Reviewer 1 report:

**Are the methods appropriate to the aims of the study, are they well described, and are necessary controls included?**

Yes

**Are the conclusions adequately supported by the data shown?**

Yes

**Are sufficient details provided to allow replication and comparison with related analyses that may have been performed?**

Yes

**Does the method perform better than existing methods (as demonstrated by direct comparison with available methods)?**

No: It underperforms some methods, but is less computationally intensive and may be faster, but this is not demonstrated directly.

**Is the method likely to be of broad utility? Is any software component easy to install and use?**

**Please indicate briefly the novel features and/or advantages of the method, and/or please reference the relevant publications and which methods, if any, it should be compared with.**

Yes, the method itself will be of broad utility. The actual installation etc don't look too bad, but are probably beyond the average experimental user and most likely to be used by other computational biology experts.

**Is the paper of broad interest to others in the field, or of outstanding interest to a broad audience of biologists?**

Yes: RNA structural probing techniques are only expanding and there is a significant need for methods to systematically analyze the data.

**Comments:**

In the manuscript entitled "patteRNA: Transcriptome-wide search for functional RNA elements via structural data signatures" the authors present a novel computational method for the analysis of RNA structural probing data that may be applied on a transcriptome-wide level. Overall this is a nice addition to an area that could use some additional methods for systematic treatment of the data rather than a combination of ad hoc methodologies. The paper is well-written and the examples utilized are well-chosen and convincing.

However, I have some concerns. The method itself is not more accurate than some of the methods it was tested against (which is too bad). As the authors point out, many of these methods are computationally intensive for large data sets, thus their method still may play an important role in the interpretation of transcriptome-wide data. However, this does decrease the significance of the work, especially since the authors do not provide any specific benchmarks. It is unclear how much faster patteRNA might be, and under what circumstances this might make a significant difference, sufficient to warrant the loss in accuracy.

1. Since the authors do make such a big deal about the speed/scalability of patteRNA as an advantage over ensemble prediction methods, it would be nice to see a better quantitative comparison in terms of speed. How much does this actually matter given differing amounts of data/length of RNAs?

2. What is the audience that the authors are trying to reach with this work? Other computational biology experts, experimentalists? I feel like experimentalists will be the ones to benefit the most from it, but the paper itself feels like it is written for other experts and while the product is accessible to experts it might not be to the experimental audience.

Reviewer 2 report:

**Are the methods appropriate to the aims of the study, are they well described, and are necessary controls included?**

Yes

**Are the conclusions adequately supported by the data shown?**

Yes

**Are sufficient details provided to allow replication and comparison with related analyses that may have been performed?**

Yes

**Does the method perform better than existing methods (as demonstrated by direct comparison with available methods)?**

Yes

**Is the method likely to be of broad utility? Is any software component easy to install and use?**

**Please indicate briefly the novel features and/or advantages of the method, and/or please reference the relevant publications and which methods, if any, it should be compared with.**

The method is novel by the fact that it mines RNA structure motifs from profiling data. I am not aware of any other method that is comparable.

**Is the paper of broad interest to others in the field, or of outstanding interest to a broad audience of biologists?**

Yes: High-throughput structure profiling experiments are becoming more popular, less expensive and hence more frequently done. Therefore the method suggested would be of outstanding interest to a broad audience of biologists.

**Comments**

**In the manuscript by Ledda and Aviran, an unsupervised pattern recognition method that mines RNA structure motifs from profiling data is put forth. The method is called patteRNA. It is a machine learning algorithm that rapidly detects RNA structural motifs in large-scale structure profiling data. It can effectively detect motifs in various data sets and can be used to narrow down a set of candidate regions which can then undergo more careful nearest neighbor thermodynamic model analyses. Using three data sets, it is demonstrated that patteRNA detects motifs with accuracy comparable to commonly used thermodynamic models. The method is compatible with diverse profiling techniques and experimental conditions.**

**In general, the method is original and promising. It fills a needed gap in mining RNA structure motifs from genomic data and as more high-throughput structure profiling experiments will be performed in the future, which is expected, the usefulness of patteRNA will increase. The manuscript is also well written and comprehensive, containing a discussion on assumptions and limitations and a detailed supplementary material. The software written in Python is easily downloadable and all of the method components are transparent. The statistics in the places where it is employed, such as in the ROC curves, is also done adequately.**

**Comments:**

1) From the organizational standpoint, in some places the manuscript appears quite condensed. The use of subsections that is implemented in the Results section should be carried out in other sections as well, for example in the Discussion section, and in general more sub-sectioning can improve the readability.

