Published in final edited form as:

Methods. 2014 May 1; 67(1): 28–35. doi:10.1016/j.ymeth.2013.10.002.

Using machine learning and high-throughput RNA sequencing to classify the precursors of small non-coding RNAs

Paul Ryvkin^{a,b}, Yuk Yee Leung^{a,c}, Lyle H. Ungar^{a,b,d}, Brian D. Gregory^{b,e,f,*}, and Li-San Wang^{a,b,c,e,g,*}

^aInstitute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

^bGenomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

^cDepartment of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

^dDepartment of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA

^ePenn Genome Frontiers Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

^fDepartment of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

^gInstitute on Aging, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract

Recent advances in high-throughput sequencing allow researchers to examine the transcriptome in more detail than ever before. Using a method known as high-throughput small RNA-sequencing, we can now profile the expression of small regulatory RNAs such as microRNAs and small interfering RNAs (siRNAs) with a great deal of sensitivity. However, there are many other types of small RNAs (<50 nt) present in the cell, including fragments derived from snoRNAs (small nucleolar RNAs), snRNAs (small nuclear RNAs), scRNAs (small cytoplasmic RNAs), tRNAs (transfer RNAs), and transposon-derived RNAs. Here, we present a user's guide for CoRAL (Classification of RNAs by Analysis of Length), a computational method for discriminating between different classes of RNA using high-throughput small RNA-sequencing data. Not only

© 2013 Elsevier Inc. All rights reserved

*Correspondence and requests for materials should be addressed to: Li-San Wang 1424 Blockley Hall University of Pennsylvania 423 Guardian Dr. Philadelphia, PA 19104 Telephone: (215) 746-7015 lswang@mail.med.upenn.edu. Brian D. Gregory 204G Lynch Laboratory Department of Biology University of Pennsylvania 433 S. University Ave. Philadelphia, PA 19104 Telephone: (215) 746-4398 bdgregor@sas.upenn.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

can CoRAL distinguish between RNA classes with high accuracy, but it also uses features that are relevant to small RNA biogenesis pathways. By doing so, CoRAL can give biologists a glimpse into the characteristics of different RNA processing pathways and how these might differ between tissue types, biological conditions, or even different species. CoRAL is available at <http://wanglab.pcbi.upenn.edu/coral/>.

Keywords

small RNAs; microRNAs; small interfering RNAs; non-coding RNAs; RNA-seq; machine learning

1. Introduction

Small RNAs are short RNA (< 50 nt) transcripts that are abundant in cells across all domains of life. In mammals one can find microRNAs as well as fragments of longer RNAs such as small nucleolar RNAs (snoRNAs), tRNAs, small nuclear RNAs (snRNAs), small cytoplasmic RNAs (scRNAs), ribosomal RNAs (rRNAs), *piwi*-interacting RNAs (piRNAs), and RNAs derived from transposable elements. MicroRNAs in particular are the best characterized type of small RNA and have a well-established role in the regulation of gene expression at the post-transcriptional level. Fragments that are processed from tRNAs are thought to be a product of stress response pathways [1,2]. While RNAs processed from snoRNAs and snRNAs are more mysterious, they are mostly thought to be degradation products, but can also sometimes demonstrate a regulatory mechanism similar to microRNAs [3-5]. There is also evidence that piRNAs and RNAs derived from transposable elements are involved in both the silencing of these elements [6] and in epigenetic regulation of gene expression [7,8].

The advent of high-throughput RNA sequencing (RNA-seq) has allowed researchers to probe the transcriptome in ever more detail and with increasing sensitivity. In particular, small RNA-sequencing, in which small RNAs are size fractionated from total RNA, is being used to study populations of RNA fragments in ever more detail. This method is most often used to profile the expression of microRNAs, largely because they are the most abundant class of small RNA (< 50 nt) and because the mechanism of their regulatory effect is well elucidated. Less well-understood small RNAs are often ignored in such studies, despite receiving a significant fraction of read coverage. High-throughput RNA-sequencing provides researchers with much more information than simply expression levels, and when examined at a deeper level by analyses such as the one presented here can help to elucidate biological processes such as RNA cleavage.

