Prediction and characterization of non-coding RNAs in *C. elegans* by integrating conservation, secondary structure and high throughput sequencing and array data

Zhi John Lu[1,2,10], Kevin Y. Yip[1,2,3,10], Guilin Wang[4], Chong Shou[1], LaDeana W. Hillier[5], Ekta Khurana[1,2], Ashish Agarwal[2,6], Raymond Auerbach[1], Joel Rozowsky[1,2], Chao Cheng[1,2], Masaomi Kato[7], David M. Miller[8], Frank Slack[7], Michael Snyder[9], Robert H. Waterston[5], Valerie Reinke[4] and Mark Gerstein[1,2,6]

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, USA

[2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, USA

[3]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong

[4]Department of Genetics, Yale University, New Haven, Connecticut, USA

[5]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington, USA

[6]Department of Computer Science, Yale University, New Haven, Connecticut, USA

[7]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut, USA

[8]Department of Cell and Developmental Biology, Vanderbilt University, Nashville, TN, USA

[9]Departments of Developmental Biology and Genetics, Stanford University Medical Center, Stanford, CA, USA

[10]These authors contributed equally to this work.

Correspondence to:
Mark B. Gerstein[1,2,6]
MB&B, 260/266 Whitney Avenue, PO Box 208114, New Haven, CT 06520-8114
e-mail: mark.gerstein@yale.edu

Running Title: Characterizing ncRNAs using an integrative approach

Keywords: modENCODE, non-coding RNA, machine learning, RNA secondary structure, high throughput sequencing and tiling array

Abstract

We present an integrative machine learning method, *incRNA*, for whole-genome identification of non-coding RNAs (ncRNAs). It combines a large amount of expression data, RNA secondary-structure stability, and evolutionary conservation at the protein and nucleic-acid level. Using the *incRNA* model and data from the modENCODE consortium, we are able to separate known *C. elegans* ncRNAs from coding sequences and other genomic elements with a high level of accuracy (97% AUC on an independent validation set), and find >7,000 novel ncRNA candidates, among which >1,000 are located in the intergenic regions of *C. elegans* genome. Based on the validation set, we estimate that 91% of the ~7000 novel ncRNA candidates are true positives. We then analyze fifteen novel ncRNA candidates by RT-PCR, detecting the expression for fourteen. In addition, we characterize the properties of all the novel ncRNA candidates and find that they have distinct expression patterns across developmental stages and tend to use novel RNA structural families. We also find that they are often targeted by specific transcription factors (~59% of intergenic novel ncRNA candidates). Overall, our study identifies many new potential ncRNAs in *C. elegans* and provides a method that can be adapted to other organisms.

Introduction

The massive amounts of data from tiling arrays and high-throughput sequencing (Margulies et al. 2005; Shendure et al. 2005; Fejes-Toth et al. 2009) have driven the discovery of novel transcripts. Many of these transcripts are functional without being translated into proteins, and hence are called non-coding RNAs (ncRNAs). ncRNAs include many well-known RNA types such as rRNA, tRNA and snoRNA, as well as small RNAs such as miRNA, siRNA, and piRNA. They also refer to more recently discovered RNA types, such as promoter-associated short RNAs (PASRs) (Fejes-Toth et al. 2009), whose function has not been well studied. Many small ncRNAs, including miRNAs and siRNAs, contribute to the complexity of regulatory networks in eukaryotes.

Before large-scale experimental data became available, genome-wide identification of ncRNAs had relied on computational approaches. For certain types of ncRNAs, specific databases and prediction methods are available, such as tRNA-SE and GtRNAdb for tRNA (Lowe and Eddy 1997; Chan and Lowe 2009), snoscan and snoRNABase for snoRNA (Schattner et al. 2005; Lestrade and Weber 2006) and miRBase for miRNA (Griffiths-Jones et al. 2008). For the more general task of identifying all ncRNAs from a genome, one common approach is based on comparative genomic analysis. For example, the methods QRNA, DDBRNA and MSARI make use of the conservation of RNA secondary structures in identifying ncRNAs (Rivas and Eddy 2001; di Bernardo et al. 2003; Coventry et al. 2004). Newly developed methods using this approach have been applied to human (Pedersen et al. 2006; Washietl et al. 2007) and *C. elegans* (Missal et al. 2006). A large amount of conserved secondary structures, identified from multiple genomes by the so-called Infernal method, are organized into structural families in Rfam (Gardner et al. 2009), which can be used to search for their structural homologues in other genomes using the Infernal software package (Nawrocki et al. 2009).

It has been shown that different types of data provide different kinds of information to the ncRNA identification process, and computational methods could perform better by integrating multiple data types. For example, when secondary structure stability and conservation information are combined using machine learning techniques, integrative methods such as RNAz and Dynalign/SVM demonstrate improved accuracy in identifying ncRNAs in various prokaryotic and eukaryotic genomes (Washietl et al. 2005; Torarinsson et al. 2006; Uzilov et al. 2006; Washietl et al. 2007; Torarinsson et al. 2008).

With the advent of high-throughput technologies, it is now also possible to experimentally survey novel transcriptomes to find ncRNAs (He et al. 2007). Other genome-wide experimental data could also potentially provide additional information about ncRNAs. For example, in recent studies, chromatin signatures have been used to identify large conserved ncRNAs (lincRNAs) in mammals (Guttman et al. 2009) and profiles of small RNA reads have been used to find novel ncRNAs in the *Drosophila* genome (Jung et al. 2010).

In this paper, we describe a comprehensive model, *incRNA* (*in*tegrated *ncRNA* finder), which integrates sequence, structure and expression data. We gathered large-scale expression datasets from the modENCODE consortium (Celniker et al. 2009; Gerstein et al. 2010), which were obtained not only by using both tiling arrays and deep sequencing, but also from different tissues and developmental stages of *C. elegans.* In addition, our model includes a carefully selected set of other features useful for separating ncRNAs from other genomic elements. We demonstrate that our integrative method significantly improves the accuracy of ncRNA identification as compared to some previous methods. Importantly, we show that no single feature could achieve the top performance of the features when combined, which strongly suggests their complementary nature and the advantage of integrative approaches.

We describe how we used *incRNA* to predict 7,237 novel ncRNA candidates (1,678 of them are intergenic ncRNA candidates) and experimentally validated the expression of a random sample. We also characterize the novel ncRNA candidates by their genomic locations, structural properties, expression patterns, binding sites of POL II and 22 transcription factors. Finally, we summarize these results in a master table to facilitate future validation and functional characterization of the candidates.

Results

Single features and feature pairs only partially separate different types of annotated genomic elements

We gathered a large compendium of sequence, structure and expression features of *C. elegans* that could potentially identify ncRNAs. The expression features were expression signals produced by small RNA-seq, poly-A+ RNA-seq, total RNA tiling arrays and poly-A+ tiling arrays in different developmental stages and conditions. The sequence and structure features include GC content, genome-wide predictions of DNA and protein conservation as well as secondary structure stability and conservation. We expected that in general, functional ncRNAs would have some expression signals and stronger GC content and DNA conservation. We also expected ncRNAs to be distinguishable from coding transcripts by having, on average, stronger small RNA-seq signals, lower poly-A+ signals, and stronger secondary structures but lower levels of predicted protein-level conservation.

Since functional ncRNAs are likely to be conserved, and RNA secondary structure prediction is more reliable in conserved regions, we only considered regions with a high conservation score (15% of *C. elegans* genome) from the pairwise alignment between *C. elegans* and *C. briggsae*. 219 known ncRNAs from Wormbase were covered by the conserved regions, including long (rRNA), small (miRNA) and medium-sized (tRNA, snoRNA, etc) ncRNAs. In order to predict local RNA secondary structure, we divided the pair-wise alignment into small bins (each bin has 150 aligned columns, see detail in Methods). Altogether we generated 439,815 bins, which cover 29,655,415 bases of the plus and minus strands. We then annotated each bin as

CDS, UTR, ncRNA or intergenic if it overlapped with the corresponding annotations from Wormbase. These annotated bins were used to construct a gold-standard set (see definition in Supplementary Methods and numbers in Supplementary Figure 3 upper panel) for further analysis. The local expression, structural and conservation properties of each genomic element (e.g. ncRNA or CDS) can be reflected by the bins within it.

