**Dear Dr. Greene**
**We appreciate the reviewers' efforts and comments on the submitted manuscript. Our detailed response to their comments is below.**

**Editor comments (Casey Greene)**
Reviewers #1 and #3 have additional points that should be addressed. In particular, please provide the literature support for the assertions noted by reviewer #1 and please address the issues raised with the validity concerns of reviewers #1 and #3.

> **We have provided the literature support and addressed the issues raised with respect to the validity of the study.**

Please also carefully define what you mean "validation" sets as noted by reviewer #3. Fields vary somewhat in terms of how this language is used (i.e., in some validation conveys an independent validation set, while in ML validation is not as strong as test).

> **We are now employing the term "test set" instead of "validation set", and have given a definition on line 243 of the meaning of "test set".**

**Reviewer 1 (Gaurav Pandey)**
**Basic reporting**
The authors have made a substantial effort to address the reviewers' comments, but several issues still remain:

1. The reasoning for the selection of the features on lines 68-71 should be supported by data and/or references.

> **References have been added.**

2. The authors state that the recall rate of promoter prediction programs isn't very high. So, what is the impact of this on the sRNA prediction results? Similar issues exist presumably for the other feature extractors as well. What is the impact of those?

> **We believe the impact is an increase of the number of false negatives of our model. Once better feature extractors become available one will be able to quantify their impact.**

3. All the statements added in lines 370-390 to explain the observed importance of the features in the RF model must be supported by data and/or references.

> **References have been added.**

4. The fonts in Figure 4 are too small to be readable.

> **The font size has been increased.**

5. The term "FDR corrected pvalue" is redundant – FDR is corrected pvalue, but it should also be mentioned how the FDR is obtained.

> **Reference has been added in lines 276-278. For clarity sake, we prefer to use the term FDR-corrected pvalue.**

6. The forward slashes in Figure 8 is an atypical way of showing the variation. Why not show a curve, or a region, whatever is appropriate?

**Lu et al (2011) provided a performance range of four methods. Thus, we are not showing variation, we are showing the range as reported by Lu et al. Nevertheless, we have replaced the forward slashes with a distinctive rectangle.**

## Experimental design

1. The reasoning for only creating a 1:3 unbalanced dataset isn't satisfactory, especially because this factor is expected to have a major impact on the prediction performance. So, this ratio has to be varied to measure this impact.

**We have varied the ratio and found that it does not have a major impact on the prediction performance of sRNARanking. A new subsection in Results presenting this finding, and additional figures 2-8 have been included, that show the PRC of RF-Combined models constructed with different ratio of positive to negative examples on all test datasets.**

## Validity of the findings

1. As stated on line 361, "sRNARanking's specificity of 88% is comparable to the 91% specificity of Barman et al's approach at a sensitivity of 85%.". So, what is the benefit of this study compared to Barman's?

**We have now executed Barman et al's approach with exactly the same test datasets (Lu's and SLT2) and our method clearly outperforms Barman et al's approach. We have updated our manuscript to reflect this (lines 290-300 and 365-375). The benefit is that we are showing that better predictive performance can be obtained using a set of genomic context features. Additionally, as mentioned by reviewer 3, this study might motivate the development of another method combining ours and Barman's.**

2. Why isn't LR included in the CD plot in Figure 6? Also, unless there is a novel interpretation of the CD plot (the order seems to be reversed), RF seems like the worst performer instead of the best, i.e., is the rightmost entry instead of the expected leftmost.

**We are not using a novel interpretation of the CD plot. In fact, we are using the same interpretation as in the original manuscript that introduced CD plots ( *Statistical Comparisons of Classifiers over Multiple Data Sets,* Demsar, 2006). As it is explained by Demsar (page 15) "The axis is turned so that the lowest (best) ranks are to the right since we perceive the methods on the right side as better".**
**As LR was clearly outperformed by the other four classifiers, we excluded LR results from further analysis (already explained in lines 322-325).**

3. The utility of the violin plot in Figure 3 is still unclear. Typically, when such a plot is shown, there is a variable on the Y-axis, e.g. individual genes' expression, and the plots show the dependence of this value on the factors listed on the X-axis. So what is the parallel interpretation here? Are the authors calculating a per-example AUPRC? This has to be explained clearly when Figure 3 is mentioned on line 309.

**The variable on the Y-axis is the AUPRC per model on a specific test set. For each classifier, four models were generated (one for each training set). Each of these models was evaluated**

**on each of the five test sets. We are showing the dependence of the AUPRC of the models on the classifier used to construct the models. This has been explained in lines 320-322.**

**Reviewer 2 (Anonymous)**

**Basic reporting**

I have no additional comments.

**Experimental design**

I have no additional comments.

**Validity of the findings**

I have no additional comments.

**Reviewer 3 (Anonymous)**

**Basic reporting**

The authors have addressed most of the concerns raised in my initial review. The comparison with existing methods improved the manuscript significantly. However, several comments remain.

**Experimental design**

- The authors added a comparison with other published methods. Although the same positive examples were used, it's not clear if the method for choosing negative examples, and their numbers were the same.