Here we present an overview and user's guide for the computational methodology called CoRAL [9] (Classification of RNAs by Analysis of Length), which can build a model that discriminates amongst classes of small non-coding RNAs and inform the user about how various characteristics of the small RNAs differ across the classes. While length turns out to be a very important feature for classification, CoRAL includes several other feature types when discriminating between and describing different classes. It takes as input a small RNA-sequencing dataset and can either build a machine learning model (implemented using

a random forest classifier) or can use an already existing model to predict RNA classes in a new RNA-seq dataset. CoRAL can also be used to classify small RNAs that differ between biological conditions, cell types, or tissues. Since CoRAL is based on features of small RNAs that are thought to be relevant to RNA processing pathways, its results lend themselves better to interpretation than black-box type features which simply examine the expression levels and genomic separation of transcribed RNA blocks. As such, CoRAL can be used to examine how these properties of different classes of small RNAs change under different conditions. It can also be used to annotate newly discovered small RNAs or RNAs in species where there is no RNA annotation. Importantly, CoRAL requires only that the input data be from a size-fractionated library preparation, thus allowing analysis to be performed on data that was only used to gather information such as expression levels. This flexibility allows the user to mine existing or publicly available RNA-seq datasets in order to glean new insights.

2. Materials and Methods

2.1. Processing small RNA reads

CoRAL operates on small RNA-seq datasets – specifically, RNA that has been size fractionated to select for small RNAs and then sequenced. This data can either come from a user's own laboratory (novel sequencing data), or analyses can be run on pre-existing public small RNA-seq datasets. Such public datasets can, if the user desires, be obtained from a data repository such as Gene Expression Omnibus (GEO) or the Sequence Read Archive (SRA). In addition to coming from size-fractionated RNA, the libraries must be constructed using a strand-specific protocol such as Illumina Truseq Small RNA, thus ensuring that strand specificity of the reads is known. In addition, the small RNA fraction should be size selected to approximately 15-30nt, although going up to 50nt yields better coverage of longer RNAs such as tRNA fragments.

Processing of small RNA-seq data prior to CoRAL consists of, in short: trimming the 3' adapter sequences, mapping to the genome, and enforcing certain alignment criteria (Fig. 1). In order to ensure that the reads we obtain from these sources correspond to fragments in the desired length range, we search for the presence of the 3' adapter sequence that was ligated during library preparation. These adapter sequences are expected to be present in most reads from a small RNA sequencing library as most of the sequenced fragments are shorter than the number of sequencing cycles. For example, for 50 bp sequencing, most of the fragments will have been shorter than 50 nt and thus sequencing will often proceed into the 3' ligated adapter. Searching for and removing this sequence was accomplished via the **cutadapt** program [10], a set of python scripts that search and remove a given adapter sequence from sequencing reads. We run cutadapt with the following parameters:

```
bowtie —error-rate=0.06 —overlap=6 —minimum-length=14 —times=2
```

These options enforce a number of criteria for detecting the 3' adapter sequence: it must be present with an error rate of at most 6%, and must be represented by at least 6 nucleotides. Its presence as a dimer should also be considered a hit, thus the “—times=2” option is selected. These options allow us to find the adapter sequence while being robust to

mismatches stemming from sequencing errors. Additionally, we require a minimum length of 14nt for the trimmed fragment sequence – this is a reasonable lower bound on what is uniquely mappable to the human genome. The result is three FASTQ files: one containing the sequences that were trimmed to shorter than 14nt, one containing the sequences where no adapter was found, and finally, the set of sequences where the adapter was found and subsequently removed by cutadapt. All downstream analyses and CoRAL are run only on this last file containing properly trimmed reads. The untrimmed reads originated from RNAs of uncertain length greater than (sequenced read length minus 6 adapter nucleotides), and so we cannot reliably use their length distribution as an feature of small-RNA biogenesis pathways.