2) On p.23 line 57, the dot-bracket representation is not entirely clear and could be explained in more detail (commonly, a simple dot-bracket represents an RNA secondary structure, and a more involved dot-bracket representation should be explained adequately). Does the number "16" in the example mean 16 repeats, which is related to some example in the text or some other realistic biological scenario?

3) I think that reference [2] in the Supplementary file that deals with SHAPE directed RNA folding should also be added to the main text.

4) On p.20 line 36, "published" is used twice in the same sentence.

5) On p.16 line 23: should be "In contrast patteRNA offers..." (instead of 'offer').

6) In the caption of Figure 4: should be "Dashed boxes highlight regions where a pseudoknot is likely present" (instead of 'were').

7) In the caption of Figure S4, the final sentence "it is more likely in 0mM NaF" should be explained a bit (referring to another place in the text).

8) In Table S3, one can observe various options that can be inserted to the command line for patteRNA. Is there a general explanation somewhere and a list of these options? It will also be helpful if in the supplementary information, an example of a simple session with a command line including input/output will be added.

Author response to reviewers

We thank the reviewers for their time, commitment and supportive comments. We address their comments below and refer to respective changes in the manuscript where relevant. In addition, and although not requested, we modified the motif scoring function used in *patteRNA*, which previously entailed an approximation, to an exact joint probability computation. The derivation of the new scoring scheme is described in Additional File 1 (see section “Motif scoring”, pages 11-13) and we updated the corresponding description in the main text (see page 21). We updated all the results and Figures 3-5 to reflect our analyses with the improved scoring function. Notably, we observed no substantial changes to the conclusions we previously reported. We observed the following:

1). Slightly better classifications for canonical structures using *patteRNA* (see Figure 3, improvement of 0.01 in the AUC).

2). For the fluoride riboswitch, the data even further supports the proposed structures (see Figure 4C-E and revised text in page 11). In particular, in Figure 4E, the third hairpin is now equally likely in both conditions up until a stem length of 7 and then becomes more likely in 0mM NaF, which is in perfect agreement with the proposed structures. For Figure 4F-G, we observed a slight increase in the background noise but the pseudoknot is still clearly different when looking at transcript's lengths between conditions.

3). We observed no substantial changes to our results for the PARS dataset (see Figures 5 and S5).

Note that all revisions made to the original manuscript and supplements are highlighted in blue and red in the revised manuscript. Our responses below are highlighted in blue.

**Reviewer # 1**

In the manuscript entitled "patteRNA: Transcriptome-wide search for functional RNA elements via structural data signatures" the authors present a novel computational method for the analysis of RNA structural probing data that may be applied on a transcriptome-wide level.  Overall this is a nice addition to an area that could use some additional methods for systematic treatment of the data rather than a combination of ad hoc methodologies. The paper is well-written and the examples utilized are well-chosen and convincing.

However, I have some concerns. The method itself is not more accurate than some of the methods it was tested against (which is too bad). As the authors point out, many of these methods are computationally intensive for large data sets, thus their method still may play an important role in the interpretation of transcriptome-wide data. However, this does decrease the significance of the work, especially since the authors do not provide any specific benchmarks. It is unclear how much faster patteRNA might be, and under what circumstances this might make a significant difference, sufficient to warrant the loss in accuracy.