We found that different features were useful in identifying different classes of genomic elements (Figure 1a). For example, known ncRNAs have, in general, higher signals from small RNA-seq experiments than the other three classes of gnomic elements. However, there are also ncRNAs with very low signals, as well as CDSs with signals even stronger than those of most ncRNAs. In fact, no single feature could distinguish all known ncRNAs from other genomic elements (Figure 1a). This classification and separation was improved by considering pairs of features (Figure 1b and Figure 1c), yet in all cases, some ncRNAs are still indistinguishable from the other classes.

ncRNAs are clearly separated from other genomic elements by machine learning methods using the integrated features

Since the different features capture different kinds of information about genomic elements, we perceived that our ability to identify ncRNAs would be maximized if we integrated all the features. A systematic way to perform this integration is to employ machine learning methods to model the subtle interactions between the features in annotated data. The learned statistical models can then be used to identify novel ncRNA candidates from the un-annotated genomic regions.

We implemented a machine learning module for this purpose and integrated it into the pipeline of *incRNA* with additional modules for preprocessing data and characterizing our predictions (Figure 2a). The annotated bins from the gold-standard set were used to train machine learning models and predict ncRNAs from the unannotated bins. We took the bins predicted with higher confidence and merged them with adjacent bins that could belong to the same ncRNAs into genomic regions we call "candidate ncRNA fragments" We then characterized them using various types of information. The details will be given in the coming sections.

To ensure an unbiased evaluation of the effectiveness of our machine learning model, we used a rigorous procedure that involves a cross-validation step for choosing the best model from those produced by a chosen set of learning methods, and a final evaluation step using an independent validation set not involved in model learning (Figure 2b, see Supplementary Methods). The procedure recommended Random Forest as the best learning method. When instructed to separate ncRNAs from the other genomic elements in the independent evaluation set, Random Forest showed high accuracy both in absolute terms and in comparison with Rfam/Infernal (Gardner et al. 2009) and RNAz (Washietl et al. 2005), two methods that identify sequences with potential RNA secondary structures (Figure 3a, Supplementary Figure 1b). The area under the receiver-operator characteristic curve (AUC) for our predictions was 97%, which indicates that our method was able to identify most of the ncRNAs in the validation set before making any false predictions. We have also separated the ncRNA examples in the independent validation set into four subsets according to their sequence identity and GC content values (lower/higher than median for the two features). The resulting AUC values of the best case (high identity, high GC) and the worst case (high identity, low GC) differ by only 0.01, which shows the robustness of our method (Supplementary Table 1b).

We noticed that while Rfam/Infernal could very accurately identify about 80% of known ncRNAs in the gold-standard set, it was unable to distinguish the remaining 20% from the other genomic elements (Figure 3a), which suggests potential undiscovered ncRNA structural families. The comparison with Rfam/Infernal and RNAz also suggests that information about RNA secondary structure alone is not sufficient to separate all ncRNAs from other genomic elements. This is possibly related to a previous finding that local RNA structures can also occur in coding regions (Katz and Burge 2003). We also need to note that Rfam/Infernal and RNAz could also pick up structured regulatory elements in UTRs, since these two programs aim to detect structured RNA sequences instead of ncRNA genes only. The comparison has some caveats because the specificities of Rfam/Infernal and RNAz will not be high by definition (UTRs containing no known ncRNA genes are negatives in our training set).

To ensure the robustness of our analysis, we repeated the machine learning procedure using various definitions of negative sets and various combinations of genomic element types (see Supplementary Methods). Our prediction accuracy for ncRNAs was very stable in the different cases (Supplementary Figure 1a and Table 1a). Interestingly, we observe that some ncRNA types form distinct clusters in the feature space (Supplementary Figure 2). For example, rRNAs have a unique signature of high conservation and high expression values in both poly-A+ and small RNA-seq experiments; miRNAs have high small RNA-seq signals but relatively low poly-A+ RNA-seq signals; snoRNAs have small ranges of medium expression values in both types of RNA-seq experiments; and tRNAs form several distinct clusters, with different RNA-seq signals and levels of conservation. Integrating different types of data could thus not only distinguish ncRNAs from other classes of genomic elements, but potentially also differentiate between different ncRNA types. Prediction at this granularity is currently limited by the small number of examples of certain types of ncRNAs.

Top prediction performance requires the integration of all features

We next studied the relative importance of the different features by checking the resulting accuracy of our machine learning procedures when we used only a subset of features. We found that the expression features were better at identifying ncRNAs than structural and sequence-based information (Figure 3b), yet combining both types of features gives an additional 5-10% precision at any given false positive rate. It is thus useful to include both types of data in identifying ncRNAs.

Since the cost of producing experimental data is high, it is interesting to see if it is sufficient to only include the expression data from a single type of experiment. We found that no single type of expression data could completely substitute for the others (Figure 3c), suggesting that each type of experiment is able to identify some unique ncRNAs.

We also used the weights of the different features in the learned Logistic Regression model as a second indicator of the importance of the features (Table 1). Consistent with the above analysis, we find that while small RNA-seq and RNA secondary structure conservation have the heaviest absolute weights in the ncRNA class, some other features, such as poly-A+ RNA-seq and tiling array data, are also heavily weighted in other classes, indicating their different roles in separating ncRNAs from other genomic element classes.

Using incRNA to predict novel ncRNA candidates

Having verified the predictive power of *incRNA* on the gold-standard set, we then extended our predictions to all the bins. For each (annotated and unannotated) bin, our model gives an “ncRNA score” that indicates the likelihood that the bin lies in an ncRNA gene. It also gives a CDS score, a UTR score, and an intergenic region score in similar ways. From the bins in the gold-standard set, we find that our model separated known ncRNAs from the other bins by a large score margin (Figure 3d). These results stand in sharp contrast to the unsatisfactory differentiations made by individual features and feature pairs (Figure 1), and provide more proof of the robustness of our method.

In particular, we found that all known ncRNA bins had predicted ncRNA scores of at least 0.69, while all the other elements had scores of 0.18 at most (Figure 3d). We called these values P_{high} and P_{low}, respectively, and used them as thresholds for defining novel ncRNA candidates. Specifically, we defined each unannotated bin with an ncRNA score of P_{high} or higher as a high-confidence candidate ncRNA bin and each unannotated bin with a ncRNA score between P_{low} and P_{high} as a medium-confidence candidate ncRNA bin (Figure 3e, see also Figure 2a). Altogether, the two sets contain 10,994 bins (covering 1,045,795 bases in total), among which 1,413 are high-confidence predictions and 9,581 are medium-confidence predictions.

To estimate the accuracy of these candidate ncRNA bins, we repeated the above procedure of defining the two thresholds using only the cross-validation set, and then examined the ncNRA scores of the bins in the independent validation set. We found that all the bins with ncRNA scores higher than P_{high}, and 221 out of 242 bins with ncRNA scores higher than P_{low}, were known ncRNA bins, corresponding to positive predictive values (PPVs) of 100% and 91%, respectively. On the other hand, since there are a total of 247 known ncRNA bins in the validation set, the two thresholds lead to prediction sensitivities of 66% and 89%, respectively.

The two sets of candidate ncRNA bins thus serve different purposes. The high-confidence set is expected to have few false positives, while the medium-confidence set provides additional coverage so that the two sets together have relatively few false negatives. We expect some of the medium-confidence candidates to be novel ncRNAs with properties different from those of known ncRNAs.

We have examined the 9% of the bins included in the medium-confidence set but are not annotated as ncRNAs (the false positives), and the 11% ncRNAs not included in the medium-confidence set (the false negatives). Among the false positives, 58% are CDSs and 42% are UTRs. They are characterized by particularly weak tiling array and poly-A+ RNA-seq signals.