Once we have obtained the trimmed reads, we map them to the reference genome sequence of the model organism (e.g. the hg19 assembly of the human genome) using the alignment program bowtie [11] with the following parameters:

```
–v 2 —best –strata –m1 –sam
```

producing a SAM file. We take only the uniquely-mapped reads (“–m1”) as we want to avoid contamination between training and testing sets during cross-validation; multiple-mapping reads could result in an overfitted training model and inflate the training accuracy. The reader can choose other modern read-alignment tool for read mapping; for example, our experience indicated that BWA [12] behaved similarly to bowtie when mapping short reads in small RNA-seq data. The resulting SAM file is then filtered in order to enforce a maximum mismatch rate of 6% for each read. This length-independent mismatch rate is necessary when reads are of heterogeneous lengths (as in small RNA-seq) in order to avoid under-penalizing mismatches on shorter reads, which would happen with a fixed number-of-mismatch criterion. The result is that we enforce a stricter absolute mismatch limit on shorter reads while allowing more mismatches on longer reads. Once all of this processing is complete, the result is a BAM file containing the mapped, trimmed read alignments that can be used by CoRAL.

2.2. CoRAL: Classification of RNAs by Analysis of Length

2.2.1. Calling transcribed loci—CoRAL begins by taking an input BAM file and predicting transcribed regions of the genome. Since no assumptions are made about transcribed regions in the genome, and discrete items are required as input to any machine learning algorithm, the RNA-seq reads must be clustered into discrete segments corresponding to genomic loci where small RNAs are being generated. Here we call these regions “small RNA loci”. For this purpose we use the **bgrSegmenter** tool, which belongs to the **RSeqTools** suite [13]. bgrSegmenter implements the minRun-maxGap algorithm that seeks to coalesce adjacent and nearly-adjacent intervals into clusters given a minimum run length and maximum allowed gap length. We use a minimum run length of 14 nt, corresponding to the length of the smallest allowed read and set the maximum gap length to 44 nt, which corresponds to the length of the longest allowed (trimmed) read. The final required parameter is the minimum coverage required to report an interval as transcribed; we use 20 reads in our analysis. Within CoRAL this locus calling step can be accomplished by running the following command:

```
call_smrna_loci.sh bam_file conf_file
```

where *bam_file* contains the aligned small RNA-seq reads and *conf_file* is the CoRAL configuration file. The result is stored in a file named “coral/loci.bed.” This file is a list of all the loci called by the algorithm, i.e., intervals corresponding to transcribed small RNA loci. In this BED file, the “score” column contains the read count for each locus.

2.2.2. Generating features for small RNAs—In order to build a classification model for small RNA loci, we must first transform the RNA-seq data into a set of features that are amenable to machine learning algorithms. CoRAL uses the following set of biologically relevant features: distribution of processed RNA lengths, positional entropy of the processed fragments, nucleotide frequencies, presence of antisense transcription, and predicted minimum free energy (MFE). The commands for generating these features are as follows:

```
feature_lengths.sh bam_file conf_file
```

```
feature_antisense.sh bam_file conf_file
```

```
feature_entropy.sh bam_file conf_file chromInfo.txt
```

```
feature_nuc.sh bam_file conf_file
```

```
feature_mfe.sh bam_file conf_file genome_fasta chromInfo.txt
```

These tasks can, if the user desires, be run in parallel. *chromInfo.txt* contains the lengths of chromosomes in the genomic sequence. *genome_fasta* is a FASTA file containing the entire genomic sequence that was used to map the reads. Both files contain the genome annotation information and are therefore specific to the genome under study.

