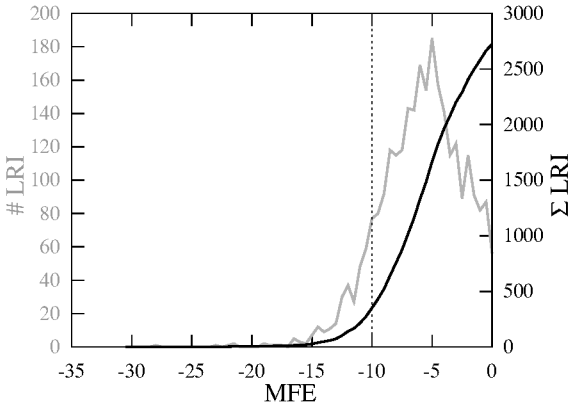

Fig. 3. Plot of predicted LRIs in the HCV alignment using different MFE thresholds. The default threshold is −10 kcal/mol resulting in 311 LRIs. With increasing MFE threshold, the number of predicted LRI increases dramatically and therewith the false positive rate. Thus, we decided to choose a conservative MFE threshold. black – cumulative sum of LRIs per MFE threshold. gray – sum of LRIs per MFE threshold

LRIs, located at the 3' end of gene M and in the center of the NS3 and NS5 gene (Fig. 4A). Different from the HCV alignment, we did not find an accumulation of LRIs in the 5' UTR and 3' UTR. This effect could be explained by different sequence conservations of the UTRs. In HCV the pairwise identity of the UTRs is $>$ 95 %, whereas the UTR pairwise identity of the Flaviviruses is $<$ 50 %. The Flavivirus alignment consists of only six sequences, including only the mosquito/vertebrate Flaviviruses. In Flavivirus three experimentally verified LRIs are known: 5'-3'-UAR, DAR and 5'-3'-CS (Khromykh *et al.*, 2001; Zhang *et al.*, 2008; Alvarez *et al.*, 2005; Corver *et al.*, 2003; You *et al.*, 2001; Friebe and Harris, 2010). All known interactions are in close proximity and can build a genome circularization, essential for the viral replication. The strongest interaction (CS), was ranked at fourth position (Fig. 4B). It was not possible to identify the 5'-3'-UAR and DAR, because the seed region does not appear to be conserved in the highly variable 5' UTR.

However, we found several new promising LRIs (Fig. 4C) in the considered Flaviviruses. All depicted interactions are ranked among the first 75 hits and show a high amount of compensatory mutations (up to four types of base-pairs) and being conserved in all sequences.

3.3 Tombusvirus

For the genomes of the Tombusviridae we used `LRIscan` with default parameters. We found 529 LRIs (213 LRIs with $p < 0.05$). Most of the LRIs in Tombusvirus are located in the p92 region (Fig. 5A). This is in line with the already known interactions, where four out of eight known LRIs start in the p92 region (Cimino *et al.*, 2011; Wu *et al.*, 2009; Lin and White, 2004). But also the intergenic regions between p41/p22 and the 3' end of the p92 coding regions show high reactivity (Fig. 5A). These areas harbor the known