They false negatives are characterized by particularly strong poly-A+ RNA-seq signals, weak predicted structures, and low sequence identities.

For bins with ncRNA scores below $P_{low,}$ we further divide them into two sets. The bins with very low ncRNA, CDS and UTR scores are grouped into a predicted set of unexpressed intergenic regions. The remaining bins form a low confidence set of ambiguous regions, which could be CDS or UTR (see Supplementary Methods and Figure 2a). We summarize the number of bins in each set in Supplementary Figure 3.

Since different predicted bins may come from the same ncRNA genes, we merged adjacent bins into longer regions. We call them candidate ncRNA fragments, as they represent candidate regions of novel ncRNA but may not cover the full-length genes. Altogether we produced 7,237 such fragments from the high-confidence and medium-confidence sets of candidate ncRNA bins. The length distribution of these fragments is comparable to that of full-length known ncRNA transcripts (Supplementary Figure 4).

Many candidate ncRNA fragments are inside the introns of coding genes, at the antisense strand of exons or close to coding sequences (see Supplementary Methods). Among the 7,237 candidate ncRNA fragments, 1,678 of them (merged from 2,469 bins) are located in strictly-defined intergenic regions based on our gold-standard annotations (see Supplementary Methods), which occupy 1,223 distinct genomic locations, not considering the strand information. These intergenic candidates are most likely to be novel ncRNAs, as their expression is unlikely to be due to nearby coding genes. On the other hand, the other candidates could also be authentic ncRNAs, as many ncRNAs have been found inside introns (Bartel 2004; Li et al. 2007) and antisense to exons (Mercer et al. 2008).

About 20% of the candidate ncRNA fragments are overlapped with repeats or inverted repeat regions. It is well known that many pi-RNAs are transcribed from transposons and other repeated sequence elements (Klattenhoff and Theurkauf 2008). Recent analysis has also identified telomeric repeat-containing RNA (TERRA) in animals and fungi (Luke and Lingner 2009). We therefore decided to keep these predictions in our set.

We overlapped our intergenic candidate ncRNA fragments with the novel genelets generated by the modENCODE consortium (Gerstein et al. 2010), where all the splice junctions and all the splice leader sites were incorporated in order to map spliced poly-A+ RNA-seq reads. A small fraction of our intergenic candidate ncRNA fragments (44 of the 1,678) are overlapped with the genelets. This is expected because most of our novel ncRNA candidates are probably not having poly-A tails. Further experiments are needed to confirm if they are unannotated coding exons or novel ncRNAs.

Experimental validation of predicted novel ncRNA candidates

In additional to computational validation of our predictions, using an independent set of annotated regions not involved in model learning, we also validated a random sample of our candidate ncRNA fragments located inside intergenic regions by means of RT-PCR experiments. To get a better sense of the lengths of these potential non-coding transcripts, we overlapped our candidate ncRNA fragments with the transcriptionally active regions (TARs) obtained from the tiling array datasets of the modENCODE consortium (Gerstein et al. 2010). We call the resulting overlapped regions the candidate ncRNA TARs. As the TARs were defined from expressed regions of at least 100nt long, many novel ncRNA candidates from short transcripts are not covered by our TARs. Only 730 of the longer fragments among our 1,678 intergenic candidate ncRNA fragments overlap with the TARs in at least one of the 41 tiling array datasets. Among them, 54 overlap in late embryos. We then picked 15 candidate ncRNA TARs (Supplementary Table 2) for RT-PCR validation using total RNA from N2 late embryos (see Methods). We detected clear bands on 2% agarose gel for 14 of the 15 candidate ncRNA TARs, and of those, 13 had significantly greater amplification than the control reactions to which no reverse transcriptase was added (Figure 4a). All the RT-PCR products were confirmed by direct DNA sequencing (see Methods). The results confirmed that the expression of our predictions (especially repeat-related ncRNA candidates 1 and 3) is not likely, due to cross-hybridization in tiling arrays.

We further estimated the expression and size of five out of 15 candidate ncRNA TARs (Supplementary Table 2) using the Northern blot analysis. Because the high throughput data we used are very sensitive, our model is able to predict many lowly expressed ncRNA candidates (Supplementary File 1). The Northern blot method is less sensitive than the RT-PCR analysis, and can only detect three out of five candidates (Supplementary Figure 5a). We estimated from the Northern blot that two of them is larger than 500 bases. We also checked their genome location and did not find any annotated genes close to them (within ~2 Kb). This proves that our model has the potential to identify novel long ncRNAs. Because of the complicated process in ncRNA (e.g. miRNA precursor), it is hard to estimate the real length of the ncRNA precursor or mature product using tiling array TARs. This is one of the reasons for the low sensitivity of the Northern blot, since the probe we selected may miss the highly expressed part (e.g. mature sequence part of miRNA).

Furthermore, we found that many novel ncRNA candidates are supported by multiple information sources. An example is shown in Figure 4b for one candidate (lane 7 in RT-PCR gel), which is located inside a transcribed region detected by the tiling array in late embryo. No significant expression was detected from Poly-A+ RNA sequencing or small RNA sequencing, indicating that it is unlikely to be a coding gene or a small interfering RNA. From ChIP-seq data, we observe a POL II binding peak immediately upstream of ncRNA (lane 7) and strong binding signals of PHA-4 across the region. These binding signals suggest potential regulation of ncRNA7 by POL II and PHA-4. Since PHA-4 is a key factor that regulates the development of the pharynx/foregut during embryogenesis, this candidate ncRNA could potentially play a role in development during the embryonic stage. Two more examples are shown in Supplementary Figure 5b,c.

Characterizing novel ncRNA candidates with RNA secondary structure and DNA conservation

We next studied various properties of the novel ncRNA candidates. We first examined the local structural and expression properties using the candidate ncRNA bins and then analyzed POL II and transcription factor binding signals using the candidate ncRNA fragments (merged bins), as the positional relationship between binding sites and ncRNAs are better captured by the fragments. All these results are presented in the coming sections.

As known ncRNAs generally have stable secondary structures (Figure 1), we first calculated the potential consensus free energy for both the known ncRNA bins and candidate ncRNA bins with the Dynalign program (Harmanci et al. 2007). We predicted about 2/3 of the known ncRNA bins and high-confidence candidate ncRNA bins to be highly structured (Figure 5). The fraction is lower for the medium-confidence bins (~1/4), yet it is still much higher than the bins predicted to be unexpressed (13%) or ambiguous (6.5%).

We checked if the highly structured novel ncRNA candidates have secondary structures similar to the known ncRNAs. When we used Infernal to predict structural homologues of known families in the *C. elegans* genome cataloged in Rfam (Gardner et al. 2009), 84% of the highly structured known ncRNA bins (56% of 67.7%) were found to overlap with the Rfam/Infernal predictions (Figure 5), which is consistent with our earlier observation that Rfam/Infernal could only predict approximately 80% of the known ncRNAs before producing substantial false positives (Figure 3a). In contrast, about 80% of the candidate ncRNA bins we predicted to be highly structured did not overlap with the Rfam/Infernal predictions (Figure 5), which suggests potential novel structural families of ncRNAs. The enrichment of novel structures is also observed for the subset of novel ncRNA candidates inside intergenic regions (Supplementary Figure 6).

We also compared the characteristics of known ncRNA and novel ncRNA candidates using other genomic features (Supplementary Figure 7 and Supplementary Table 3). The high-confidence candidate ncRNA bins are found to share many common properties with known ncRNA bins, including high conservation between *C. elegans* and *C. briggsae* at DNA and RNA level. The conservation of the medium-confidence candidate ncRNA bins is relatively low (Supplementary Figure 7).

As conservation is an indicator of potential function of ncRNAs, we also used a multiple sequence alignment between five nematode species to double-check the conservation of our predictions. We found 94% of our candidate ncRNA bins overlap the conserved aligned regions (see Supplementary Methods), which provides extra support for these predictions.