2.2.3. Building a classification model—In order to build a classification model of small RNA loci, one first needs to create a training dataset which contains the annotations of all known transcribed RNAs. We have generated GFF files that can be downloaded from the CoRAL website (<http://wanglab.pcbi.upenn.edu/coral/>) and these can be used for training. In short, the annotation is created by conglomerating various UCSC genome browser [14] tracks (e.g., knownGene, wgRna, tRNA, rnsk) and prioritizing overlapping annotations by the expected abundance of a given class of RNA in the transcriptome. We use the following prioritization in decreasing order in CoRAL: rRNA, as-rRNA, tRNA, mt-tRNA, as-tRNA, as-mt-tRNA, miRNA, snRNA, snoRNA, srpRNA, scRNA, piRNA, as-miRNA, as-snRNA, as-snoRNA, as-srpRNA, as-scRNA, as-piRNA, mRNA_exon, lincRNA_exon, ncRNA_exon, 7SK, transposon, LTR, rnsk_repeat, rnsk_tRNA, asmRNA_exon, as-lincRNA_exon, as-ncRNA_exon, as-7SK, as-transposon, as-LTR, as-rnsk_repeat, mRNA_intron, lincRNA_intron, ncRNA_intron, as-mRNA_intron, as-lincRNA_intron, as-ncRNA_intron, intergenic. The prefix “as” denotes intervals that are antisense with respect to the given annotation type. To create the training data by labeling the called small RNA loci with known annotations, CoRAL provides the following commands:

```
annotate_loci.sh coral/loci.bed genome_gff class_pri.txt
```

Here, *genome_gff* is the aforementioned genomic annotation GFF and *class_pri.txt* contains the RNA class priorities described above.

```
make_data_matrix.sh coral
```

The `make_data_matrix.sh` command combines all the computed features into one matrix file (*data_x.txt*) and all of the annotation labels into one text file (*data_y.txt*).

In order to build a predictive model using the training data, CoRAL provides the following command:

```
coral_train.R -r min_coverage -c classes coral/data_x.txt \ coral/data_y.txt coral
```

Here *min_coverage* is the inclusion/exclusion criterion – the minimum coverage (number of reads) required at a locus in order to include it in the analysis. Replace the word *classes* with a comma separated list of classes to be included in the model, e.g.

“miRNA,tRNA,snoRNA_CD” will build a model that can distinguish between microRNAs, tRNAs, and C/D box snoRNAs. The resulting model will be stored in a subdirectory called “run_xxx” where xxx is a hashed version of the parameters used; in other words, this string will be unique for each set of parameters passed to `coral_train.R`.

CoRAL also reports the importance of particular features for each class used in the model. The output file “feature_importance.txt” under the “run_xxx” subdirectory contains a matrix in which each row is a feature and each column is a class. The value in a cell corresponds to the number of times the feature was selected in the one-vs.-all random forest model for that particular class. As random forest consists of an internal bootstrap step, the loci used for building each tree in a forest are randomly selected. Therefore the importance of features in each of the trees will vary. As a result, multiple feature selection runs are performed to minimize the potential bias. The number of feature selection runs can be specified by the `-varsel-runs` option, and this is the denominator used when computing the proportion of times a feature was selected (the feature importance). Another file, “feature_direction.txt,” contains a similarly-designed matrix consisting of +1 or –1 in each cell. These values specify whether the value of the feature was, on average, lower or higher in that particular class, when compared to the others.

2.2.4. Prediction of RNA precursor class—Given the set of features and trained classification model, one can use these to predict the classes of the loci in the second dataset (leave-out or testing dataset) (Table 1). The following command uses a previously generated model to predict classes on a given dataset:

```
coral_predict.R -r 15 [-y coral/data_y.txt] coral/data_x.txt model_dir output_dir
```

where the `-r` option again specifies the minimum number of reads at a locus, *model_dir* is the “run_xxx” directory generated by a prior training run and *output_dir* is the destination directory for the prediction output. The “-y” option is only used if one wishes to compare the predictions to already-known annotations. In the output directory a file will be created called “pred.txt”, which contains the locus information and class prediction for each locus. If

“-y” was specified, another file called “pred_class_performance.txt” that contains performance measures for each locus type identified by the classifier will be created. The performance measures include sensitivity, positive predictive value (PPV), and specificity.