Response: The reviewer is correct in pointing out that *patteRNA* does not necessarily perform better than ensemble-based predictions. This is in fact something we discussed in our original manuscript and we acknowledge his/her concern (see pages 8-9 and 16). Before we elaborate on the new benchmarks we performed to address this comment, it should be noted that the performance results shown in Figure 3 were obtained for a dataset (the Weeks set) containing RNAs for which NNTM-based predictions are known to remarkably improve when directed by SHAPE data. However, this is not universally true, as performance gains for data-driven NNTM predictions have been shown to vary significantly depending on RNAs. Therefore, it remains unclear how data-driven NNTM would perform across large sets of diverse RNAs. We feel this point was not well conveyed in the original version and we now clarify it in the main text (page 10).

It is also worth noting that thermodynamic models fall short when it comes to considering inter- and intra-molecular interactions, which are common *in vivo*. Such interactions are absent from the Weeks set, as it was obtained from *in vitro* purified RNAs. In other words, it is currently unknown to which extent predictions from thermodynamic models can be trusted in the complex environment of a cell. In contrast, SP experiments capture snapshot of all RNAs in a cell simultaneously and as a result are more likely to mine a realistic picture of all RNA structures at transcriptome-wide scale. We clarified this point in the Results section (see pages 9-10).

In summary, we believe that for these two reasons, *patteRNA* would still confer an advantage over NNTM methods in many scenarios that are of current interest to a large community of biologists. Furthermore, as we stated in the original text, *patteRNA* can be used to rapidly narrow down a motif’s search into a manageable set of candidate regions that can then undergo more careful investigations using NNTM-based methods if required, hence combining fast detection with comparable or higher accuracy. What we mean by fast detection as well as our response to the “runtime vs accuracy” question is provided below.

1. Since the authors do make such a big deal about the speed/scalability of patteRNA as an advantage over ensemble prediction methods, it would be nice to see a better quantitative comparison in terms of speed. How much does this actually matter given differing amounts of data/length of RNAs?

Response: Thank you for your suggestion. We agree that our statements were a bit vague and that we were missing on specific practical examples. Furthermore, we believe that a more in-depth analysis of runtime requirements would be *per se* of interest to the community. To address this, we extended our original runtime benchmark in several ways. First, we report runtimes for much longer RNAs and various datasets ranging in size (Supplementary Figure S3). Second, we include additional algorithms for MFE prediction and ensemble sampling from the ViennaRNA and RNAstructure packages (Supplementary Figure S3 and Table S4). Third, we developed a formula that would allow a user to predict runtimes based on a dataset size and its transcript-length composition (see section “Runtime benchmarks” and Supplementary Table S5 in Additional File 2). Fourth, we provide runtime estimates for the PARS dataset and the human transcriptome (hg19) (see Supplementary Table S6 and S7).

Our simulations show that *patteRNA* is significantly faster compared to alternative methods for transcriptome-wide datasets. *patteRNA* reduces average time requirements from months/years, using typical NNTM-based methods, to only a few hours/days (see Supplementary Figure S3 and Tables S4-7). The dramatic improvement in speed we observed would apply to any dataset that contains long RNAs (>2500 nt). In other words, *patteRNA* is the only algorithm that can practically detect motifs in transcriptome-wide SP data, while still maintaining the entire sequence-context of transcripts.

We hope that the new plots and numbers help put *patteRNA*’s performance in the context of relevant biological datasets and would resonate with a broader audience of experimentalists. We clarified these points in both the Results and Discussion sections (see pages 8, 9 and 16).

2. What is the audience that the authors are trying to reach with this work? Other computational biology experts, experimentalists? I feel like experimentalists will be the ones to benefit the most from it, but the paper itself feels like it is written for other experts and while the product is accessible to experts it might not be to the experimental audience.

Response: We also feel that experimentalists are the ones that will likely benefit the most from our work. However, this is a Method article and the methodology we present is novel, especially in the context of SP experiments. As such, it cannot be found elsewhere in the literature that is commonly read by our targeted audience. For this reason, we felt that it should be adequately described in the text. While we agree that some subsections might not be readily accessible to researchers with little mathematical or computational background, we believe that the main ideas brought forth are understandable, in particular the biological applications of *patteRNA*.