Characterizing novel ncRNA candidates with expression patterns

Next, we compared the expression profiles of the known ncRNA transcripts and novel ncRNA candidates. In particular, we analyzed whether they exhibit developmental stage-specific expression patterns. Since our small RNA-seq data cover the largest number of developmental stages (see Supplementary Methods), in this part of analysis we focus on this type of expression data. Also, since miRNAs form a major class of small RNAs among all known ncRNAs, we separate them from other known ncRNAs to determine whether they have distinct expression patterns as reported by the small RNA-seq experiments.

By clustering the bins according to their expression patterns across the developmental stages, we observe that there are three sub-classes of ncRNAs (Figure 6a):

[1] ncRNAs that are universally expressed in all stages. This subclass contains 58% of the known ncRNAs, with a slight enrichment of miRNAs. Some of the ncRNAs in this subclass exhibit fluctuations of expression levels across stages (Figure 6a, right panel).

[2] ncRNAs that are expressed only in some stages. This subclass contains only 2% of the known ncRNAs, with a significant enrichment of novel ncRNA candidates (p-value from Fisher's exact test less than 1e-15).

[3] ncRNAs with no detectable expression by small RNA-seq in all stages.

The numbers of high-confidence and medium-confidence candidate ncRNA bins in the three sub-classes are shown in Supplementary Table 4, with the numbers of candidates inside intergenic regions shown separately in the table. In general, the expression levels of novel ncRNA candidates are lower than the universally expressed known ncRNAs, such as rRNAs and tRNAs. The finding that most differentially expressed ncRNAs come from the novel candidates is intriguing. It could suggest that the novel ncRNA candidates play more specialized roles in specific stages.

To gain more insight into how differential expression affects the identification of ncRNAs, we identified the bins with detectable expressions in each combination of developmental stages (see Supplementary Methods). Then for each combination of stages (e.g. L2 + L4 + Young Adult) we computed the fractions of expressed known miRNAs, other known ncRNAs, and candidate ncRNA bins. Finally, we grouped all combinations with the same number of stages (e.g. both L2 + L3 and L4 + Young Adult have two stages) to form a distribution of expressed fractions. The saturation plot is shown in Figure 6b.

We observe that the amount of novel ncRNA candidates with detectable expressions keeps increasing when small RNA-seq datasets of more stages are added, suggesting that for each stage there is a set of novel ncRNA candidates that express only in that stage (Figure 6b). The same trend is observed when we consider all our expression datasets (Supplementary Figure 8). In contrast, for known ncRNAs, saturation is reached when only a few stages are considered, regardless of the exact combination of the stages (Figure 6b).

We notice that in the embryonic stage there is a much larger fraction of expressed novel ncRNA candidates (Figure 6b). This observation is consistent with a finding in a previous study (Kato et al. 2009) showing that 65.6% of small RNA sequencing reads in embryos could not be mapped to known miRNAs or 21U-RNAs, whereas the percentages of unmapped reads in other hermaphrodite stages were only 27-38%. Additional experiments in embryo replicas are needed to provide further confirmation.

Characterizing intergenic novel ncRNA candidates with binding sites of POL II and different transcription factors

As a first step to understanding the underlying mechanisms that caused the observed expression patterns, we studied the potential regulation of the novel ncRNA candidates by POL II and transcription factors. We obtained the binding signals of POL II and 22 transcription factors from ChIP-seq experiments produced by the modENCODE consortium (Gerstein et al. 2010). We looked for POL II binding signals near the starting site (±150 base pairs) of the intergenic candidate ncRNA fragments (see Methods), and found that about 15% of them had significant POL II binding signals in at least one of the seven developmental stages (Supplementary Figure 9). Using previous results from modENCODE consortium we also located the binding sites of 22 transcription factors, with a total of 27 experiments across different developmental stages (see Methods). Many of these binding sites are located at potential promoter regions of ~59% of intergenic candidate ncRNA fragments (Figure 6c and Supplementary Figure 9). Compared to random genome locations with the same size, the binding of TFs has 2-fold enrichment for the intergenic candidate ncRNA fragments (p value < 0.001, Figure 6d). However, the enrichment or depletion of Pol II binding is not significant (p value > 0.05, Supplementary Figure 9). This is consistent with the known ncRNAs, where only certain types of ncRNAs (e.g. miRNA) are transcribed by Pol II. The binding sites of Pol II as found from ChIP-seq data may still be true, although we did not find any enrichment.

Master table of novel ncRNA candidates

We summarize our whole list of novel ncRNA candidates (Supplementary Table 5) predicted from *incRNA* in a master table (Supplementary File 1), and the subset in intergenic regions in another table (Supplementary File 2). Each candidate is associated with its prediction scores, predicted class, feature values, genomic coordinate and location (such as inside intron or intergenic region), structural properties, expression pattern class, binding signals of POL II and transcription factors, and the conservation score among the five nematodes. In order to facilitate follow-up studies, we have ranked our predictions into 9 levels based on these associated pieces of evidence (See Supplementary Methods). The predictions in higher levels are more likely to be functional ncRNAs.

All datasets, prediction results, and the prediction software can be found at our supplementary web site: http://incrna.gersteinlab.org/ .

Discussion

We have shown that none of the individual expression, sequence and structural features is able to clearly separate known ncRNAs from protein coding sequences and UTRs. Instead, by integrating all the features using a machine learning framework, *incRNA*, we were able to identify ncRNAs with high accuracy. The learned model is robust, with prediction accuracy above 97% AUC, regardless of the expression thresholds for defining the unexpressed intergenic regions (Supplementary Figure 1a). The massive experimental data from the modENCODE consortium (Gerstein et al. 2010) contributed significantly to our model.

We have validated our predictions by multiple means. The validation RT-PCR experiments have confirmed that the expression levels of our novel ncRNA candidates are not likely due to the cross-hybridization of the tiling array or a mapping error in RNA-seq. The sequences of the RT-PCR products were verified by direct DNA sequencing, and sizes of the candidate ncRNA transcripts were estimated by Northern blots. Most of our candidate ncRNA bins (94%) were found to be conserved based on a multiple sequence alignment of five nematodes that was not used in model learning. We have also used cross-validation and an independent validation set of known ncRNAs to show that our models are able to identify known ncRNAs that were not included in the model training process. To further identify the subset of our predictions corresponding to functional non-coding genes, more biological functional assays need to be performed in addition to these computational and biochemical validations.

The accuracy of RNA secondary structure predictions (percentage of known base pairs that are correctly predicted) using Dynalign (Harmanci et al. 2007) and RNAz (Washietl et al. 2005) depends heavily on the quality of the genome alignment. Research has shown that the sensitivity of ncRNA sequence identification gets worse as the DNA sequence identity drops below 60% in the alignment (Gorodkin et al. 2009). Among our aligned bins, 26.5% have a sequence identity of less than 60%. However, we still successfully identified most ncRNAs

(Figure 3), because other features involved in the machine learning models are not directly affected by the degree of conservation. We also remark that methods that use local alignments, such as FOLDALIGN (Havgaard et al. 2007), have the potential of making better structure predictions for short RNAs, which may in turn increase the relative contribution of the structural features in the overall ncRNA identification process.

In the *C. elegans* genome, only a small portion (13-15%) could be aligned with *C. briggsae* (Missal et al. 2006) at a quality level sufficient for Dynalign or RNAz to make meaningful predictions. The small portion of aligned genome illustrates a tradeoff between precision and coverage, that although the information about secondary structures and conservation could improve prediction accuracy (Figure 3b), it also excluded a lot of genomic regions that could potentially contain novel ncRNA candidates. In this study our goal is to identify ncRNA candidates with high confidence, and thus we decided to focus on the highly conserved regions with a 73% median DNA identity. For studies that aim at high coverage, one possible alternative approach is to use weakly aligned regions from the multiple alignment of MultiZ, which provides a better coverage (~25% of the genome) (Kiontke and Fitch 2005), with the tradeoff of a lower conservation (66% median DNA identity). This approach would generate a more permissive set of ncRNA candidates.