**We have now executed Barman et al's approach with exactly the same test datasets (Lu's and SLT2) and the new results are included in the manuscript.**
**As already explained in the previous version of the manuscript (lines 109-114), we are not using the same method for choosing negative examples. Arnedo et al and Barman et al used shuffled genomic sequences while we are using sequences from random genomic locations. In other words, our negative instances are real genomic sequences present in the bacterial genomes, while Barman et al's negative instances are artificial sequences <u>not</u> present in the bacterial genomes.**

- I agree with reviewer Pandey remark about the evaluation procedure employed by the authors, and have additional comments in that regard. In machine learning the term "validation set" refers to data that is used for choosing classifier hyperparameters. The authors appear to use this term to describe the test set on which performance is evaluated.

**We are now employing the term “test set” instead of “validation set”, and have given a definition on line 243 of the meaning of “test set”.**

Furthermore, the fact that the authors mention specific hyperparameter values as optimal when discussing the various classifiers is inconsistent with the mention that multiple training sets were used, as each choice of training set might end up with different optimal hyperparameters.

**In the manuscript in lines 167-172, we have explained that “For each classifier, the ‘best’ parameters were obtained by maximizing the average area under the ROC curve (AUC) when performing leave-one-out cross-validation (LOO CV) on the training data (Fig.2). The final ‘best’ parameters per classifier used were those given the largest average AUC.”**

The model selection/evaluation proposed by the reviewer Pandey makes a lot more sense than the one used by the authors in view of the small number of examples. Measuring the variability of the performance over several training runs has limited usefulness, and in algorithms such as logistic regression which has a single global optimal solution that is even more the case.

**As our results show, classifiers have large variability of performance over several training runs; thus we still believe that having an understanding of the variability of the classifiers over several training runs is useful.**

Finally, using LOO cross validation for parameter selection is clearly overkill. Nested cross-validation is the standard technique for classifier evaluation and model selection in the area of machine learning, and is well supported by packages such as scikit-learn. The following paper is a good reference on this topic, and goes even further, suggesting repeated nested cross validation as a good approach for small datasets: https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-6-10

**Nested cross-validation is recommended to estimate the prediction error (model assessment); however, we are calculating test error on three completely independent test sets. Using independent test sets for model assessment provides an unbiased estimate of test error. Krstajic et al, in the manuscript referred above by the reviewer, wrote "In an ideal situation we would have enough data to train and validate our models (training samples) and have separate data for assessing the quality of our model (test samples)". Since this is the case in this research, we are assessing the quality of our model on separate data (test sets).**
**We used LOO cross-validation for parameter tuning (model selection). Leave-one-out cross-validation is also a standard technique in the area of machine learning. Chapter 7 of *The Elements of Statistical Learning* (2009) by Hastie et al (a book cited by Krstajic et al) discusses cross-validation in detail and explains that LOO cross-validation has lower bias than 5- or 10-fold cross-validation in small datasets.**

## Validity of the findings

- In my initial review I mentioned the fact that the method is likely of limited use when applied genome-wide, because it is likely to produce a large number of false positives.
My comment was:

The accuracy of the method is likely not sufficient to be useful for applying the method on a genome-wide scale, where even a small false positive rate can lead to many erroneous predictions.

The answer of the authors was:

As we demonstrated in the revised version of the manuscript, our method outperforms or is comparable to existing approaches in terms of precision/recall and sensitivity/specificity.

While that may be true, it does not refute my initial comment. It is very unfortunate that researchers developing methods for sRNA detection choose to use relatively balanced datasets while testing their methods. This evaluation methodology gives no indication of method usability when used genome-wide (area under the precision recall curve is not invariant to the ratio of positive to negative examples), and can give biologists the wrong impression about the usability of the method. Furthermore, the majority of the features computed by the authors' method (except for the free energy feature) are insensitive to shifting the region over which they are computed. That is another reason for the limited usability genome-wide. Perhaps when combined with a method such as Barman et al, which uses primary sequence, more specificity can be obtained. Overall, it appears like the problem of sRNA identification directly from genomic sequence is a hard unsolved problem, and this should be indicated.

**We already had a sentence indicating this in the manuscript in lines 39-41: "Computational prediction of sRNAs in genomic sequences remains a challenging problem, even though tools to tackle this problem have been around since early 2000s". We have changed the wording to**

**clarify this: “Despite the fact that tools to tackle this problem have been around since early 2000s, computational prediction of sRNAs in genomic sequences remains a challenging unsolved problem” and have added a sentence in lines 455-456 of the manuscript to indicate that there is still room for improvement in computational sRNA detection: “Although sRNARanking outperformed the other published methods in the benchmark datasets, there is still room for improvement of computational identification of sRNAs from genomic sequences. ”.**

## Comments for the Author

Minor comments:

- In line 100 n is used to denote the number of negative examples, but in lines 102-105, n is used differently to denote the relative fraction of negative to positive examples.

**The sentences have been re-written to consistently use *n* as the number of negative examples.**

- In the description of logistic regression you reference "balanced" mode of some implementation of LR, but no mention is made which software was used. For the other classifiers specific software/version should be mentioned as well.

**In lines 167-172 we had already mentioned that: “All the machine learning classification approaches were implemented in the Python programming language version 3.6. Scikit-learn (version 0.19.1) (Pedregosa et al., 2011) was used for the implementation of all the classifiers.”**