Finally, to facilitate access to the software itself, we prepared an extended Github documentation that describes in details the installation and usage of *patteRNA* (see the Github repo at https://github.com/AviranLab/patteRNA). In summary, we strongly believe that labs having the capabilities to generate SP data will also be able to use *patteRNA* with no particular difficulties.

Reviewer #2: In the manuscript by Ledda and Aviran, an unsupervised pattern recognition method that mines RNA structure motifs from profiling data is put forth. The method is called patteRNA. It is a machine learning algorithm that rapidly detects RNA structural motifs in large-scale structure profiling data. It can effectively detect motifs in various data sets and can be used to narrow down a set of candidate regions which can then undergo more careful nearest neighbor thermodynamic model analyses. Using three data sets, it is demonstrated that patteRNA detects motifs with accuracy comparable to commonly used thermodynamic models. The method is compatible with diverse profiling techniques and experimental conditions.

In general, the method is original and promising. It fills a needed gap in mining RNA structure motifs from genomic data and as more high-throughput structure profiling experiments will be performed in the future, which is expected, the usefulness of patteRNA will increase. The manuscript is also well written and comprehensive, containing a discussion on assumptions and limitations and a detailed supplementary material. The software written in Python is easily downloadable and all of the method components are transparent. The statistics in the places where it is employed, such as in the ROC curves, is also done adequately.

Comments:

1) From the organizational standpoint, in some places the manuscript appears quite condensed. The use of subsections that is implemented in the Results section should be carried out in other sections as well, for example in the Discussion section, and in general more sub-sectioning can improve the readability.

Response: We agree with the reviewer and added subsections within the Discussion section (see pages 14-16).

2) On p.23 line 57, the dot-bracket representation is not entirely clear and could be explained in more detail (commonly, a simple dot-bracket represents an RNA secondary structure, and a more involved dot-bracket representation should be explained adequately). Does the number "16" in the example mean 16 repeats, which is related to some example in the text or some other realistic biological scenario?

Response: Thank you for pointing this out. We added an “Extended dot-bracket notation” subsection in Methods (see page 20) that explains in detail the convention used by our notation. We added a cross-reference to this new subsection in the text highlighted by the reviewer (page 24) as well as in the Methods subsection “ROC analysis of motif predictions” (page 25) to ensure clarity. As to the specific example raised by the reviewer, “16” represents the length of the consecutive run of the symbol “(“, i.e., 16 repeated opening pairs “(“.

3) I think that reference [2] in the Supplementary file that deals with SHAPE directed RNA folding should also be added to the main text.

Response: Thanks for the suggestion. We added that reference to the main text. See pages 8 and 18.

4) On p.20 line 36, "published" is used twice in the same sentence.

Response: Thanks, we fixed the typo.

5) On p.16 line 23: should be "In contrast patteRNA offers..." (instead of 'offer').

Response: Thanks, we fixed the typo.

6) In the caption of Figure 4: should be "Dashed boxes highlight regions where a pseudoknot is likely present" (instead of 'were').

Response: Thanks, we fixed the typo.

7) In the caption of Figure S4, the final sentence "it is more likely in 0mM NaF" should be explained a bit (referring to another place in the text).

Response: We rephrased the sentence in an attempt to make it more understandable. Please see the new caption of Supplementary Figure S4, which now reads “A red color indicates that the data better supports the presence of pseudo-hairpins in the presence of fluoride, and inversely, blue in the absence of fluoride.”

8) In Table S3, one can observe various options that can be inserted to the command line for patteRNA. Is there a general explanation somewhere and a list of these options? It will also be helpful if in the supplementary information, an example of a simple session with a command line including input/output will be added.

Response: Thank you for this suggestion. An extended tutorial of *patteRNA* as well as all currently supported options are now publicly available as Github documentation (see Github repo at https://github.com/AviranLab/patteRNA). This documentation also includes a detailed description of typical inputs/outputs and several examples. While we agree with the reviewer that this is of critical importance, we believe such information should not be included in the manuscript directly as those could be subject to changes as we improve and add features to *patteRNA*. Moreover, we believe it will be easier for users to refer to the Github page for questions regarding software usage and to the manuscript for mathematical and implementation details of the method.