NAR Genomics and Bioinformatics, 2025, **7**, lqaf066
https://doi.org/10.1093/nargab/lqaf066
Application Notes

OXFORD

VariantFoldRNA: a flexible, containerized, and scalable pipeline for genome-wide riboSNitch prediction

Kobie J. Kirven [1,2,*], Philip C. Bevilacqua [2,3,4,*], Sarah M. Assmann [2,5,*]

[1]Graduate Program in Bioinformatics and Genomics, Pennsylvania State University, University Park, PA 16802, United States
[2]Center for RNA Molecular Biology, Pennsylvania State University, University Park, PA 16802, United States
[3]Department of Chemistry, Pennsylvania State University, University Park, PA 16802, United States
[4]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802, United States
[5]Department of Biology, Pennsylvania State University, University Park, PA 16802, United States

[*]To whom correspondence should be addressed. Email: kjk6173@psu.edu
Correspondence may also be addressed to Philip C. Bevilacqua. Email: pcb5@psu.edu
Correspondence may also be addressed to Sarah M. Assmann. Email: sma3@psu.edu

Abstract

Single nucleotide polymorphisms (SNPs) can alter RNA structure by changing the proportions of existing conformations or leading to new conformations in the structural ensemble. Such structure-changing SNPs, or riboSNitches, have been associated with diseases in humans and climate adaptation in plants. While several computational tools are available for predicting whether an SNP is a riboSNitch, these tools were generally developed to analyze individual RNAs and are not optimized for genome-wide analyses. To fill this gap, we developed VariantFoldRNA, a flexible, containerized, and automated pipeline for genome-wide prediction of riboSNitches. Our pipeline automatically installs all dependencies, can be run locally or on high-performance clusters, and is modular, enabling the user to customize the analysis for the research question of interest. VariantFoldRNA can predict riboSNitches genome-wide at user-specified temperatures and splicing conditions, opening the door to novel analyses. The pipeline is an open-source command-line tool that is freely available at https://github.com/The-Bevilacqua-Lab/variantfoldrna.

Introduction

Across all domains of life, RNAs are vital to cellular functions, including gene expression, condensate formation, translation, and catalysis. These diverse functions are possible in part because RNA can fold into complex structures through both intramolecular interactions, such as hydrogen bonding and base stacking, and intermolecular interactions with proteins, metal ions, other RNAs, and metabolites [1–3]. Rather than adopting single structures, RNAs, particularly long ones such as messenger RNAs (mRNAs), tend to use the above interactions to fold into multiple conformations, collectively known as the structural ensemble [4, 5]. The weight of a particular conformation in the ensemble is a function of the relative free energy of that conformation. Free energies of an RNA fold can be affected by single-nucleotide changes to the sequence, shifting proportions of particular conformations in the ensemble, or even leading to new conformations [6]. Such structure-altering single nucleotide polymorphisms (SNPs), known as riboSNitches, can alter RNA function and regulation. For instance, some riboSNitches are associated with genetic disease in humans [7–9], while for sessile organisms, such as plants, some riboSNitches may be climate-adaptive, altering the stability of a particular RNA conformation to favor functionality under local climate conditions [10]. Thus, riboSNitches represent a significant class of sequence variation that could underlie mechanisms of disease and adaptation across all domains of life.

It is currently challenging to experimentally identify riboSNitches from among the many SNPs detected across a population. This difficulty arises because determining whether each SNP alters RNA structure requires data on the structure of the respective RNA with the reference nucleotide and with the alternative nucleotide. While methods such as X-ray crystallography, cryogenic electron microscopy, and nuclear magnetic resonance can measure RNA structure with near-atomic-level resolution, their low throughput makes them unsuitable for large-scale analyses, typically involving millions of SNPs. By contrast, structure-probing methods such as parallel analysis of RNA structure [11], Structure-seq [12], SHAPE-MaP [13], and DMS-MaPseq [14] can produce measurements of RNA structure genome-wide. While the throughput of these wet-bench methods allows whole-genome assessment, identifying riboSNitches population-wide would require probing of hundreds to thousands of genomes, which is prohibitive. The above experimental limitations, combined with the large-scale nature of current SNP datasets comprised of thousands of genomes, such as the 3K Rice Genomes project [15] and gnomAD [16], underscore the need for a pipeline that provides efficient computational identification of candidate riboSNitches.

Received: December 21, 2024. **Revised:** March 5, 2025. **Editorial Decision:** May 12, 2025. **Accepted:** May 15, 2025

