VeGETA: whole viral genome multiple sequence alignments based on RNA secondary structures

Kevin Lamkiewicz[1,2] Michèle Kayser[1], Emanuel Barth[1,3] and Manja Marz[1,2,3,4]

[1]Faculty of Mathematics and Computer Science, Friedrich Schiller University Jena, Germany
[2]European Virus Bioinformatics Center, Jena, Germany
[3]FLI Leibniz Institute for Age Research, Beutenbergstraße 11, Jena, Germany
[4]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

European VIRUS BIOINFORMATICS Center

Problem - Finding representative sequences from million of viruses to create RNA secondary structure alignments

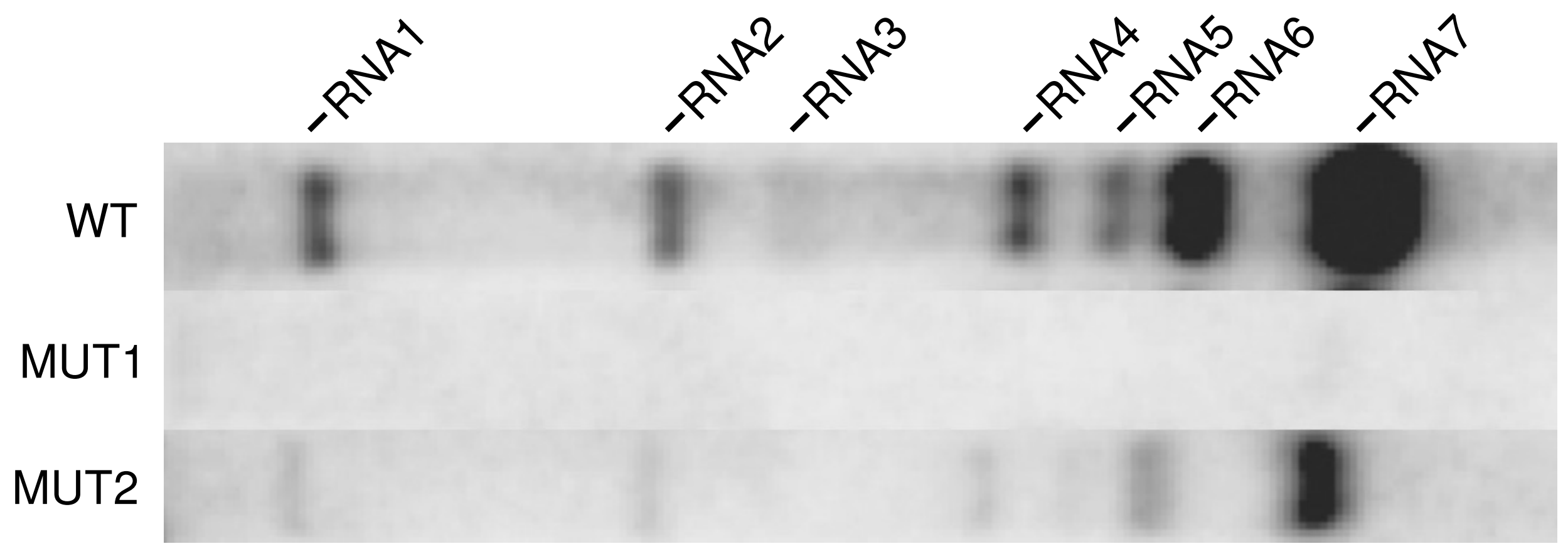
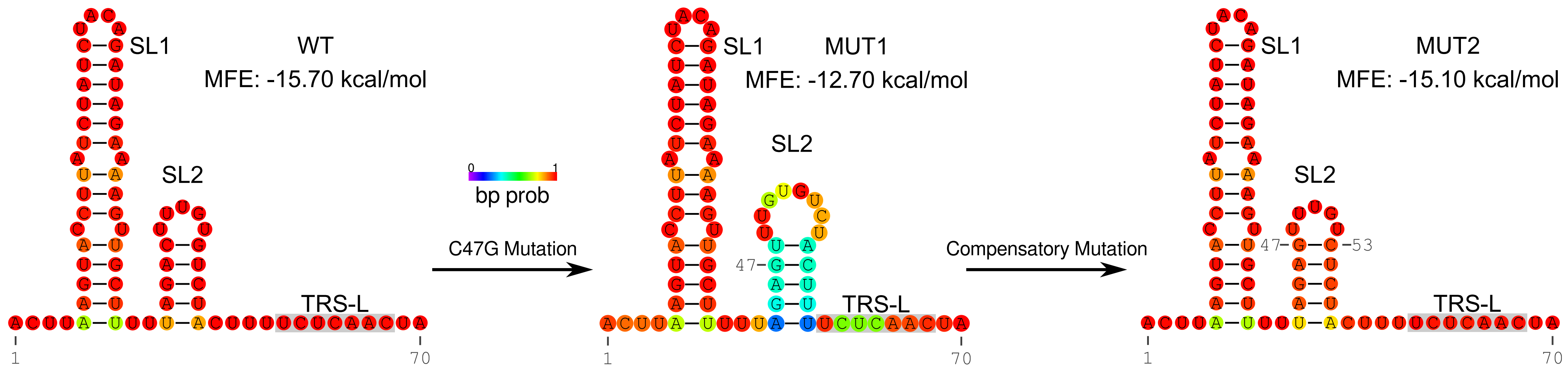