2.2.7. Visualization of small RNA reads and secondary structure—Visualization of the reads and loci in the UCSC genome browser can be accomplished by uploading the “loci.bed” file that CoRAL produces as a custom track. This will display the intervals corresponding to expressed small RNA loci. In addition, one can upload the input BAM file to view the reads and their correspondence with the called loci (Fig. 3). To take a closer look at some of the loci (especially unannotated, uncharacterized ones), one can visualize the small RNA reads and how they fall onto the RNA secondary structure by utilizing our **SAVoR** (Sequencing Annotation and Visualization of RNA structures) web application [15]. This can be performed by entering the coordinates of a locus of interest as well as providing a link to the BAM file at the SAVoR website. The output will be the RNA secondary structure (whether as annotated by RFAM or predicted on the fly by RNAfold), along with colors denoting the abundance of small RNA read coverage. SAVoR is useful for visualizing, among other things, the pattern of small RNA reads (as they are processed from their precursor RNA) and how this correlates with the secondary structure (Fig. 4).

3. Theory

3.1. RNA processing pathways

The impetus for CoRAL’s selection of features comes from the specificities of known RNA processing pathways. The most well characterized small RNA biogenesis pathway in mammals is that of microRNAs (miRNA). microRNAs are short regulatory RNAs, that are processed from longer primary microRNA (pri-miRNA) transcripts into pre-miRNAs in the nucleus by a protein complex composed of DGCR8 and Drosha. The resulting pre-miRNAs are then exported to the cytoplasm, and processed by Dicer into mature miRNAs. These Dicer products are almost always 22 nucleotides in length, unlike small RNAs coming from other types of loci. For example, piRNAs tend to be considerably longer [6]. For this reason, we use the distribution of small RNA lengths produced from a transcript as one of the principal features in CoRAL. Additionally, we look for the presence of antisense transcription at a locus, which is a phenomenon that often occurs in the silencing of transposable elements [16]. Antisense reads can also indicate the presence of natural antisense transcription, an important source of double-stranded RNA and potential small RNA precursor in the eukaryotic transcriptome [17]. Another feature of small RNA biogenesis pathways is the distribution of cleavage sites. We encode this aspect of small RNAs as a feature called “positional entropy.” The justification for looking at the positions of cleavage sites comes from the observation that microRNAs tend to have a fairly consistent 5’ cleavage site but they show considerably more variation on the 3’ end [18]. This is predicted to be a result of the particular secondary structure of each specific microRNA, and therefore can be used to differentiate amongst even this specific class of smRNA. Given that we are using smRNA-sequencing data, we specifically obtain the sequences of the small RNAs themselves – we leverage this information by including nucleotide frequency as another feature. This is because it is reasonable to expect that the

smRNA-producing cleavages will have sequence-specific biases. Finally, we also use the genomic sequence of the loci to predict secondary structure; since RNA processing proteins are often structure-specific, it is reasonable to expect that structural motifs at and around a small RNA locus can be used to differentiate between small RNA processing pathways. As a proxy for secondary structure at a locus we use the minimum free energy (MFE) of the predicted structure. This value should reflect the proportion of nucleotides that are base-paired at each given locus.

3.2. Small RNA feature computations

Motivated by the observations above we introduce five classes of features for classifier construction: distribution of RNA fragment lengths, presence of anti-sense short RNA fragments, specificity of cleavage sites as endpoints of RNA fragments, nucleotide frequency, and minimum free energy from *in silico* RNA secondary structure folding. The first four classes are extracted from the small RNA-seq data and the last feature is obtained by the RNA sequence of the entire transcribed locus.