Existing computational tools for riboSNitch prediction, such as SNPfold [7], remuRNA [17], RNAsnp [18], and Riprap [19], predict whether an SNP affects RNA structure and have been valuable for identifying candidate riboSNitches for experimental validation. These methods use thermodynamic-based predictions of RNA secondary structure ensembles to estimate whether introducing an SNP would alter the ensemble of the RNA, although each method employs a different metric for scoring the structural change. Specifically, the scores for both SNPfold and remuRNA are calculated using local RNA sequence flanking the SNP since structure is typically influenced more by local than long-range interactions [20]; moreover, experimental data also support a model in which riboSNitches tend to alter local rather than global structure [21]. The scoring metric for SNPfold is the Pearson correlation coefficient between the base-pairing probabilities of the RNA with and without the alternative nucleotide, where a lower correlation coefficient reflects changes in pairing probabilities and, thus, changes to RNA structure [7]. On the other hand, remuRNA uses relative entropy as the scoring metric, where a higher relative entropy predicts a greater structural change introduced by the SNP [17]. While the scores from both SNPfold and remuRNA quantify the predicted structural effect of the alternative nucleotide on the entire user-defined local region of the RNA, RNAsnp and Riprap find and use the subregion within a user-defined local region where the base-pairing probabilities differ the most between the reference and alternative alleles. For RNAsnp, the default scoring metric is the Euclidean distance [18], while in Riprap the score is derived from a combination of the fold change between the reference and alternative pairing probabilities and the significance of the fold change based on the Kolmogorov–Smirnov test [19]. For both RNAsnp and Riprap, higher scores predict larger SNP-induced effects on RNA structure. Thus, one has to be mindful that changes to RNA structure can be indicated by lower (SNPfold) or higher (remuRNA, RNAsnp, and Riprap) scores. Regardless of the scoring metric, all of these tools require users to provide the local reference sequence of the RNA surrounding the SNP; this sequence is kept the same for the reference and alternative nucleotides. While manually providing these inputs is manageable for a few SNPs, it becomes overwhelming when the number of SNPs is large. At large scales, SNPs are often defined in variant call format (VCF), which details the positions of SNPs relative to the genome. However, VCF format is not currently incorporated in any of the above tools, and we are currently not aware of a pipeline that allows users to go directly from SNPs in a VCF file to predictions of the effects of those SNPs on RNA structure. Additionally, the computing time needed to run the thermodynamic calculations that underlie riboSNitch prediction can be prohibitively long at such large scales.

Here, we introduce VariantFoldRNA, a user-friendly pipeline for computationally predicting riboSNitches genome-wide across any species of interest, starting from a VCF file, or from user-designated sequences, starting with a CSV file. Our automated pipeline annotates SNPs, extracts the local flanking sequence, and predicts the impact of each SNP on RNA structure, all from a single command. The pipeline is also flexible, allowing users to choose from the aforementioned riboSNitch prediction tools (SNPfold, remuRNA, RNAsnp, and Riprap), to designate the temperature for RNA folding, which is particularly important for poikilothermic organisms, and to select whether the sequence surrounding the SNP is in the context of spliced or unspliced transcripts. We also added the ability to split or "chunk" the input files. Because of the underlying Snakemake [22] architecture and our addition of the ability to chunk the input files, VariantFoldRNA is parallelizable. This parallelization allows VariantFoldRNA to be scaled to as many cores as the user has available, facilitating riboSNitch prediction at the genome-wide scale. An overview of the pipeline is shown in Fig. 1. We demonstrate the utility of VariantFoldRNA by predicting riboSNitches genome-wide in thousands of *Arabidopsis thaliana* and *Oryza sativa* (rice) genomes over a range of biologically relevant temperatures.

Materials and methods

Automatic handling of dependencies

Table 1 shows versions of the software discussed herein. VariantFoldRNA manages all software dependencies internally using predefined Conda environments [23]. Our Conda environments contain specific versions of the software dependencies, ensuring reproducibility for the future even as new versions of these dependencies are released. The Conda environments are automatically created the first time the pipeline is run by the user, removing the burden to install the software dependencies manually. Users have the option to run the pipeline inside a Docker [24] container, which provides a consistent operating system regardless of the user's host system and thus further increases reproducibility. The Docker option requires only one additional command-line flag.

Pipeline overview

The following sections describe the components of the pipeline shown in Fig. 1. Steps 1, 2, 6, and 7 are common to the VCF and CSV modules, while steps 3–5 are unique to the VCF module. All command-line flags are defined in the VariantFoldRNA documentation (https://github.com/The-Bevilacqua-Lab/variantfoldrna/wiki).

1. Pipeline inputs (VCF and CSV modules)
   Users can input SNPs to the pipeline in either a VCF or CSV file. The VCF option is intended for naturally occurring SNPs, while the CSV option also allows the user to test sequences from other sources, such as synthetic sequences. When using SNPs from a VCF file, users must provide three inputs: SNPs of interest, gene annotations (in GFF format), and a reference genome (in FASTA format). For the CSV option, users need to include the reference sequence and the position and nucleotide identities of the reference and the alternative variants. Sample input files for running the pipeline in both VCF and CSV modes are provided in the GitHub repository (https://github.com/The-Bevilacqua-Lab/variantfoldrna/tree/main/example_input).
2. Split the VCF/CSV file into chunks (VCF and CSV modules)
   Computational prediction of riboSNitches from many SNPs is time-consuming since riboSNitch prediction tools rely on thermodynamic predictions of RNA secondary structure partition functions. Moreover, this issue with runtime is amplified since the partition function needs to be computed for both the reference and alternative sequences. To enable the later steps in the pipeline, including those in the riboSNitch module (steps 6 and 7), to be performed in parallel, this step