Figure 1: WT, a SNP in the SL2 structure (MUT1) and the compensatory SNP mutant (MUT2) of HCoV-229E 5'UTR. Overall viral RNA abundance highly correlates with changes of the structure stability. Adapted from [1].



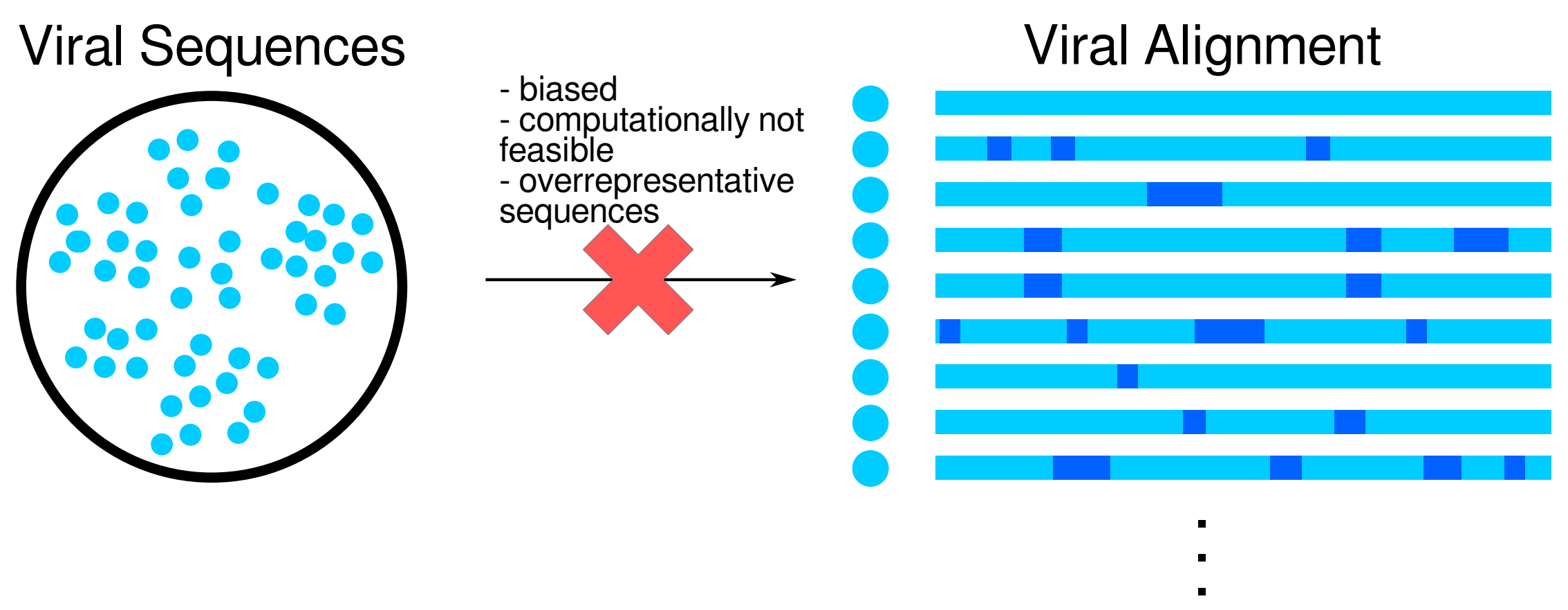
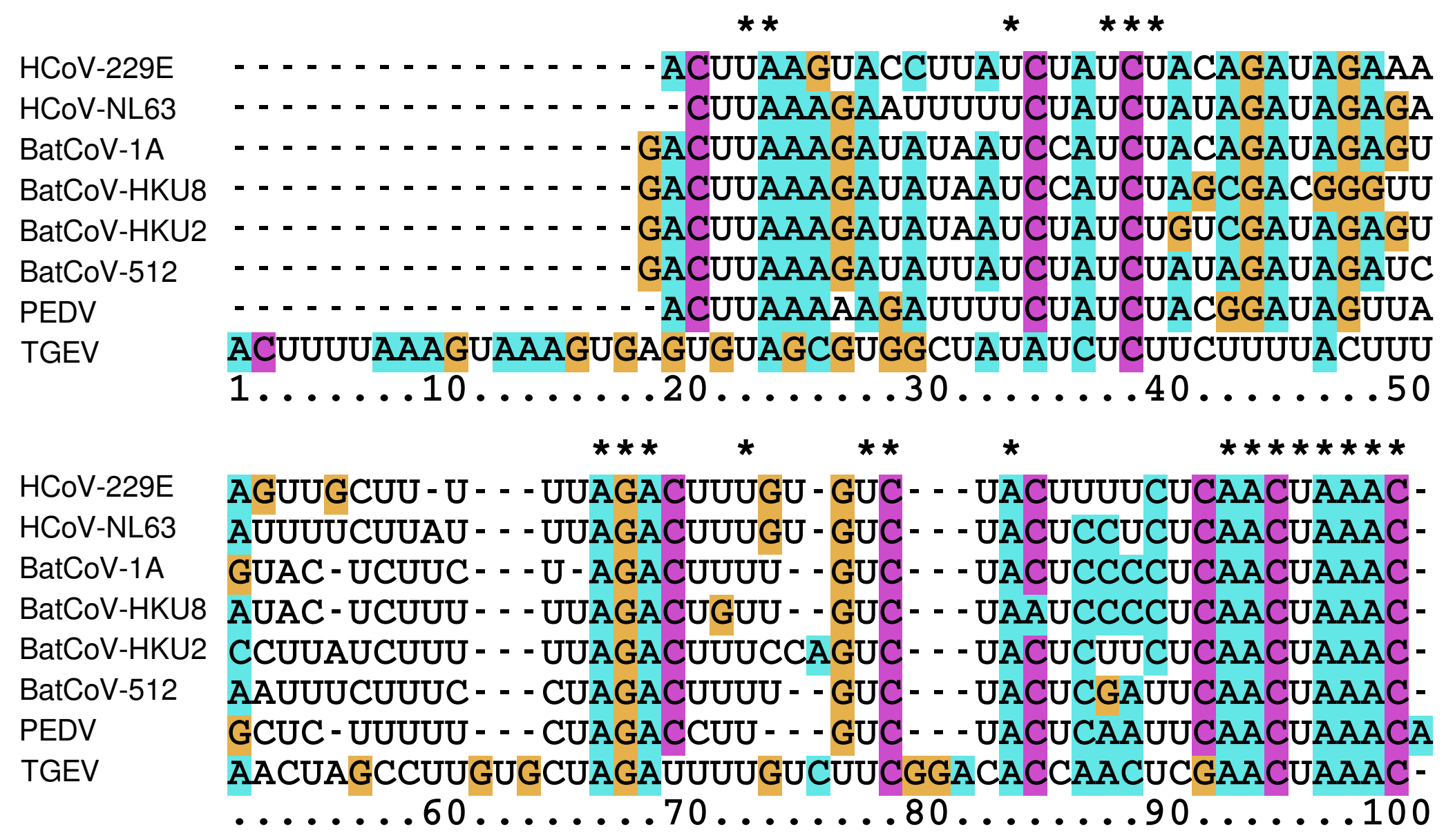

Viral Sequences
- biased
- computationally not feasible
- overrepresentative sequences
Viral Alignment