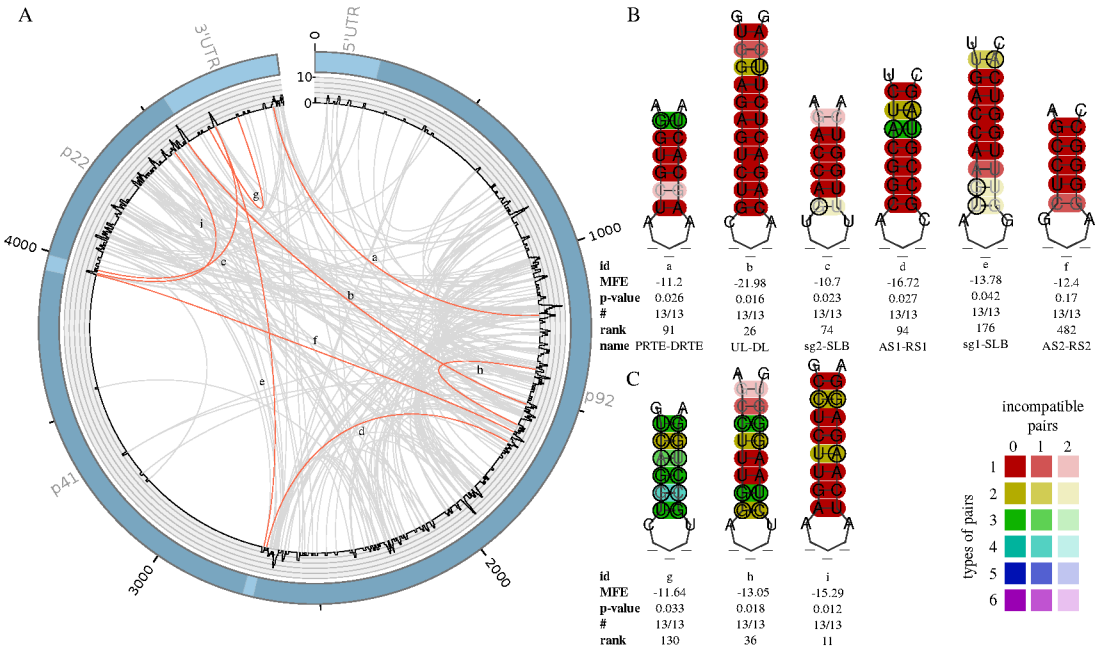

Fig. 5. **(A)** Plot of all predicted LRIs with $p < 0.05$ (213) found in the Tombusvirus alignment of 13 sequences. The outer circle represents the genome. The histogram represents the number of LRIs per alignment position. The inner circle shows all predicted interactions between all genome positions. gray – all new LRIs; colored – LRIs corresponding to B and C. The plot was created with `Circos` (Krzywinski *et al.*, 2009). **(B)** Experimentally verified LRIs, which have been also predicted by `LRIscan`, named SL3-SLB (Fabian and White, 2004, 2006; Nicholson and White, 2008), PRTE-DRTE (Cimino *et al.*, 2011), UL-DL (Wu *et al.*, 2009), sg2-SLB (Fabian and White, 2004), AS1-RS1 (Lin and White, 2004), AS2-RS2 (Lin and White, 2004). **(C)** Highly interesting new LRIs predicted by `LRIscan`. A complete list including all predicted LRIs can be found at the supplemental page.

interacting regions of the AS1/RS1, DE/CE, sg1/SLB, AS2/RS2, sg2/SLB (Wu *et al.*, 2009; Lin and White, 2004; Fabian and White, 2004). In contrast, the p41 region contains only a few predicted LRIs. This is due to the very variable sequence of this gene, which encodes the coat protein of the Tombusviruses.

In addition to the mentioned five, three more LRIs are known from experimental data for Tombusviruses: SL3-SLB (Fabian and White, 2004, 2006; Nicholson and White, 2008), PRTE-DRTE (Cimino *et al.*, 2011), UL-DL (Wu *et al.*, 2009). We detected six of the eight known LRIs (Fig. 5B). AS2-RS2 was ranked very low due to a very short seed of 5 bp, a low p-value and a poly-G stretch. The missing SL3-SLB and DE-CE have been only described for one species and are also manually not discoverable in other species. Both interactions are located in variable regions with low conserved RNA sequences. Here, we present three novel LRIs (Fig. 5C). The LRI *5g* with strong compensatory mutations in 6 of 6 base-pairs and up to four types is conserved in all sequences. The LRIs *5h* and *5i* are conserved in all sequences and have a very high rank, as well as compensatory base-pairs with up to three types of base-pairs mutated at both interaction sites. The introduced LRIs are interesting candidates for further wet lab studies.

3.4 HIV

For the HIV alignment, `LRIscan` was used with default parameters. Only 20% of HIV genomes (40 sequences) contained a complete 5' and 3' UTR, therefore we decreased the minimum involved sequences to 20%. With these settings we identified 314 LRIs (115 LRIs with $p < 0.05$).