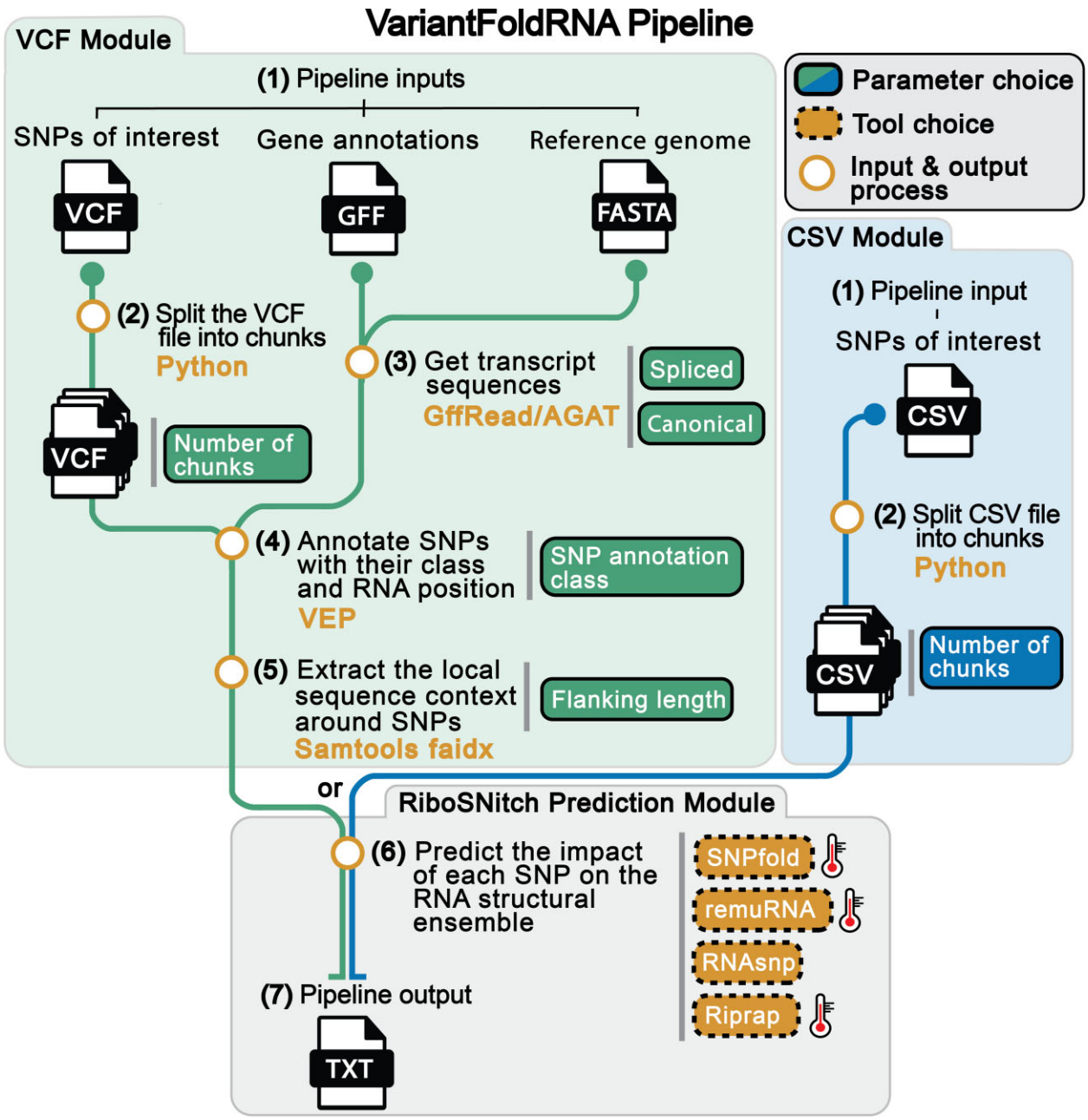

Figure 1. Overview of the VariantFoldRNA pipeline. The pipeline is comprised of three modules: the VCF (lighter green shading) and CSV (lighter blue shading) input modules and the riboSNitch prediction (lighter gray shading) output module. Corresponding boxes for VCF (darker green shading) and CSV (darker blue shading) indicate parameter choices, while dashed orange boxes indicate riboSNitch prediction method choices. For the VCF module, VariantFoldRNA first splits the input VCF file into chunks. Users can specify the number of chunks, or it will be automatically determined. Then, the sequences of RNAs that are defined in the GFF file are retrieved from the reference genome in FASTA format with GffRead [25]. SNPs are then annotated with class (e.g. synonymous) and position in the transcript using the Ensembl Variant Effect Predictor (VEP) [27]. Next, the sequence flanking the SNP is extracted using Samtools faidx [28]. For the CSV module, users provide the flanking sequences and the position and identities of the SNP in the RNA, so chunking of the CSV file is the only step performed before riboSNitch prediction. Finally, the SNPs and flanking sequences from both the VCF and CSV modules are passed to the riboSNitch prediction module, and the user-selected riboSNitch prediction tool(s) (SNPfold [7], remuRNA [17], RNAsnp [18], and/or Riprap [19]) is used to predict the impact of each SNP on the structural ensemble of the RNA. Three of the four tools allow riboSNitch prediction at a user-defined temperature(s), as indicated by the thermometer icon.