Fragment length distribution—From the input file CoRAL examines the trimmed lengths of all trimmed reads covering a locus and compute features corresponding to the distribution of these lengths. These features measure the tendency, at a given locus, for reads to be of particular lengths and not of other lengths. This can be quite important for picking out microRNAs, since they have a very specific length distribution around 22nt.

Presence of antisense transcription—Since CoRAL requires that the RNA-sequencing library must be strand-specific, we can distinguish between reads mapping to the sense strand of a locus and reads mapping to the antisense strand. Given a transcribed region, the sense strand is defined as the strand that contains more reads. CoRAL first computes the number of mapped reads from the anti-sense strand. The strand specificity feature is set to 1 when the number of antisense reads at a locus passes a particular threshold (specified as a parameter) and 0 otherwise.

Specificity of cleavage sites by positional entropy—In order to quantify the “cleavage specificity” of the RNA fragments, in terms of position along the precursor, CoRAL computes two features: 5′ positional entropy and 3′ positional entropy. These features measure the entropy of the smRNA reads relative to the transcribed locus start position. We distinguish between 5′ ends of fragments and 3′ ends of fragments because these distributions can be different: in the case of microRNAs, 5′ cleavage tends to be more consistent. Entropy is a convenient quantity that measures the uniformity of a distribution. For example, assume p_i is the proportion the 5′ end a RNA fragment coincides with the i^{th} nucleotide of a N-nt-long transcript, the 5′ positional entropy is defined as

$$\sum_{i=1}^N p_i \log_2 p_i.$$

We use base-2 log to follow common conventions but other log bases can be used as well. The 3' positional entropy is defined similarly. These features measure specificity versus randomness when cleavage locations along the transcript is considered.

Nucleotide frequencies in fragments—CoRAL computes the nucleotide frequencies of the mapped reads at a locus weighted by abundance of the reads. This feature is intended to measure bias for or against particular nucleotides, which could have an impact on RNA processing specificity.

Minimum free energy of secondary structure—CoRAL computes the minimum free energy (MFE) for each locus by using **RNAfold** [19] to predict the secondary structure given the genomic sequence at the locus. RNA secondary structure is likely to be an important determinant in RNA processing. The sequence used for folding consists of the locus plus 40 nt of genomic sequence up and downstream.

3.3. Visualizing data and classifier modeling using the random forest algorithm

The data matrix can be visualized by projecting them into a 2-dimensional space using a method such as PCA (principle component analysis), MDS (multidimensional scaling), or CVA (canonical variate analysis), and coloring each locus by RNA class. Statistical software such as R [20] can generate two-dimension plots (e.g. the biplot command) while drawing lines indicating how individual features correlate with each other and how informative they are regarding the separation of RNA loci. This is an informative visualization for how separable the classes are. For example, Fig. 5 shows that the proportion of 21-22nt-long RNA fragments are the best features that separate miRNA from the other two RNA classes, since the maximal separation between miRNA and the other two classes is along the corresponding feature lines. If it is clear that the features are adequate at separating the classes, then it is reasonable to continue with building the machine learning model. CoRAL uses the *random forest* (RF) algorithm for RNA locus classification. RF is an ensemble-learning classification method [21] that is widely used in bioinformatics. In short, the RF algorithm builds a large number of randomized decision trees constructed by optimally splitting nodes on a subset of the input features. The final prediction is produced by a majority vote of the output of the trees. RF further computes the out-of-bag error as an estimate to the classification accuracy, and the importance score for each input feature by shuffling the input data. CoRAL outputs the importance score for each input feature for each class of RNA, so the user can identify features that are specific to each RNA class. Readers who are interested in the implementation can refer to [19].

4. Conclusion

The method described here, CoRAL, not only allows for more accurate classification of small RNA loci than existing methods, but also has the ability to provide biologically relevant data to the user. By treating the features not as a black box but as a window into small RNA biogenesis pathways, the user can see how these pathways differ between different conditions that have been analyzed by RNA sequencing. In addition, since the *in silico* method requires no special treatment at the library preparation stage, it can be applied

retroactively to any existing small RNA datasets, thus yielding new biological insights from these previously published studies.