In HIV, five LRIs are known termed R-GAG, LDI, U5-AUG, TAR-TAR and GAG-U3R (Huthoff and Berkhout, 2001; Abbink and Berkhout, 2003; Andersen *et al.*, 2004; Ooms *et al.*, 2007; Beerens and Kjems, 2010). We identified three of the known interactions: U5-AUG, TAR-TAR and LDI, (Fig. 6B). Due to the conservative MFE threshold it was not possible to detect the conserved R-GAG interaction. For the GAG-U3R interactions no conserved seed interaction exists.

We also suggest three novel LRIs (Fig. 6C). These LRIs are conserved in all 200 sequences and have several compensatory mutations. LRI *6d* is ranked at position one and has compensatory mutations in 7 out of 9 base-pairs with up to 3 different base-pair types.

3.5 Comparison to `CovaRNA`

We compare `LRIscan` to `CovaRNA`, designed for large eukaryotic genomes. `CovaRNA` outputs no reliable results for the small viral genomes. For the HCV and for the Tombusvirus alignment, `CovaRNA` found only one covariation cluster consisting of only two base-pairs. This cluster covers no known LRI. No cluster could be identified for the HIV alignment. The output of the Flavivirus alignment includes five covariation clusters. All clusters contain only two base-pairs and cover none of the known LRIs.

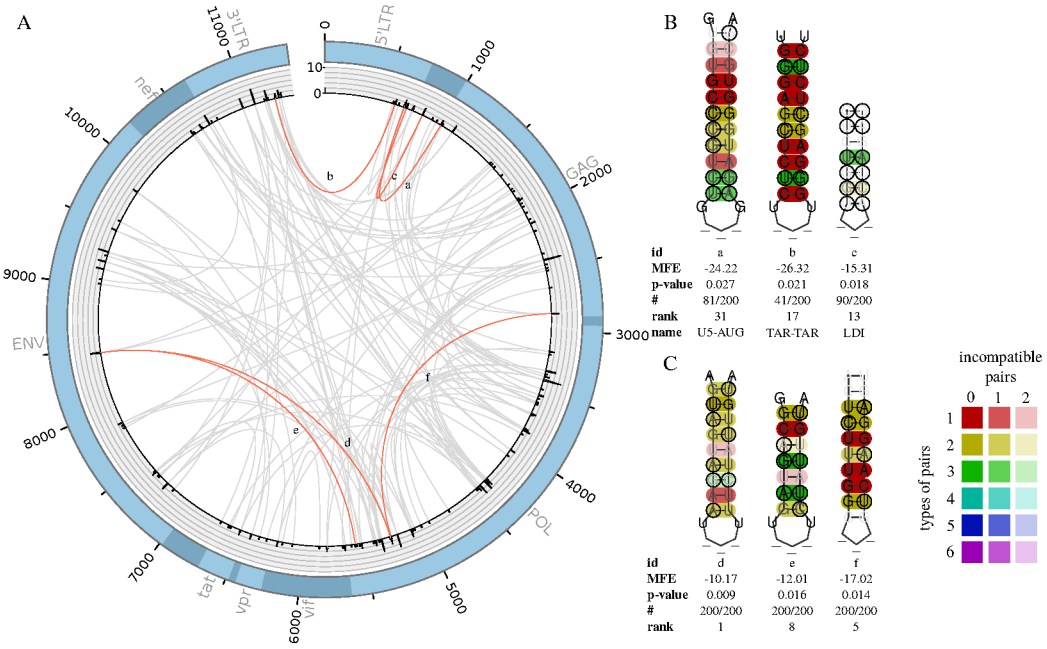

Fig. 6. **(A)** Plot of all predicted LRIs with $p < 0.05$ (115) found in the HIV-1 alignment of 200 sequences. The outer circle represents the genome. The histogram represents the number of LRIs per alignment position. The inner circle shows all predicted interactions between all genome positions. gray – all new LRIs; colored – LRIs corresponding to B and C. The plot was created with `Circos` (Krzywinski *et al.*, 2009). **(B)** Experimentally verified LRIs, which can be predicted by `LRIscan`, named LDI, U5-AUG and TAR-TAR (Huthoff and Berkhout, 2001; Abbink and Berkhout, 2003; Andersen *et al.*, 2004; Beerens and Kjems, 2010). **(C)** Highly interesting new LRIs predicted by `LRIscan`. A complete list including all predicted LRIs can be found at the supplemental page.