Table 1. Software versions of the primary tools used in VariantFoldRNA. Relevant references are indicated

Software	Version	Reference
AGAT	1.4.1	[26]
GffRead	0.12.1	[25]
remuRNA	1.0	[17]
Riprap	Our Golang implementation	[19]
RNAsnp	1.2	[18]
Samtools	1.16.1	[28]
Snakemake	7.24.2	[22]
Snk	0.22.2	[30]
SNPfold	Our Golang implementation	[7]
VEP	112.0	[27]

splits or "chunks" the input VCF file using a custom Python script. Users can specify the number of chunks or allow the number to be determined automatically based on the size of the input file. For large files, it is often helpful to use more chunks than available cores as this prevents a greater number of predictions from needing to be repeated in the case of a failed run. The Snakemake [22] architecture underlying VariantFoldRNA enables parallelization on desktop computers with multiple cores and on high-performance computing (HPC) systems with job management systems like Slurm. Parallelization from both chunking and Snakemake itself can significantly reduce the runtime of the pipeline, especially as the size of the SNP dataset increases.

3. Get transcript sequences (VCF module)
   With VariantFoldRNA, users can set the splicing flag on or off to predict riboSNitches in either spliced or unspliced transcripts as extracted from the GFF file with GffRead [25]. Since a gene may have many annotated isoforms in the GFF, users can focus their analysis on the single canonical transcript per gene identified by Another GTF/GFF Analysis Toolkit (AGAT) [26] by activating the canonical flag. [Canonical transcripts are defined for a given gene as the transcript with the longest coding sequence (CDS) or, if the gene has no CDS, the longest transcribed RNA [26].] As compared to using all isoforms annotated in the provided GFF file, using the single canonical transcript decreases the complexity of the analysis and reduces the runtime.
4. Annotate SNPs with their RNA position and class (VCF module)
   Any class of SNPs in RNA has the potential to be riboSNitches. For instance, in mRNAs such classes include missense and synonymous SNPs, and those in noncoding regions (5′UTR (untranslated region), 3′UTR, and intronic SNPs). VariantFoldRNA allows the user to focus on any one of these classes as annotated by the Ensembl VEP (e.g. only SNPs in 5′UTRs or only synonymous SNPs [27]) by specifying the "variant annotation type" flag. Users can also choose to focus their analysis on SNPs in a particular chromosome(s), by specifying the "chromosome" flag. VEP also annotates the nucleotide position in the RNA where the SNP occurs, which is needed for extracting the flanking sequence.
5. Extract the local sequence context around SNPs (VCF module)
   To extract the local RNA sequence surrounding an SNP with a user-defined flanking length, VariantFoldRNA uses faidx from Samtools [28] within a custom Python script. Here, flanking length is defined as the length of one of the two symmetric flanks. The default flanking length is 50 nt on either side of the SNP (total length of 101 nt), as in previous work by Corley and colleagues [29]. It is important to consider that for each prediction tool, especially remuRNA, a choice of increasing flanking length will exponentially increase runtime. In the VCF module, SNPs are excluded from analysis if their distance from the 5′ or 3′ end of the RNA is less than the chosen flanking length. However, these SNPs can be handled by SNPfold, remuRNA, and Riprap in the CSV module because the user provides the exact flanking sequence.
6. Predict the impact of each SNP on the RNA structural ensemble (riboSNitch prediction module)
   Next, the SNPs and flanking sequences from either the VCF or the CSV module are passed to the riboSNitch prediction module. As mentioned, VariantFoldRNA incorporates four tools for riboSNitch prediction: SNPfold [7], remuRNA [17], RNAsnp [18], and Riprap [19]. Users can select one or more of these tools when running the pipeline. Since SNPfold and Riprap were only available as Python implementations, they were re-implemented in the compiled language Golang to increase computational efficiency and thus decrease runtime. Additionally, VariantFoldRNA can perform riboSNitch prediction at any temperature or range of temperatures for SNPfold, remuRNA, and Riprap. (This is not possible with RNAsnp since the distributions underlying the *P*-value calculations are based on predictions at 37°C [18].)
7. Pipeline output (riboSNitch prediction module)
   The pipeline output is a tab-separated text file with a row for each SNP. Example output files for both VCF and CSV modules can be found in the GitHub repository (https://github.com/The-Bevilacqua-Lab/variantfoldrna/tree/main/example_output). When multiple riboSNitch prediction tools are selected, the results from each tool are combined into a single output file. Since Riprap and RNAsnp find the subregion with the greatest structural change, for those two tools VariantFoldRNA also outputs the position of that region for each SNP. After the pipeline has finished running, intermediate files from the pipeline run are retained. Since Snakemake can recognize these intermediate files, their retention can prevent steps in the pipeline from being unnecessarily repeated when the pipeline is run with new command-line flags or when re-running after a failed run.