We have performed a comparative study [9] on the classification accuracy of CoRAL with two other RNA-seq-based tools for RNA classification, DARIO [22] and miRDeep2 [23]. Since DARIO only classifies between three non-coding RNA classes (miRNAs, C/D box snoRNAs and tRNAs), we tested CoRAL and DARIO on these three datasets using a human brain smRNA-seq dataset. CoRAL's overall predictive performance was 91%, or 4% improvement over the same analysis performed by DARIO. A comparison with miRDeep2 showed that CoRAL had a lower precision (91% vs 98%) but better sensitivity (88% vs 81%) for miRNA detection, although the lower precision may be due to miRNA loci that are previously unannotated. On top of the better performance, CoRAL classifies six non-coding RNAs by default and can be generalized to more classes. Additionally, the ability to visualize the differences of the non-coding RNA classes under study using CoRAL makes it an extremely appealing and powerful tool. This makes CoRAL a more generalizable algorithm and a better choice for RNA-seq-based non-coding class RNA annotation.

Acknowledgments

This work was supported by the National Institute of General Medical Sciences [R01-GM099962], the National Human Genome Research Institute [T32-HG000046 to PR], the National Institute on Aging [U24-AG041689, U01-AG032984, and P30-AG010124 to LSW], the National Institute on Neurological Disorders and Stroke [P50-NS053488 to LSW], the Penn Alzheimer's Disease Center [P30-AG010124 to BDG, LSW and YYL], and the National Science Foundation [MCB-1053846 and MCB-1243947 to BDG]. Brain samples were obtained from the Center for Neurodegenerative Disease Research.

References

1. Lee YS, Shibata Y, Malhotra A, Dutta A. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev.* 2009; 23:2639–2649. [PubMed: 19933153]
2. Gebetsberger J, Zywicki M, Künzi, Andrea, Polacek N. tRNA-Derived Fragments Target the Ribosome and Function as Regulatory Non-Coding RNA in *Haloferax volcanii*. *Archaea.* 2012; 2012
3. Brameier M, Herwig A, Reinhardt R, Walter L, Gruber J. Human box C/D snoRNAs with miRNA like functions: expanding the range of regulatory RNAs. *Nucleic Acids Res.* 2011; 39:675–686. [PubMed: 20846955]
4. Ono M, Scott MS, Yamada K, Avolio F, Barton GJ, Lamond AI. Identification of human miRNA precursors that resemble box C/D snoRNAs. *Nucleic Acids Res.* 2011; 39:3879–3891. [PubMed: 21247878]
5. Scott MS, Avolio F, Ono M, Lamond AI, Barton GJ. Human miRNA precursors with box H/ACA snoRNA features. *Plos Comput. Biol.* 2009; 5:e1000507. [PubMed: 19763159]
6. Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, Zamore PD. A Distinct Small RNA Pathway Silences Selfish Genetic Elements in the Germline. *Science.* 2006; 313:320–324. [PubMed: 16809489]
7. Huang XA, Yin H, Sweeney S, Raha D, Snyder M, Lin H. A major epigenetic programming mechanism guided by piRNAs. *Dev. Cell.* 2013; 24:502–516. [PubMed: 23434410]
8. Di Giacomo M, Comazzetto S, Saini H, De Fazio S, Carrieri C, Morgan M, et al. Multiple Epigenetic Mechanisms and the piRNA Pathway Enforce LINE1 Silencing during Adult Spermatogenesis. *Mol. Cell.* 2013; 50:601–608. [PubMed: 23706823]
9. Leung YY, Ryvkin P, Ungar LH, Gregory BD, Wang L-S. CoRAL: predicting non-coding RNAs from small RNA-sequencing data. *Nucleic Acids Res.* 2013

10. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. 2011; 17
11. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009; 10:R25. [PubMed: 19261174]
12. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics*. 2010; 26:589–95. [PubMed: 20080505]
13. Habegger L, Sboner A, Gianoulis TA, Rozowsky J, Agarwal A, Snyder M, et al. RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics*. 2011; 27:281–283. [PubMed: 21134889]
14. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, et al. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res*. 2011; 39:D876–882. [PubMed: 20959295]
15. Li F, Ryvkin P, Childress DM, Valladares O, Gregory BD, Wang L-S. SAVoR: a server for sequencing annotation and visualization of RNA structures. *Nucleic Acids Res*. 2012; 40:W59–64. [PubMed: 22492627]
16. Malone CD, Hannon GJ. Small RNAs as Guardians of the Genome. *Cell*. 2009; 136:656–668. [PubMed: 19239887]
17. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, et al. Antisense Transcription in the Mammalian Transcriptome. *Science*. 2005; 309:1564–1566. [PubMed: 16141073]
18. Ebhardt HA, Tsang HH, Dai DC, Liu Y, Bostan B, Fahlman RP. Meta-analysis of small RNA sequencing errors reveals ubiquitous post-transcriptional RNA modifications. *Nucleic Acids Res*. 2009; 37:2461–2470. [PubMed: 19255090]
19. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. *Monatshefte Für Chemiechemical Mon*. 1994; 125:167–188.
20. The R Core Team. R Foundation for Statistical Computing. Vienna; Austria: 2013. R: a language and environment for statistical computing. ISBN: 3-900051-07-0
21. Breiman L. Random forests. *Mach. Learn*. 2001; 45:5–32.
22. Fasold M, Langenberger D, Binder H, Stadler PF, Hoffmann S. DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res*. 2011; 39:W112–117. [PubMed: 21622957]
23. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res*. 2012; 40:37–52. [PubMed: 21911355]

Highlights

- We have developed CoRAL, a method for classifying small non-coding RNAs using high-throughput sequencing of small RNAs (small RNA-seq)
- Post-transcriptionally generated non-coding RNAs can be classified using features that are relevant to small RNA biogenesis pathways such as fragment lengths, strand bias, endpoint specificity
- In addition to classifying unknown small ncRNAs, our method can provide insight into which features are specific to which types of RNA
- CoRAL is a method for classifying small non-coding RNAs using small RNA-seq data
- CoRAL uses features associated with small RNA biogenesis such as fragment length
- These features are informative for classifying post-transcriptionally generated RNAs
- Our method can provide insight into which features are specific to which RNA types

Fig 1.

The upstream processing of small RNA-seq datasets prior to analysis by CoRAL: trimming of 3' adapter sequences and mapping to the genome.

Fig. 2.

A flow diagram depicting CoRAL, from processing the smRNA-seq reads to building a model of the RNA classes and using that model to predict the classes of loci in a completely separate RNA-seq dataset.

a)

b)

c)

Fig.3.

UCSC genome browser plots of smRNA-seq read coverage for a) a microRNA, b) a C/D box snoRNA, and c) a transposable element-derived small RNA locus

a)

Fig. 4.

Plots of read coverage onto the secondary structures of three small RNA loci: a) a microRNA locus (RNAfold-based structure), b) a C/D box snoRNA locus (RFAM-based structure), and c) a transposable element-derived RNA locus (RNAfold-based structure).

Fig.5.

CVA (canonical variates analysis) projection of the various features used in a 3-class classification of miRNAs, C/D box snoRNAs, and transposable-element-derived RNAs. The axes are labeled on their positive sides.

Table 1

Example output from CoRAL's class predictions:

chr	start	end	name	reads	strand	label
chr10	328025	328066	SL0000000001	34	-	scRNA
chr10	687636	687710	SL0000000002	102	-	miRNA
chr10	1094791	1094809	SL0000000003	15	-	transposon
chr10	5895678	5895747	SL0000000004	173	-	snoRNA_CD