Table 1. Plot of sensitivity and specificity ($p < 0.05$) of the four shuffled genome alignments of HCV, Flaviviruses, Tombusviruses and HIV.

	#seq	identity %	TP	sensitivity	specificity
HCV	106	62.2	10	0.81	0.69
Flaviviruses	6	58.3	10	0.83	0.46
Tombusviruses	13	63.5	23	0.88	0.75
HIV	200	80.0	16	0.81	0.69
Mean				0.83	0.64

3.6 Sensitivity and specificity

An accurate sensitivity/specificity calculation is difficult due to the limited number of experimentally verified LRIs. We reach a mean sensitivity of 0.83. Most of the non-detected true positive LRIs are specific and experimentally verified only for single isolates. We investigated their conservation throughout the individuals of the alignment manually, resulting in isolate-specific LRIs. The sensitivity is independent on the sequence number and alignment identity (see Tab. 1). The mean specificity of 0.64 depicts a rather high number of false positives. A more stringent p-value would remove a large fraction of false positives, however, results also in a loss of true positive LRIs, see Fig. 2, 4-6. In practice, known LRIs have very small p-values (high ranks) and/or high compensatory scores and/or can be extended in length. The user selects the LRIs based on all metrics and on the regions of interests. A minimal p-value threshold including almost all experimental verified LRIs can be found at 0.05.

4 CONCLUSIONS

The identification of long-range RNA-RNA interactions is experimentally and bioinformatically a challenging task. The huge amount of theoretically possible interactions makes it impossible to verify all possible interactions by wet lab experiments. `LRIscan` offers the opportunity to predict possible LRIs under certain criteria and reduces dramatically the number of candidates for wet lab studies. As shown in part *A* of Fig. 2-6, also well studied viruses provide a huge list of highly ranked LRIs, which could be involved in viral translation and transcription mechanisms. Further wet lab studies, which verify the functionality of these interactions, could improve our understanding of the viral replication.

ACKNOWLEDGEMENTS

This work has been partially financed by Carl Zeiss Stiftung.

Supporting Data

Supporting data is available at `http://www.rna.uni-jena.de/en/supplements/lriscan/`.

Conflict of interest statement

None declared.

REFERENCES

Abbink, T. E. and Berkhout, B. (2003). A novel long distance base-pairing interaction in human immunodeficiency virus type 1 RNA occludes the Gag start codon. *J Biol Chem*, **278**(13), 11601–11611.

Alkan, C., Karako, E., Nadeau, J. H., Sahinalp, S. C., and Zhang, K. (2006). RNA-RNA interaction prediction and antisense RNA target search. *J Comput Biol*, **13**(2), 267–282.

Alvarez, D. E., Lodeiro, M. F., Ludueña, S. J., Pietrasanta, L. I., and Gamarnik, A. V. (2005). Long-range RNA-RNA interactions circularize the dengue virus genome. *J Virol*, **79**(11), 6631–6643.

Anandam, P., Torarinsson, E., and Ruzzo, W. L. (2009). Multiperm: shuffling multiple sequence alignments while approximately preserving dinucleotide frequencies. *Bioinformatics*, **25**(5), 668–669.

Andersen, E. S., Contera, S. A., Knudsen, B., Damgaard, C. K., Besenbacher, F., and Kjems, J. (2004). Role of the trans-activation response element in dimerization of HIV-1 RNA. *J Biol Chem*, **279**(21), 22243–22249.

Andronescu, M., Zhang, Z. C., and Condon, A. (2005). Secondary structure prediction of interacting RNA molecules. *J Mol Biol*, **345**(5), 987–1001.