Example command-line usage

VariantFoldRNA is made accessible to those unfamiliar with Snakemake using the Snk workflow management system [30]. With Snk, parameters that would be included in the Snakemake config file are instead turned into command-line arguments, removing the need to manually edit a config file. The command below demonstrates this simplicity, as it is all that is needed to predict riboSNitches from canonical spliced transcripts in a genome-wide dataset of SNPs from the 3024 distinct inbred cultivars of rice (*O. sativa*) that comprise the rice 3K dataset [15].

```
variantfoldrna run -vcf rice_3k_snps.vcf \
-gff Oryza_sativa.IRGSP-1.0.60.gff3 \
-ref-genome Oryza_sativa.IRGSP-1.0.dna.toplevel.fa \
-spliced -canonical
```

Results

Development and optimization of the VariantFoldRNA pipeline

To improve the throughput of riboSNitch identification in genome-wide datasets, we developed a flexible, containerized, and automated pipeline for riboSNitch prediction that we call VariantFoldRNA. As shown in Fig. 1, this pipeline is comprised of three modules. Two are input modules, which handle the VCF and CSV file formats, and one is an output module. The output module is flexible in allowing analysis by four riboSNitch prediction programs, three of which (all except RNAsnp) can perform this analysis as a function of temperature.

In addition to ease of use, a major advance in our pipeline is its parallelization, made possible by Snakemake and chunking. Parallelization markedly reduces the runtime, especially as the size of the SNP dataset increases. For instance, as shown in Supplementary Fig. S1, going from one to five cores with just 1000 SNPs decreased the runtime by 1 min (from 2 to 1 min), but the same change in core number with 50 000 SNPs decreased the runtime by nearly an hour (from 70 to 20 min, as quantified in Supplementary Fig. S1). We note that, for each prediction tool, especially remuRNA, increasing flanking length exponentially increases runtime (Supplementary Fig. S2), further emphasizing the importance of parallelization. More details on pipeline performance can be found in the "Runtime analyses" section of the Supplementary data, where we describe the effects of parallelization (Supplementary Fig. S1 and Supplementary Table S1) and flanking length (Supplementary Fig. S2 and Supplementary Table S2) on runtime, and report the average runtime per SNP for each of the tools included in VariantFoldRNA (Supplementary Table S3).

Effects of temperature on predicted riboSNitches

VariantFoldRNA can be used to uncover genome-wide trends in predicted riboSNitches. To illustrate its utility, we tested the impacts of temperature on the distributions of predicted riboSNitches in two plant species. To this end, we obtained SNP datasets from *A. thaliana* (*Arabidopsis*) and *O. sativa* (rice) populations. *Arabidopsis* is the premier model plant species, and rice is a globally important food crop. For *Arabidopsis*, we used the SNPs detected in 1135 distinct accessions that were part of the 1001 Genomes Project [31], while for rice, we used the SNPs identified from 941 landraces that were sequenced as part of the 3K Rice Genomes Project [15]. We applied VariantFoldRNA to these SNP datasets and obtained riboSNitch predictions on ∼2.85M SNPs for *Arabidopsis* and ∼1.30M SNPs for rice. When running VariantFoldRNA, we varied the folding temperatures in 5°C increments within biologically relevant temperature ranges for each species. The temperature range used for *Arabidopsis* was −35°C to 40°C, while the range for rice was −15°C to 40°C. (see Supplementary data and Supplementary Fig. S3 for details on the choice of temperature ranges). For these analyses, the SNPfold and Riprap riboSNitch prediction tools were chosen as illustrative of the two approaches used to quantify predicted structural effects: SNPfold (and remuRNA) calculates a score for the entire input region, while Riprap [and RNAsnp (at 37°C only)] calculates the score for the subregion that the algorithm finds to have the most significant predicted structural effect.

For both *Arabidopsis* and rice populations, we found that as folding temperature decreased, the distributions of SNPfold scores tended toward lower values (i.e. greater riboSNitch tendency). Most notably, with decreasing temperature, the lower whiskers of the boxplots of SNPfold scores shifted downward, while the medians and the top whiskers remained relatively similar (Fig. 2A and B). Therefore, the boxplots indicate that more SNPs are predicted to be riboSNitches at low temperatures for both *Arabidopsis* and rice. Likewise, we found that the upper whiskers of the Riprap score boxplots increased (i.e. greater riboSNitch tendency) with decreasing temperatures in both plant species (Fig. 2C and D), in accordance with the SNPfold predictions. Studies from Mathews and coworkers have tested the impact on RNA structure prediction of temperature extrapolation from 37°C, and their results suggest that predictions hold between 10°C and 60°C [32]. We note that if we limit our data to this temperature range, the conclusions remain unchanged (Fig. 2).

Discussion

Herein, we developed VariantFoldRNA, a computational pipeline to facilitate riboSNitch prediction at scale, which incorporates the tools SNPfold [7], remuRNA [17], RNAsnp [18], and Riprap [19]. We are aware of two other pipelines, SNIPPER and RNA-stability, that perform riboSNitch prediction, but these were designed for very specific applications in human genomics [33, 34]. Neither pipeline accepts VCF or CSV files as input nor incorporates any of the four tools that are incorporated in VariantFoldRNA. VariantFoldRNA is a user-friendly command-line tool that simplifies the complex process of generating riboSNitch predictions for millions of SNPs of interest to just a single command. All software dependencies are managed internally through Conda environments, which are automatically created the first time the pipeline is run. Additionally, the underlying Snakemake architecture, along with the option for "chunking" the input VCF/CSV files, allows VariantFoldRNA to be parallelized on resources ranging from a desktop computer with multiple cores to a large HPC system. Parallelization can significantly reduce the runtime. For example, on a computer with 32 GB of RAM, running the full pipeline with 50 000 input SNPs and SNPfold as the riboSNitch prediction tool takes over an hour on one core while running it with five cores takes only ∼20 min (Supplementary Fig. S1). This combination of usability, automatic dependency handling, and scalability poises VariantFoldRNA for large-scale riboSNitch prediction.