One direct consequence of using a genome alignment is the possibility of having a novel ncRNA only partially covered by the aligned regions. We expect the potential ncRNA fragments we constructed by merging candidate ncRNA bins to be shorter than the full-length ncRNA transcripts. Nevertheless, these fragments form a detailed map of genomic regions that could be used to locate novel ncRNAs at a more refined level.

A significant portion of our novel ncRNA candidates are overlapped with repeats and introns, or are antisense to exons. This is partially due to the compactness of the *C. elegans* genome. While many interesting ncRNA candidates have been previously found in these kinds of regions, caution should be taken when studying them, as the expression levels of intronic or antisense ncRNA candidates could have been affected by the enclosing genes. In order to facilitate follow-up studies that require a more conservative set of predictions with high precision, we have defined a set of intergenic novel ncRNA candidates, which are far away from annotated coding sequences (see Supplementary Methods).

The tradeoff between precision and coverage is also illustrated by our choice of unannotated regions. In order to minimize the number of false positives in our predicted set of intergenic ncRNAs, we used a permissive set of genome annotations, such that bins that overlap with any confirmed, unconfirmed and predicted genes from multiple models were all excluded from our final set of predictions.

Since our machine learning models were trained on known types of ncRNAs, they tend to predict novel ncRNA candidates that exhibit properties similar to these types of annotated ncRNAs. Expressed regions with more distinct properties are likely predicted as CDS-like or UTR-like ambiguous regions. For instance, since the annotated ncRNAs have low poly-A+ RNA signals in general (Figure 1), this feature was used as a negative indicator of ncRNAs, as reflected by its negative weight in the Logistic Regression model (Table 1). As a result, our novel ncRNA candidates tend to have weak poly-A+ RNA signals. Also, while the majority of known ncRNAs are small- or medium-sized, we also included some non-mRNA-like long ncRNAs (i.e. rRNAs) in our training set. In our predictions, there are indeed fragments of long candidate ncRNA transcripts (>500 nt) (Northern blot result in Supplementary Figure 5a).

Methods

Machine learning methods

We combined the sequence, structure and expression datasets and generated nine genomic features for training the machine models and making the predictions. Among the nine genomic features, four of them are expression features, corresponding to the maximum signals of 1) the six poly-A+ RNA-seq experiments; 2) the 11 small RNA-seq experiments; 3) the 29 total RNA tiling arrays; and 4) the 12 poly-A+ RNA tiling arrays. Three of the features are related to sequence information, including GC%, DNA conservation and predicted protein sequence conservation. The remaining two features are related to RNA structures, namely predicted secondary structure free energy and predicted secondary structure conservation. The details of the nine gnomic features and machine learning method are described in Supplementary Materials.

RT-PCR and Northern blot confirmation

We used a three-step process to pick candidate ncRNAs for validation. First, we overlapped candidate ncRNA bins with the tiling array TARs (transcriptionally active regions) defined and optimized by the modENCODE consortium (Agarwal et al. 2010; Gerstein et al. 2010). We then discarded TARs that overlap with any exonic or intronic regions using a permissive set of annotation from Wormbase, and those predicted as ncRNAs by Rfam. Finally, we retained only TARs with a positive small RNA-seq read count or with a log2-transformed, normalized total RNA tiling array signal larger than 8 at late embryo. Subsequently, fifteen candidates (Supplementary Table 3) were randomly picked from the remaining set for validation.

Total RNA was isolated from N2 late embryos and treated with Dnase I from Ambion. Reverse transcription was then performed using random hexamers following the Omniscript RT kit instruction from Qiagen. Subsequently, primers were designed within the candidate regions and amplified by PCR. The RT-PCR products were purified using the PCR purification kit (Qiagen Inc.). The purified products (10-20 ng DNA) were then sent to the W.M. Keck Facility for direct sequencing. The details of sequences, PCR primers and products are provided in Supplementary Materials.

We then manually picked five ncRNA candidates (out of fifteen candidates validated above) with clear signals on the PCR gel to undergo a Northern blot assay. The total RNA were extracted from late embryos and hybridized with labeled probes (see detail in Supplementary Materials).

Scoring novel ncRNA candidates with POL II and transcription factor signals

POL II and transcription factor binding data from ChIP-seq experiments were scored using the method from the modENCODE consortium (Gerstein et al. 2010) with a default PeakSeq q-value cut-off (Rozowsky et al. 2009) of 0.001. We defined the sets of candidate ncRNA fragments potentially targeted by POL II and the various transcription factors using a previous target calling method (Zhong et al. 2010). Random locations from the genome with the same size as candidate ncRNA fragments were tested using the same binding peaks. We repeated the random process twenty times and the average chances of being targeted were plotted in Figure 6c and Supplementary Figure 9.

Acknowledgements

We would like to thank David H. Mathews and Shan Zhao at University of Rochester for providing the shuffled RNA sequence free energy calculation method. We would like to thank the Yale University Biomedical High Performance Computing Center for its support, which is funded by NIH grant RR19895. We especially thank Nicholas Carriero from the Computing Center for helping us to map small RNA sequencing reads.

Figure Legends

Figure 1. Distributions of nine genomic feature values. The distributions of values of the nine features are shown for the gold-standard set (see Supplementary Methods for the definition of the gold-standard set) of the four types of genomic elements: known ncRNAs, CDSs (Coding Sequences), UTRs and intergenic regions. The values of each expression feature are the maximum of the corresponding values from all the expression datasets of the same type. **(a)** Box plots of individual features (normalized values). **(b)** Two-dimensional scatter-plot of the maximum small RNA-seq signal against the maximum poly-A+ RNA-seq signal. **(c)** Two-dimensional scatter-plot of the maximum poly-A+ RNA tiling array signal against the predicted secondary structure conservation. Expression values in **(b)** and **(c)** are the log-transformed normalized read counts (DCPM: Depth of Coverage Per Million reads).

Figure 2. A flowchart of *incRNA* (*in*tegrated *ncRNA* finder) for predicting and characterizing novel ncRNA candidates in *C. elegans.* **(a)** We looked for ncRNAs from conserved regions from the genome alignment between *C. elegans* and *C. briggsae*, and divided them into small bins. Annotated bins from the gold-standard set were used to build a machine learning model based on nine expression, sequence and structural features. The model was then used to score each unannotated bin by its likelihood of belonging to four genomic element classes (ncRNA, CDS, UTR and unexpressed intergenic region). Adjacent bins predicted to be novel ncRNA candidates with high or medium confidence were merged into candidate ncRNA fragments, which were further characterized by their predicted RNA secondary structures, expression patterns, and the binding signals of POL II and different transcription factors. **(b)** We used an unbiased procedure to build and evaluate our machine model. Multiple models were trained and tested using cross-validation, and the one with the highest cross-validation accuracy was evaluated using an independent validation set. See Methods and Supplementary Methods for details.

Figure 3. Prediction performance of *incRNA*. **(a)** Comparison between the performance of *incRNA* and two previously published methods, Rfam/Infernal and RNAz. **(b)** Comparison between the performance of our method using all features, only expression data (tiling array and RNA-seq features), and sequence and structural features only (GC%, DNA conservation, RNA

secondary structure and protein conservation). **(c)** Comparison between the performance of our method using all features, and sequence and structural features in addition to only small RNA-seq, tiling array, or poly-A+ RNA-seq data. (**d**) Predicted ncRNA scores and CDS scores of annotated genomic elements (gold-standard set) assigned by our full model. All known ncRNA bins have ncRNA scores of at least P_{high} and all other genomic elements (bins) have ncRNA scores of, at most, P_{low}. (**e**) Unannotated bins with ncRNA scores of at least P_{high} form our high-confidence candidate ncRNA bins, while bins with ncRNA scores between P_{low} and P_{high} form our medium-confidence candidate ncRNA bins. All predictions are applied on the pairwise alignments of *C. elegans* and *C. briggsae*.