Figure 2: Calculating multiple sequence alignments on all available viruses (genus, family) is not feasible due to overrepresentative sequences, bias and limited computational resources.



```
                                      **           *    ***
HCoV-229E    ------------------ACUUAAGUACCUUAUCUAUCUACAGAUAGAAA
HCoV-NL63    -------------------CUUAAAGAAUUUUUCUAUCUAUAGAUAGAGA
BatCoV-1A    -----------------GACUUAAAGAUAUAAUCCAUCUACAGAUAGAGU
BatCoV-HKU8  -----------------GACUUAAAGAUAUAAUCCAUCUAGCGACGGGUU
BatCoV-HKU2  -----------------GACUUAAAGAUAUAAUCUAUCUGUCGAUAGAGU
BatCoV-512   -----------------GACUUAAAGAUAUUAUCUAUCUAUAGAUAGAUC
PEDV         ------------------ACUUAAAAAGAUUUUCUAUCUACGGAUAGUUA
TGEV         ACUUUUAAAGUAAAGUGAGUGUAGCGUGGCUAUAUCUCUUCUUUUACUUU
             1.......10........20........30........40........50

                              ***    *    **    *         ********
HCoV-229E    AGUUGCUU-U---UUAGACUUUGU-GUC---UACUUUUCUCAACUAAAC-
HCoV-NL63    AUUUUCUUAU---UUAGACUUUGU-GUC---UACUCCUCUCAACUAAAC-
BatCoV-1A    GUAC-UCUUC---U-AGACUUUU--GUC---UACUCCCCUCAACUAAAC-
BatCoV-HKU8  AUAC-UCUUU---UUAGACUGUU--GUC---UAAUCCCCUCAACUAAAC-
BatCoV-HKU2  CCUUAUCUUU---UUAGACUUUCCAGUC---UACUCUUCUCAACUAAAC-
BatCoV-512   AAUUUCUUUC---CUAGACUUUU--GUC---UACUCGAUUCAACUAAAC-
PEDV         GCUC-UUUUU---CUAGACCUU---GUC---UACUCAAUUCAACUAAACA
TGEV         AACUAGCCUUGUGCUAGAUUUUGUCUUCGGACACCAACUCGAACUAAAC-
             ........60........70........80........90.......100
```

Figure 3: Multiple sequence alignment of the 5' UTR of different *Alphacoronaviruses*. This region is known to be highly conserved on structure level but not on sequence level.