Users can customize VariantFoldRNA for a specific research question due to its built-in flexibility. The VCF option is designed for identification of riboSNitches genome-wide and across many genotypes, making it optimal for assessing the properties of naturally occurring SNPs in populations, wherein there are typically many millions of SNPs. The CSV option is ideal for evaluation of designed RNAs or sequences from species lacking whole-genome sequencing data. In addition, because the CSV option does not require symmetric flanking sequences, it can be applied to assess riboSNitch propensities of SNPs in the context of asymmetric flanks, including SNPs occurring near the ends of mRNAs or present

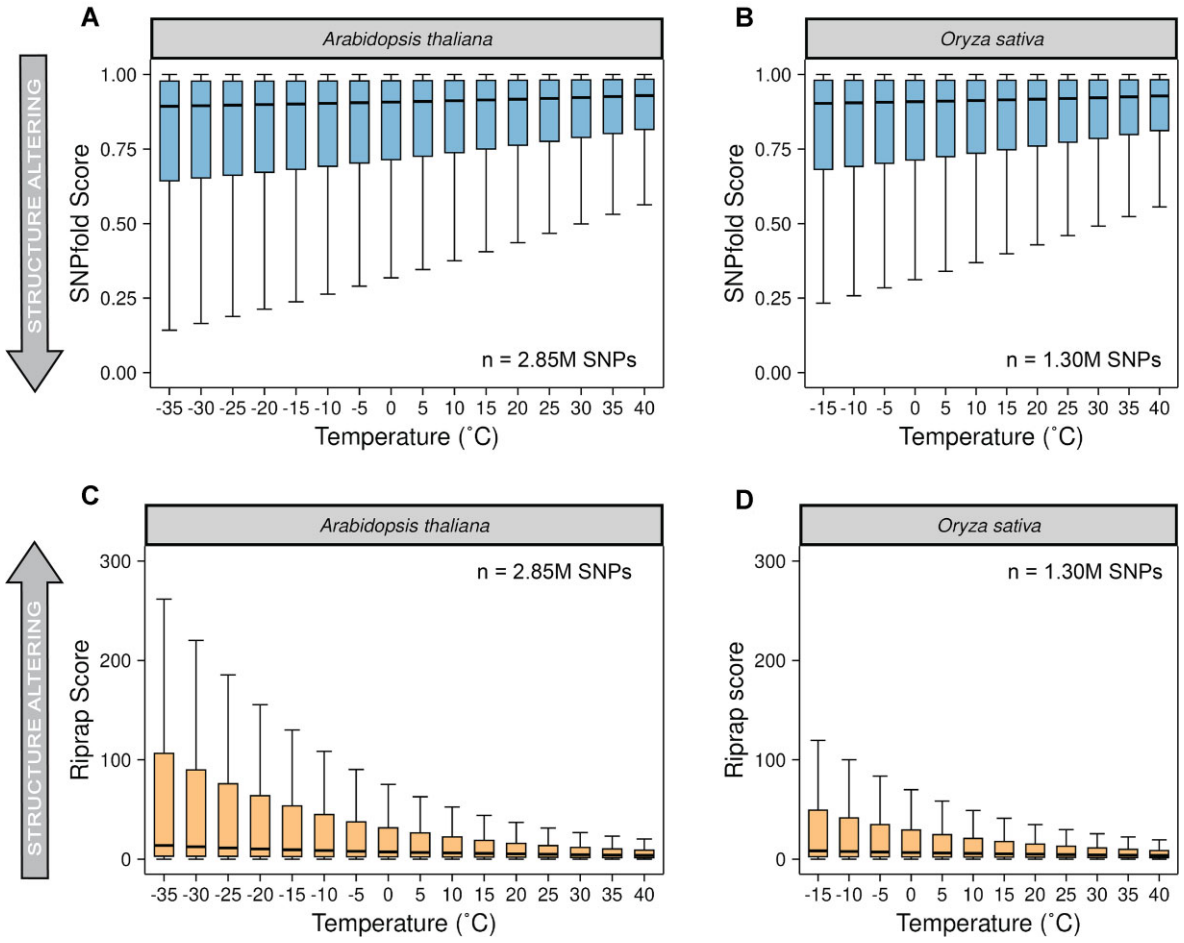

Figure 2. Increasing temperature decreases the overall extent of SNP-induced structural changes for both *A. thaliana* and *O. sativa*. Box and whisker plots of SNPfold scores across all SNPs calculated at different temperatures for (**A**) *A. thaliana* and (**B**) *O. sativa*. (**C**, **D**) Same as for panels (A) and (B) but for Riprap scores. SNPfold [7] and Riprap [19] were chosen as representative of the four riboSNitch prediction tools included in VariantFoldRNA.

in short transcripts such as transfer RNAs, microRNAs, and other functional small RNAs.