Figure 4. Validation of our novel ncRNA candidates. **(a)** Fifteen novel ncRNA candidates were tested using RT-PCR. Lanes 1 to 15 on the left (+RT) correspond to the PCR results of the novel ncRNA candidates, and lane 16 is the positive control, HEX-3. The right lanes (-RT) are the negative controls without reverse transcriptase. The PCR sizes are around 150 bp in length. Fourteen of the candidates (except lane 15) were detected on the gel, among which 13 (except lane 11) showed a clear enrichment of expression signals in contrast to the negative control. **(b)** Example of a validated novel ncRNA candidate (ncRNA at lane 7 in (**a**)) with support from multiple information sources. The first and second rows correspond to the ChIP-seq reads from PHA-4 and POL II, respectively. The heights of signals are normalized by their total mapped reads. The third and fourth rows are the log-transformed values of the tiling array in late embryo and its TARs. The next two rows are the reads from small RNA sequencing and the reads from poly-A+ RNA sequencing. The last row is the annotated genes from Refseq.

Figure 5. Structural properties of the known ncRNAs and novel ncRNA candidates. The percentages of highly structured ncRNA bins (with predicted z-score of folding free energy less than the mean by at least one standard deviation) are shown for the known ncRNA bins, high-confidence and medium-confidence candidate ncRNA bins. Among the highly structured bins, the percentages that overlap with structural homologues of Rfam families are also shown. The same calculations were performed for the bins predicted to be unexpressed intergenic regions and the low-confidence ambiguous regions (bins) for comparison.

Figure 6. Expression patterns of the novel ncRNA candidates and binding signals of POL II and 22 transcription factors around their genomic regions. (**a**) Expression patterns of known miRNA transcripts, other known ncRNA transcripts and the candidate ncRNA bins based on our small RNA sequencing data at eleven developmental stages. Expression values are the log-transformed normalized read counts (DCPM: Depth of Coverage Per Million reads). Three sub-classes formed according to the expression patterns are shown in the bottom row in three different colors. All known ncRNA transcripts are shown, while 1,000 bins were randomly sampled from a total of 10,994 candidate ncRNA bins for this heat map visualization. The right panel shows a magnified view of class 1 with the colors rescaled to show the fluctuation of expression patterns across the different developmental stages. **(b)** Saturation plots of expressed known ncRNA transcripts and candidate ncRNA bin in different developmental stages. The fractions of expressed regions (with small RNA-seq signals stronger than the average signal of gold-standard intergenic regions) at the 11 developmental stages are computed using all possible combinations of the stages. The x-axis corresponds to the number of stages considered, and each point at a given number of stages corresponds to a different combination of stages. **(c)** Fractions of intergenic candidate ncRNA fragments potentially targeted by a selected subset of transcription factors. The total fractions targeted by any of the transcription factors in any of the stages are also shown. Each bar is labeled by the name of the transcription factor followed by the stage at which the binding experiment was performed. The bindings on random genome locations with the same size are also shown. Abbreviations: EMB – embryo; YA – young adult.

Tables

Table 1. Weights of nine genomic features in the trained logistic regression models for known ncRNA, CDS and UTR

	Known ncRNA	CDS	UTR
GC%	0.58	1.56	0.18
DNA Conservation [a]	0.03	0.12	0.19
Secondary Structure Free Energy [b]	-0.50	0.03	0.16
Secondary Structure Conservation [c]	1.78	-0.53	-0.50
Protein Sequence Conservation [d]	0.81	2.31	0.64
Poly-A+ RNA-seq (max) [e]	0.49	2.47	3.46
Small RNA-seq (max) [e]	2.23	0.49	0.14
Total RNA Tiling Array (max) [e]	1.57	0.29	-0.62
Poly-A+ RNA Tiling Array (max) [e]	1.75	4.47	4.67

[a] DNA conservation is the nucleotide identity in each window of the genome alignment between *C. elegans* and *C. briggsae.*

[b] The free energy of RNA secondary structure is measured by the z-score of RNA's folding $\Delta G°_{37}$ calculated by Dynalign. A stable structure favors low free energy.

[c] RNA secondary structure conservation is measured by the SCI (structure conservation index) between *C. elegans* and *C. briggsae*.

[d] Protein sequence conservation is the tblastx score divided by DNA identity in the *C. elegans* and *C. briggsae* DNA alignment.

[e] The maximum expression value from different biological samples produced by the same technology is used.

References

Agarwal A, Koppstein D, Rozowsky J, Sboner A, Habegger L, Hillier LW, Sasidharan R, Reinke V, Waterston RH, Gerstein M. 2010. Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC Genomics* **11**(1): 383.

Bartel DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**(2): 281-297.

Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM et al. 2009. Unlocking the secrets of the genome. *Nature* **459**(7249): 927-930.

Chan PP, Lowe TM. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res* **37**(Database issue): D93-97.

Coventry A, Kleitman DJ, Berger B. 2004. MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. *Proc Natl Acad Sci USA* **101**(33): 12102-12107.

di Bernardo D, Down T, Hubbard T. 2003. ddbRNA: detection of conserved secondary structures in multiple alignments. *Bioinformatics* **19**(13): 1606-1611.

Fejes-Toth K, Sotirova V, Sachidanandam R, Assaf G, Hannon GJ, Kapranov P, Foissac S, Willingham AT, Duttagupta R, Dumais E et al. 2009. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* **457**(7232): 1028-1032.

Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR et al. 2009. Rfam: updates to the RNA families database. *Nucleic Acids Res* **37**(Database issue): D136-140.

Gerstein MB Lu ZJ Nostrand ELV Cheng C Arshinoff BI Liu T Yip K Robilotto R Rechtsteiner A Ikegami K et al. 2010. Integrative Analysis of Functional Elements in the *Caenorhabditis elegans* Genome by the modENCODE Project. *Submitted to Science*.

Gorodkin J, Hofacker IL, Torarinsson E, Yao Z, Havgaard JH, Ruzzo WL. 2009. De novo prediction of structured RNAs from genomic sequences. *Trends Biotechnol* **28**(1): 9-19.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36**(Database issue): D154-158.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**(7235): 223-227.

Harmanci AO, Sharma G, Mathews DH. 2007. Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. *BMC Bioinformatics* **8**: 130.

Havgaard JH, Torarinsson E, Gorodkin J. 2007. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Comput Biol* **3**(10): 1896-1908.

He H, Wang J, Liu T, Liu XS, Li T, Wang Y, Qian Z, Zheng H, Zhu X, Wu T et al. 2007. Mapping the C. elegans noncoding transcriptome with a whole-genome tiling microarray. *Genome Res* **17**(10): 1471-1477.

Jung CH, Hansen MA, Makunin IV, Korbie DJ, Mattick JS. 2010. Identification of novel non-coding RNAs using profiles of short sequence reads from next generation sequencing data. *BMC Genomics* **11**: 77.

Kato M, de Lencastre A, Pincus Z, Slack FJ. 2009. Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during Caenorhabditis elegans development. *Genome Biol* **10**(5): R54.

Katz L, Burge CB. 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res* **13**(9): 2042-2051.

Kiontke K, Fitch DH. 2005. The phylogenetic relationships of Caenorhabditis and other rhabditids. *WormBook*: 1-11.

Klattenhoff C, Theurkauf W. 2008. Biogenesis and germline functions of piRNAs. *Development* **135**(1): 3-9.

Lestrade L, Weber MJ. 2006. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res* **34**(Database issue): D158-162.

Li SC, Tang P, Lin WC. 2007. Intronic microRNA: discovery and biological implications. *DNA Cell Biol* **26**(4): 195-207.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**(5): 955-964.

Luke B, Lingner J. 2009. TERRA: telomeric repeat-containing RNA. *EMBO J* **28**(17): 2503-2510.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**(7057): 376-380.

Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS. 2008. Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci USA* **105**(2): 716-721.

Missal K, Zhu X, Rose D, Deng W, Skogerbo G, Chen R, Stadler PF. 2006. Prediction of structured non-coding RNAs in the genomes of the nematodes Caenorhabditis elegans and Caenorhabditis briggsae. *J Exp Zool B Mol Dev Evol* **306**(4): 379-392.

Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**(10): 1335-1337.

Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* **2**(4): e33.

Rivas E, Eddy SR. 2001. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* **2**: 8.

Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. 2009. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* **27**(1): 66-75.

Schattner P, Brooks AN, Lowe TM. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* **33**(Web Server issue): W686-689.

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**(5741): 1728-1732.

Torarinsson E, Sawera M, Havgaard JH, Fredholm M, Gorodkin J. 2006. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res* **16**(7): 885-889.

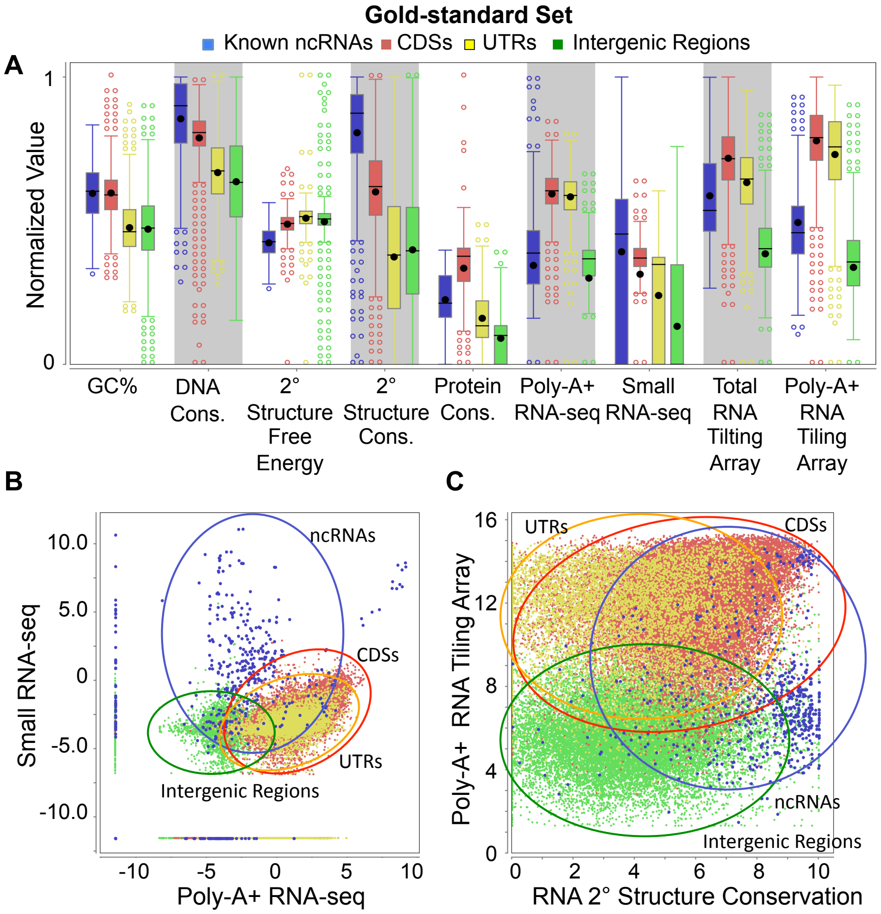
Torarinsson E, Yao Z, Wiklund ED, Bramsen JB, Hansen C, Kjems J, Tommerup N, Ruzzo WL, Gorodkin J. 2008. Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. *Genome Res* **18**(2): 242-251.

Uzilov AV, Keegan JM, Mathews DH. 2006. Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics* **7**(1): 173.

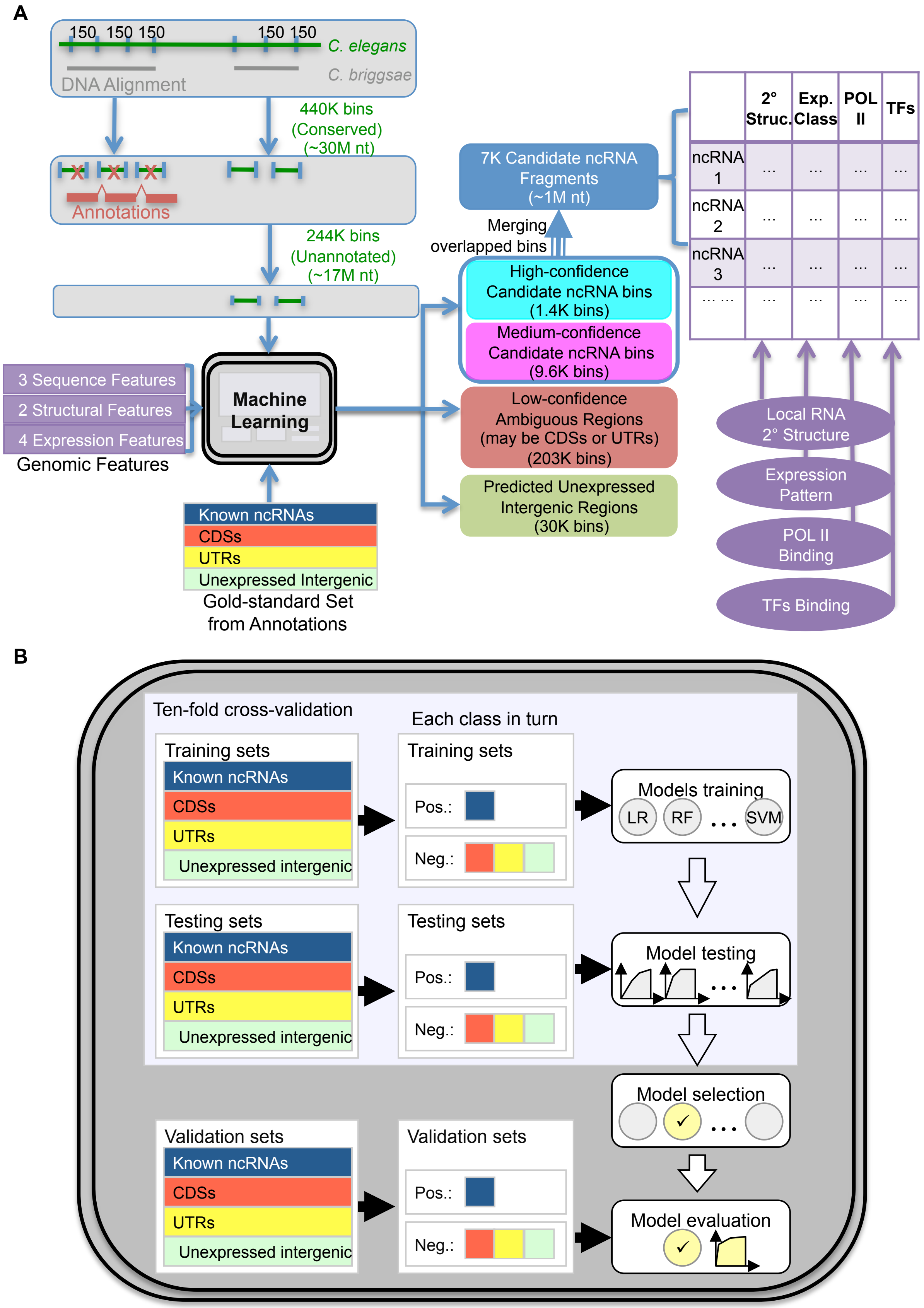
Washietl S, Hofacker IL, Stadler PF. 2005. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci USA* **102**(7): 2454-2459.

Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermuller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A et al. 2007. Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res* **17**(6): 852-864.