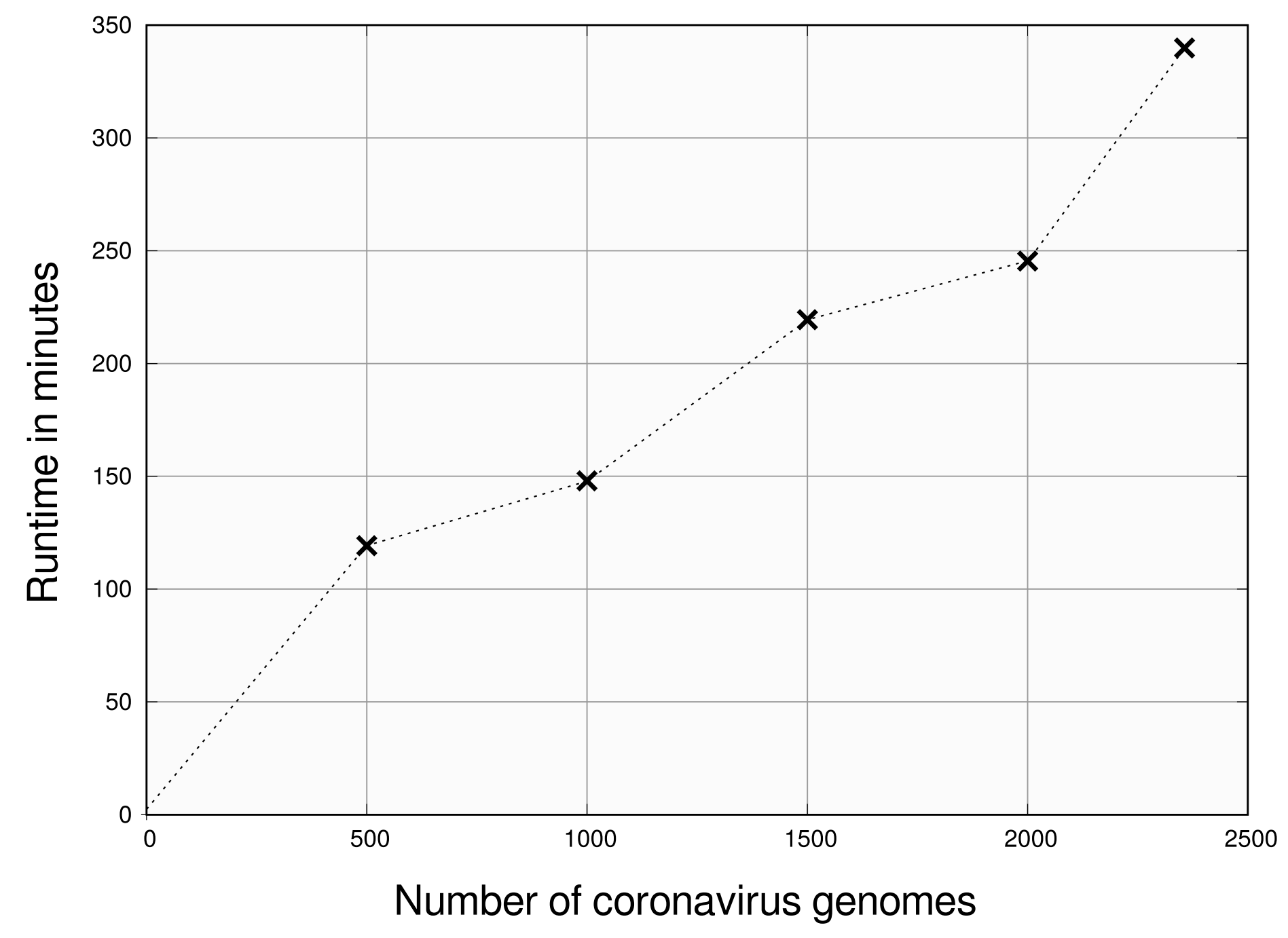
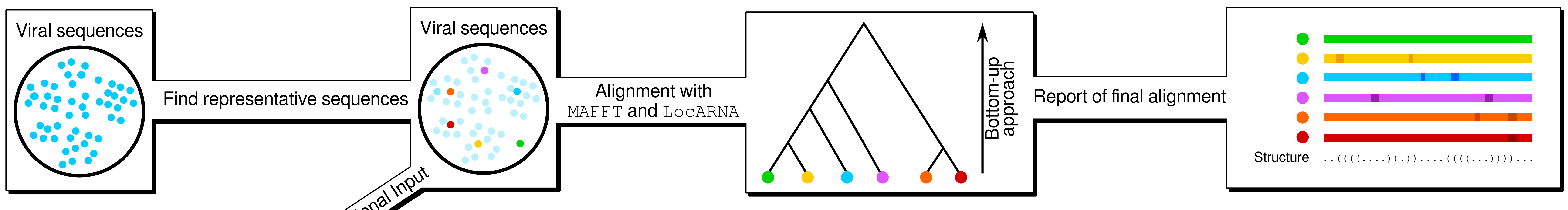
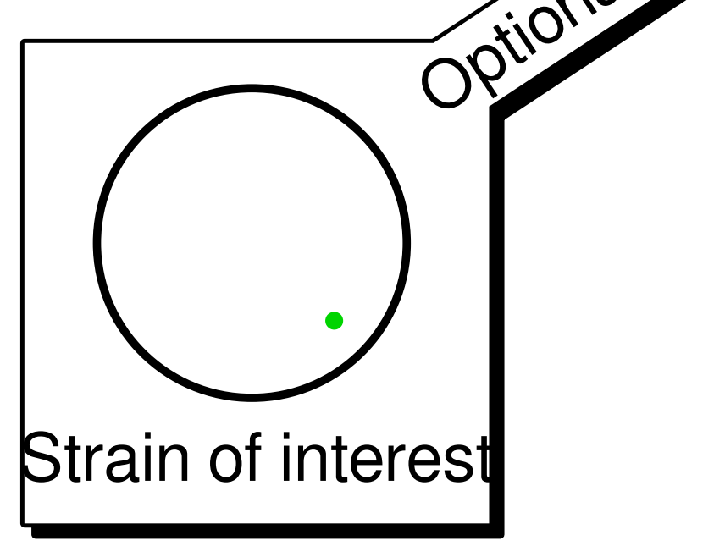


Figure 4: Runtime analysis for the calculation of multiple sequence alignments of coronaviruses using MAFFT [3].