Beerens, N. and Kjems, J. (2010). Circularization of the HIV-1 genome facilitates strand transfer during reverse transcription. *RNA*, **16**(6), 1226–1235.

Beguiristain, N., Robertson, H. D., and Gómez, J. (2005). RNase III cleavage demonstrates a long range RNA: RNA duplex element flanking the hepatitis C virus internal ribosome entry site. *Nucleic Acids Res*, **33**(16), 5250–5261.

Bernhart, S. H., Tafer, H., Mückstein, U., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2006). Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol*, **1**(1), 3.

Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R., and Stadler, P. F. (2008). RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.

Bindewald, E. and Shapiro, B. A. (2013). Computational detection of abundant long-range nucleotide covariation in Drosophila genomes. *RNA*, **19**(9), 1171–1182.

Busch, A., Richter, A. S., and Backofen, R. (2008). IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, **24**(24), 2849–2856.

Cimino, P. A., Nicholson, B. L., Wu, B., Xu, W., and White, K. A. (2011). Multifaceted regulation of translational readthrough by RNA replication elements in a tombusvirus. *PLoS Pathog*, **7**(12).

Corver, J., Lenches, E., Smith, K., Robison, R. A., Sando, T., Strauss, E. G., and Strauss, J. H. (2003). Fine mapping of a cis-acting sequence element in yellow fever virus RNA that is required for RNA replication and cyclization. *J Virol*, **77**(3), 2265–2270.

Fabian, M. R. and White, K. A. (2004). 5'-3' RNA-RNA interaction facilitates cap- and poly(A) tail-independent translation of tomato bushy stunt virus mRNA: a potential common mechanism for tombusviridae. *J Biol Chem*, **279**(28), 28862–28872.

Fabian, M. R. and White, K. A. (2006). Analysis of a 3'-translation enhancer in a tombusvirus: a dynamic model for RNA-RNA interactions of mRNA termini. *RNA*, **12**(7), 1304–1314.

Filbin, M. E. and Kieft, J. S. (2011). HCV IRES domain IIb affects the configuration of coding RNA in the 40S subunit's decoding groove. *RNA*, **17**(7), 1258–1273.

Fricke, M., Dünnes, N., Zayas, M., Bartenschlager, R., Niepmann, M., and Marz, M. (2015). Conserved RNA secondary structures and long-range interactions in hepatitis C viruses. *RNA*, **21**(7), 1219–1232.

Friebe, P. and Harris, E. (2010). Interplay of RNA elements in the dengue virus 5' and 3' ends required for viral RNA replication. *J Virol*, **84**(12), 6103–6118.

Friebe, P., Boudet, J., Simorre, J. P., and Bartenschlager, R. (2005). Kissing-loop interaction in the 3' end of the hepatitis C virus genome essential for RNA replication. *J Virol*, **79**(1), 380–392.

Honda, M., Beard, M. R., Ping, L. H., and Lemon, S. M. (1999). A phylogenetically conserved stem-loop structure at the 5' border of the internal ribosome entry site of hepatitis C virus is required for cap-independent viral translation. *J Virol*, **73**(2), 1165–1174.

Huthoff, H. and Berkhout, B. (2001). Mutations in the TAR hairpin affect the equilibrium between alternative conformations of the HIV-1 leader RNA. *Nucleic Acids Res*, **29**(12), 2594–2600.

Kato, Y., Sato, K., Hamada, M., Watanabe, Y., Asai, K., and Akutsu, T. (2010). RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming. *Bioinformatics*, **26**(18), 460–466.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*, **30**(14), 3059–3066.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., and Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**(12), 1647–1649.

Khromykh, A. A., Meka, H., Guyatt, K. J., and Westaway, E. G. (2001). Essential role of cyclization sequences in flavivirus RNA replication. *J Virol*, **75**(14), 6719–6728.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res*, **19**(9), 1639–1645.

Kuiken, C., Yusim, K., Boykin, L., and Richardson, R. (2005). The Los Alamos hepatitis C sequence database. *Bioinformatics*, **21**(3), 379–384.