Moreover, users can choose one or more of four riboSNitch prediction tools incorporated in the pipeline: SNPfold [7], remuRNA [17], RNAsnp [18], and Riprap [19]. In one analysis [9], SNPs predicted to be riboSNitches by several tools were found to have increased reliability of prediction, as benchmarked by comparison with experimental data. With VariantFoldRNA, users can additionally perform riboSNitch predictions in the context of either spliced or unspliced transcripts. Because splicing can change the sequences flanking an SNP, this could uncover "splicing-conditional" riboSNitches. Another tunable parameter of VariantFoldRNA is temperature. Most work on riboSNitches has focused on humans, who tightly regulate internal temperature within a few degrees; however, poikilothermic organisms like plants can harbor SNPs that are riboSNitches only at certain temperatures (another type of "conditional riboSNitch") [10]. Thus, the ability of VariantFoldRNA to predict riboSNitches at multiple temperatures could aid in predicting temperature-conditional riboSNitches genome-wide in plants, fungi, and non-homeothermic animal species of Eukarya, as well as in Bacteria and Archaea.

Overall, our pipeline considerably reduces computational barriers for genome-wide riboSNitch prediction and thus facilitates discovery of genome-wide trends in predicted riboSNitches that are not apparent when studying riboSNitches individually. To illustrate the beneficial aspects of VariantFoldRNA, we predicted riboSNitches in *Arabidopsis* and rice over a range of biologically relevant temperatures. For both species, SNPs were predicted to be more likely to alter RNA structure at lower temperatures. This trend may occur because unfolded conformations, which are similar between the reference and alternative sequences, populate the ensembles more at higher temperatures, especially for weak structures [5]. Future work could test for differing selective pressures on riboSNitches between genotypes of the same species that experience different temperatures or occupy different climate niches.

In summary, VariantFoldRNA is a resource for those interested in predicting riboSNitches genome-wide. Based on the output predictions from VariantFoldRNA, specific putative riboSNitches can be identified for experimental validation. Additionally, VariantFoldRNA provides a platform to increase the usability and scalability of future riboSNitch prediction tools.

Acknowledgements

We would like to thank Megan Sylvia, Gabriela Hohenwarter, Reuben Kern, and Drs. Ángel Ferrero-Serrano and Elizabeth Jolley for beta testing and helpful feedback on the pipeline.

Author contributions: Software development: K.J.K.; analysis: K.J.K. with contributions from all authors; writing: all authors; funding acquisition: P.C.B. and S.M.A.

Supplementary data

Supplementary data is available at NAR Genomics & Bioinformatics online.

Conflict of interest

None declared.

Funding

This work was supported by National Science Foundation PGRP grant IOS-2122357 to P.C.B. and S.M.A., with additional support from National Institutes of Health grant R35 GM127064 to P.C.B. K.J.K. gratefully acknowledges funding by the Penn State Computation, Bioinformatics, and Statistics (CBIOS) training program (5T32GM102057).

Data availability

The data generated as a part of this manuscript are available on FigShare (https://figshare.com/articles/dataset/VariantFoldRNA_Analysis/28050719). VariantFoldRNA is freely available on GitHub (https://github.com/The-Bevilacqua-Lab/VariantFoldRNA), PyPI (https://pypi.org/project/variantfoldrna/), and FigShare (https://figshare.com/articles/software/VariantFoldRNA/28057367).