Methods - Implementation and workflow of VeGETA

Viral sequences

Find representative sequences

Viral sequences

Alignment with MAFFT and LocARNA

Bottom-up approach

Report of final alignment

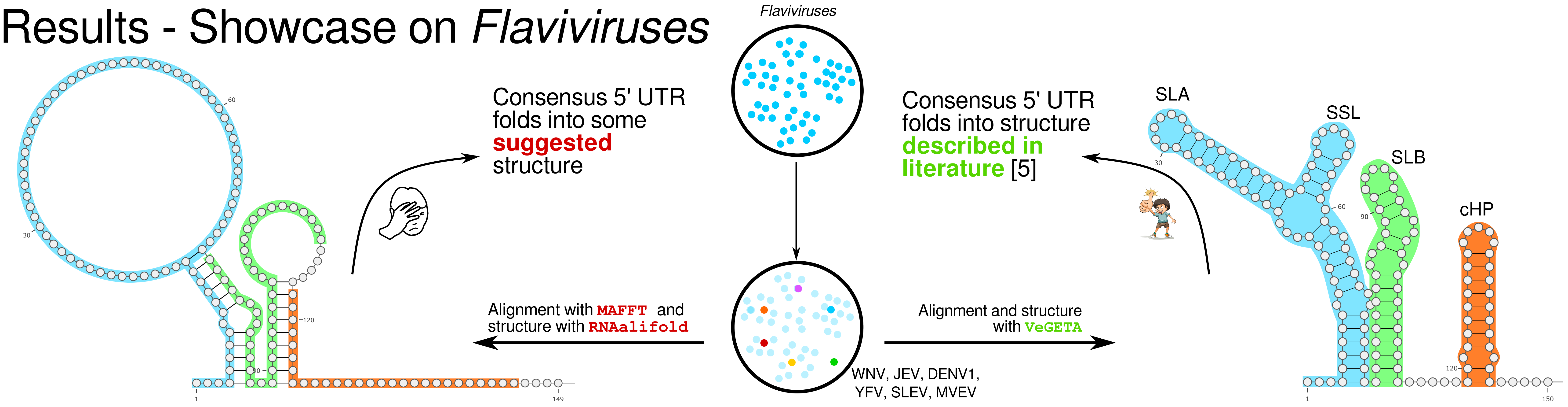
Structure `..((((....)).)).....((((...)))).......`

Optional Input

Strain of interest

Figure 5: Workflow of VeGETA. RNA secondary structure prediction benefits from alignment information, however, generating good alignments is a challenging task and usually does not capture structural information. With our pipeline we filter all input sequences for representative viruses and build a structure-based whole genome alignment. This is realized by combining MAFFT [3] and LocARNA [4] in a progressive way. All intermediate alignments and the final alignment are reported and provide an annotation of local RNA structures for further downstream analysis.

Results - Showcase on *Flaviviruses*



Consensus 5' UTR folds into some **suggested** structure

Alignment with **MAFFT** and structure with **RNAalifold**

WNV, JEV, DENV1, YFV, SLEV, MVEV

Alignment and structure with **VeGETA**

Consensus 5' UTR folds into structure **described in literature** [5]

Conclusion - VeGETA for good alignments

- RNA structures are more conserved than the genomic sequence
- Function of ncRNAs derived from structure
- Alignment-based analysis preferable, but computationally limiting
- Viral genome data exceeds these limits

⟶ Appropiate filters and selection needed

VeGETA

- **Filters sequences** for representative viruses
- Considers **sequence and structure** information for alignments
- Calculates **whole genome alignments** for viruses
- Allows inclusion of **own virus** of interest
- First results highly **agree with literature**

RNA BIOINFORMATICS & HIGH-THROUGHPUT ANALYSIS

References:
[1] Madhugiri R. et al. (2018) Structural and functional conservation of cis-acting RNA elements in coronavirus 5′-terminal genome regions, Virology
[2] Fricke M. et al. (2015) Conserved RNA secondary structures and long-range interactions in hepatitis C viruses, RNA
[3] Nakamura T. et al. (2018) Parallelization of MAFFT for large-scale multiple sequence alignments, Bioinformatics
[4] Sebastian W. et al. (2007) Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering, PLoS Comp. Bio.
[5] Fernández-Sanlés A. et al. (2017) Functional Information Stored in the Conserved Structural RNA Domains of Flavivirus Genomes. Front. Microbiol
[6] Fricke M., Marz M. (2016), Prediction of conserved long-range RNA-RNA interactions in full viral genomes, Bioinformatics

Thumbs-up Boy Vector created by brgfx - www.freepik.com

Contact
kevin.lamkiewicz@uni-jena.de
@k_lamkiewicz
www.rna.uni-jena.de

Federal Ministry of Education and Research