Li, Y. P., Gottwein, J. M., Scheel, T. K., Jensen, T. B., and Bukh, J. (2011). MicroRNA-122 antagonism against hepatitis C virus genotypes 1-6 and reduced efficacy by host RNA insertion or mutations in the HCV 5' UTR. *Proc Natl Acad Sci U S A*, **108**(12), 4991–4996.

Lin, H. X. and White, K. A. (2004). A complex network of RNA-RNA interactions controls subgenomic mRNA transcription in a tombusvirus. *EMBO J*, **23**(16), 3365–3374.

Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms Mol Biol*, **6**, 26–26.

Meyer, I. M. and Miklós, I. (2007). SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS Comput Biol*, **3**(8).

Mückstein, U., Tafer, H., Hackermüller, J., Bernhart, S. H., Stadler, P. F., and Hofacker, I. L. (2006). Thermodynamics of RNA-RNA binding. *Bioinformatics*, **22**(10), 1177–1182.

Nicholson, B. L. and White, K. A. (2008). Context-influenced cap-independent translation of Tombusvirus mRNAs in vitro. *Virology*, **380**(2), 203–212.

Ooms, M., Abbink, T. E., Pham, C., and Berkhout, B. (2007). Circularization of the HIV-1 RNA genome. *Nucleic Acids Res*, **35**(15), 5253–5261.

Pervouchine, D. D. (2014). IRBIS: a systematic search for conserved complementarity. *RNA*, **20**(10), 1519–1531.

Rehmsmeier, M., Steffen, P., Hochsmann, M., and Giegerich, R. (2004). Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**(10), 1507–1517.

Romero-López, C. and Berzal-Herranz, A. (2009). A long-range RNA-RNA interaction between the 5' and 3' ends of the HCV genome. *RNA*, **15**(9), 1740–1752.

Romero-López, C. and Berzal-Herranz, A. (2012). The functional RNA domain 5BSL3.2 within the NS5B coding sequence influences hepatitis C virus IRES-mediated translation. *Cell Mol Life Sci*, **69**(1), 103–113.

Romero-López, C., Barroso-Deljesus, A., García-Sacristán, A., Briones, C., and Berzal-Herranz, A. (2014). End-to-end crosstalk within the hepatitis C virus genome mediates the conformational switch of the 3'X-tail region. *Nucleic Acids Res*, **42**(1), 567–582.

Salari, R., Möhl, M., Will, S., Sahinalp, S. C., and Backofen, R. (2010). Time and Space Efficient RNA-RNA Interaction Prediction via Sparse Folding. In *RECOMB*, volume 6, pages 473–490. Springer.

Seemann, S. E., Richter, A. S., Gesell, T., Backofen, R., and Gorodkin, J. (2011). PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences. *Bioinformatics*, **27**(2), 211–219.

Tafer, H. and Hofacker, I. L. (2008). RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*, **24**(22), 2657–2663.

Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*, **10**(3), 512–526.

Tuplin, A., Struthers, M., Simmonds, P., and Evans, D. J. (2012). A twist in the tail: SHAPE mapping of long-range interactions and structural rearrangements of RNA elements involved in HCV replication. *Nucleic Acids Res*, **40**(14), 6908–6921.

Wu, B., Pogany, J., Na, H., Nicholson, B. L., Nagy, P. D., and White, K. A. (2009). A discontinuous RNA platform mediates RNA virus replication: building an integrated model for RNA-based regulation of viral processes. *PLoS Pathog*, **5**(3).

You, S., Falgout, B., Markoff, L., and Padmanabhan, R. (2001). In vitro RNA synthesis from exogenous dengue viral RNA templates requires long range interactions between 5'- and 3'-terminal regions that influence RNA structure. *J Biol Chem*, **276**(19), 15581–15591.

Zhang, B., Dong, H., Stein, D. A., and Shi, P. Y. (2008). Co-selection of West Nile virus nucleotides that confer resistance to an antisense oligomer while maintaining long-distance RNA/RNA base pairings. *Virology*, **382**(1), 98–106.