References

1. Brion P, Westhof E. Hierarchy and dynamics of RNA folding. *Annu Rev Biophys Biomol Struct* 1997;**26**:113–37. https://doi.org/10.1146/annurev.biophys.26.1.113
2. Vicens Q, Kieft JS. Thoughts on how to think (and talk) about RNA structure. *Proc Natl Acad Sci USA* 2022;**119**:e2112677119. https://doi.org/10.1073/pnas.2112677119
3. Assmann SM, Chou H-L, Bevilacqua PC. Rock, scissors, paper: how RNA structure informs function. *Plant Cell* 2023;**35** 1671–701 https://doi.org/10.1093/plcell/koad026
4. Martin J. Describing the structural diversity within an RNA's ensemble. *Entropy* 2014;**16**:1331–48. https://doi.org/10.3390/e16031331
5. Bonilla SL, Jones AN, Incarnato D. Structural and biophysical dissection of RNA conformational ensembles. *Curr Opin Struct Biol* 2024;**88**:102908. https://doi.org/10.1016/j.sbi.2024.102908
6. Solem AC, Halvorsen M, Ramos SBV *et al.* The potential of the riboSNitch in personalized medicine. *WIREs RNA* 2015;**6**:517–32. https://doi.org/10.1002/wrna.1291
7. Halvorsen M, Martin JS, Broadaway S *et al.* Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet* 2010;**6**:e1001074. https://doi.org/10.1371/journal.pgen.1001074
8. Linnstaedt SD, Riker KD, Rueckeis CA *et al.* A functional riboSNitch in the 3' untranslated region of FKBP5 alters microRNA-320a binding efficiency and mediates vulnerability to chronic post-traumatic pain. *J Neurosci* 2018;**38**:8407–20. https://doi.org/10.1523/JNEUROSCI.3458-17.2018
9. R G, Mitra A, Pk V. Predicting functional riboSNitches in the context of alternative splicing. *Gene* 2022;**837**:146694. https://doi.org/10.1016/j.gene.2022.146694
10. Ferrero-Serrano Á, Sylvia MM, Forstmeier PC *et al.* Experimental demonstration and pan-structurome prediction of climate-associated riboSNitches in *Arabidopsis*. *Genome Biol* 2022;**23**:101. https://doi.org/10.1186/s13059-022-02656-4
11. Kertesz M, Wan Y, Mazor E *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* 2010;**467**:103–7. https://doi.org/10.1038/nature09322
12. Ding Y, Tang Y, Kwok CK *et al.* *In vivo* genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* 2014;**505**:696–700. https://doi.org/10.1038/nature12756
13. Smola MJ, Rice GM, Busan S *et al.* Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat Protoc* 2015;**10**:1643–69. https://doi.org/10.1038/nprot.2015.103
14. Zubradt M, Gupta P, Persad S *et al.* DMS-MaPseq for genome-wide or targeted RNA structure probing *in vivo*. *Nat Methods* 2017;**14**:75–82. https://doi.org/10.1038/nmeth.4057
15. Wang W, Mauleon R, Hu Z *et al.* Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 2018;**557**:43–9. https://doi.org/10.1038/s41586-018-0063-9
16. Chen S, Francioli LC, Goodrich JK *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* 2024;**625**:92–100. https://doi.org/10.1038/s41586-023-06045-0
17. Salari R, Kimchi-Sarfaty C, Gottesman MM *et al.* Sensitive measurement of single-nucleotide polymorphism-induced changes of RNA conformation: application to disease studies. *Nucleic Acids Res* 2013;**41**:44–53. https://doi.org/10.1093/nar/gks1009
18. Sabarinathan R, Tafer H, Seemann SE *et al.* RNAsnp: efficient detection of local RNA secondary structure changes induced by SNPs. *Hum Mutat* 2013;**34**:546–56. https://doi.org/10.1002/humu.22273
19. Lin J, Chen Y, Zhang Y *et al.* Identification and analysis of RNA structural disruptions induced by single nucleotide variants using Riprap and RiboSNitchDB. *NAR Genom Bioinform* 2020;**2**:lqaa057. https://doi.org/10.1093/nargab/lqaa057
20. Lange SJ, Maticzka D, Möhl M *et al.* Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res.* 2012;**40**:5215–26. https://doi.org/10.1093/nar/gks181
21. Wan Y, Qu K, Zhang QC *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* 2014;**505**:706–9. https://doi.org/10.1038/nature12946
22. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 2012;**28**:2520–2. https://doi.org/10.1093/bioinformatics/bts480
23. Grüning B, Dale R, Sjödin A *et al.* Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods* 2018;**15**:475–6. https://doi.org/10.1038/s41592-018-0046-7
24. Merkel D. Docker: lightweight Linux containers for consistent development and deployment. *Linux J* 2014;**2014**:2.
25. Pertea G, Pertea M. GFF utilities: gffRead and GffCompare. *F1000Research* 2020;**9**: ISCBComm J-304. https://doi.org/10.12688/f1000research.23297.1
26. Dainat J, Hereñú D, Murray KD *et al.* NBISweden/AGAT: AGAT-v1.4.1. Zenodo, 2024. https://zenodo.org/records/13799920
27. McLaren W, Gil L, Hunt SE *et al.* The Ensembl variant effect predictor. *Genome Biol* 2016;**17**:122. https://doi.org/10.1186/s13059-016-0974-4
28. Li H, Handsaker B, Wysoker A *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9. https://doi.org/10.1093/bioinformatics/btp352
29. Corley M, Solem A, Qu K *et al.* Detecting riboSNitches with RNA folding algorithms: a genome-wide benchmark. *Nucleic Acids Res* 2015;**43**:1859–68. https://doi.org/10.1093/nar/gkv010
30. Wirth W, Mutch S, Turnbull R. Snk: a Snakemake CLI and workflow management system. *JOSS* 2024;**9**:7410. https://doi.org/10.21105/joss.07410

31. Weigel D, Mott R. The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol* 2009;**10**:107. https://doi.org/10.1186/gb-2009-10-5-107
32. Lu ZJ, Turner DH, Mathews DH. A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. *Nucleic Acids Res* 2006;**34**:4912–24. https://doi.org/10.1093/nar/gkl472
33. He F, Wei R, Zhou Z *et al*. Integrative analysis of somatic mutations in non-coding regions altering RNA secondary structures in cancer genomes. *Sci Rep* 2019;**9**:8205. https://doi.org/10.1038/s41598-019-44489-5
34. Gaither JBS, Lammi GE, Li JL *et al*. Synonymous variants that disrupt messenger RNA structure are significantly constrained in the human population. *GigaScience* 2021;**10**: giab023. https://doi.org/10.1093/gigascience/giab023

Received: December 21, 2024. **Revised:** March 5, 2025. **Editorial Decision:** May 12, 2025. **Accepted:** May 15, 2025