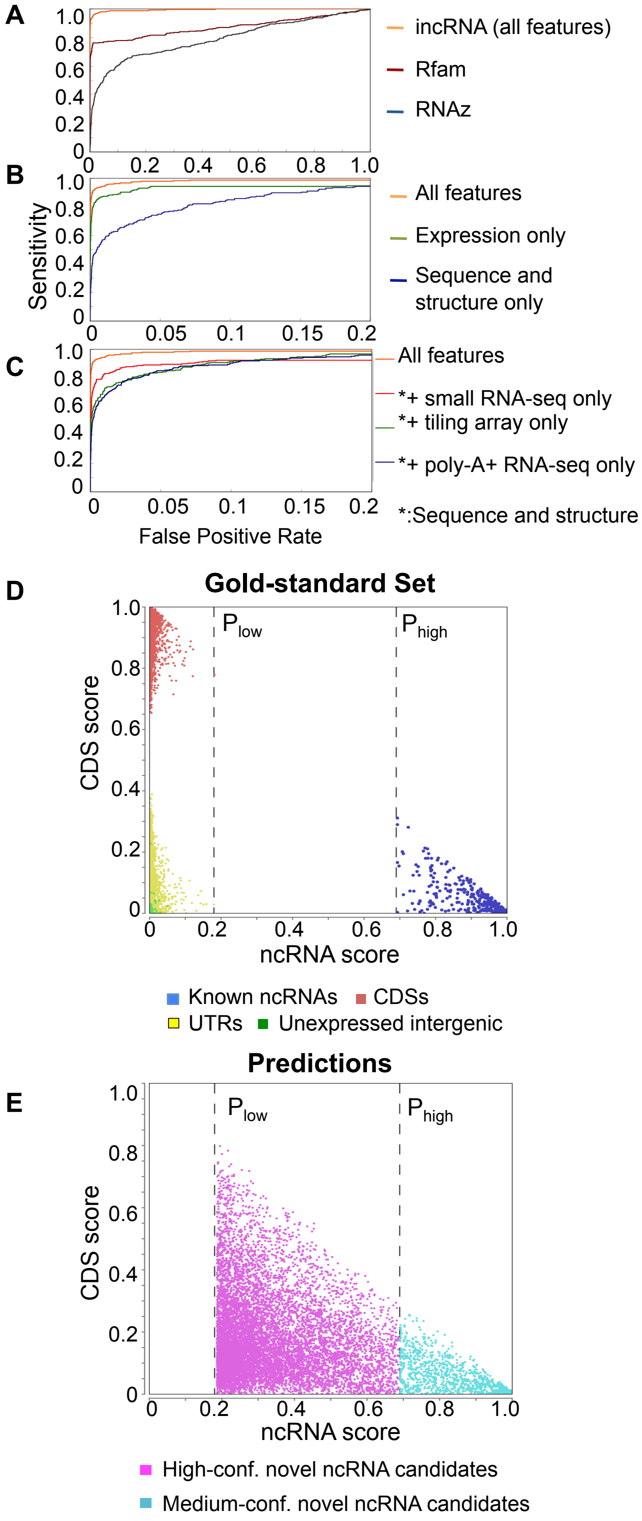
Zhong M, Niu W, Lu ZJ, Sarov M, Murray JI, Janette J, Raha D, Sheaffer KL, Lam HYK, Preston E et al. 2010. Genome-Wide Identification of Binding Sites Defines Distinct Functions for <italic>Caenorhabditis elegans</italic> PHA-4/FOXA in Development and Environmental Response. *PLoS Genet* **6**(2): e1000848.

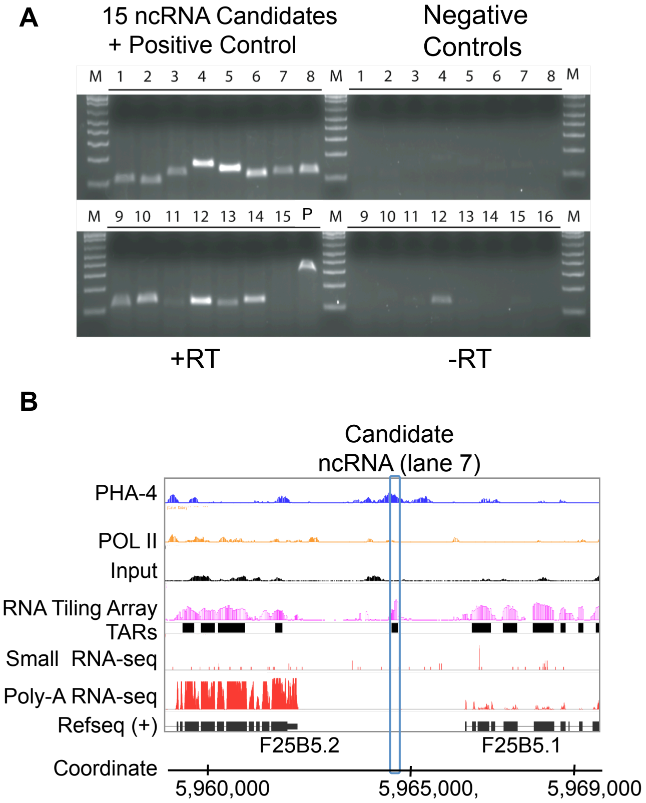


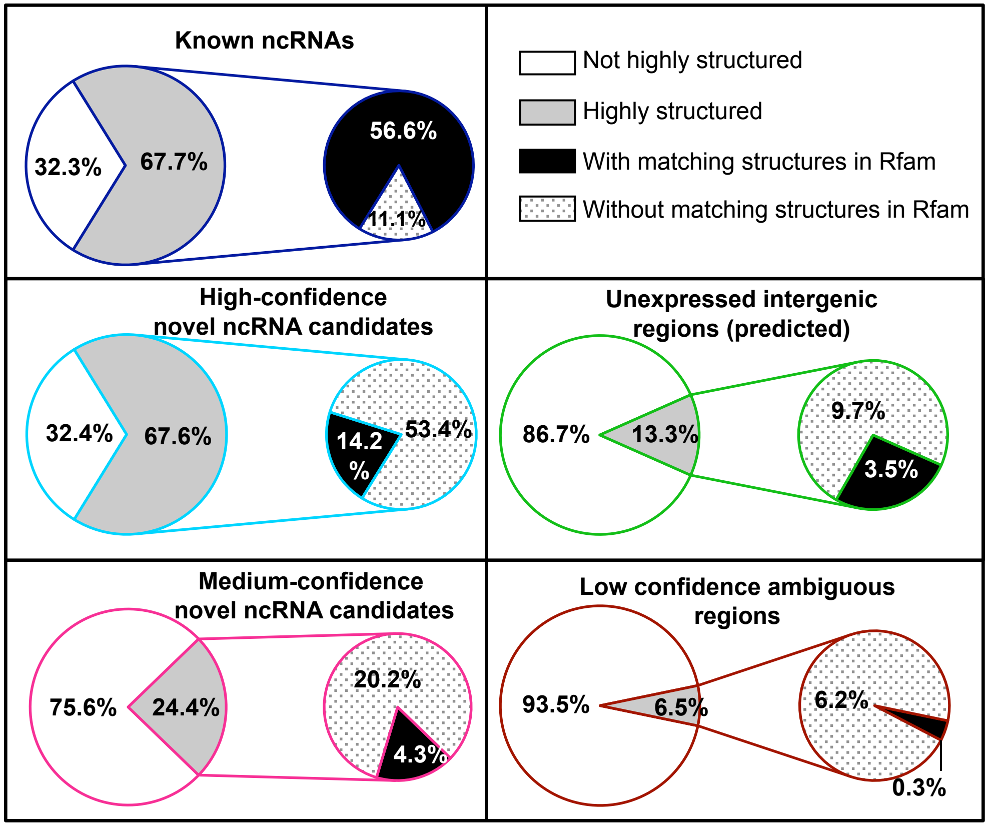
Gold-standard Set
Known ncRNAs
CDSs
UTRs
Intergenic Regions
A
Normalized Value
1
0
GC%
DNA Cons.
2° Structure Free Energy
2° Structure Cons.
Protein Cons.
Poly-A+ RNA-seq
Small RNA-seq
Total RNA Tilting Array
Poly-A+ RNA Tiling Array
B
Small RNA-seq
10.0
5.0
0
-5.0
-10.0
-10
-5
0
5
10
Poly-A+ RNA-seq
ncRNAs
CDSs
UTRs
Intergenic Regions
C
Poly-A+ RNA Tiling Array
16
12
8
4
0
0
2
4
6
8
10
RNA 2° Structure Conservation
UTRs
CDSs
ncRNAs
Intergenic Regions

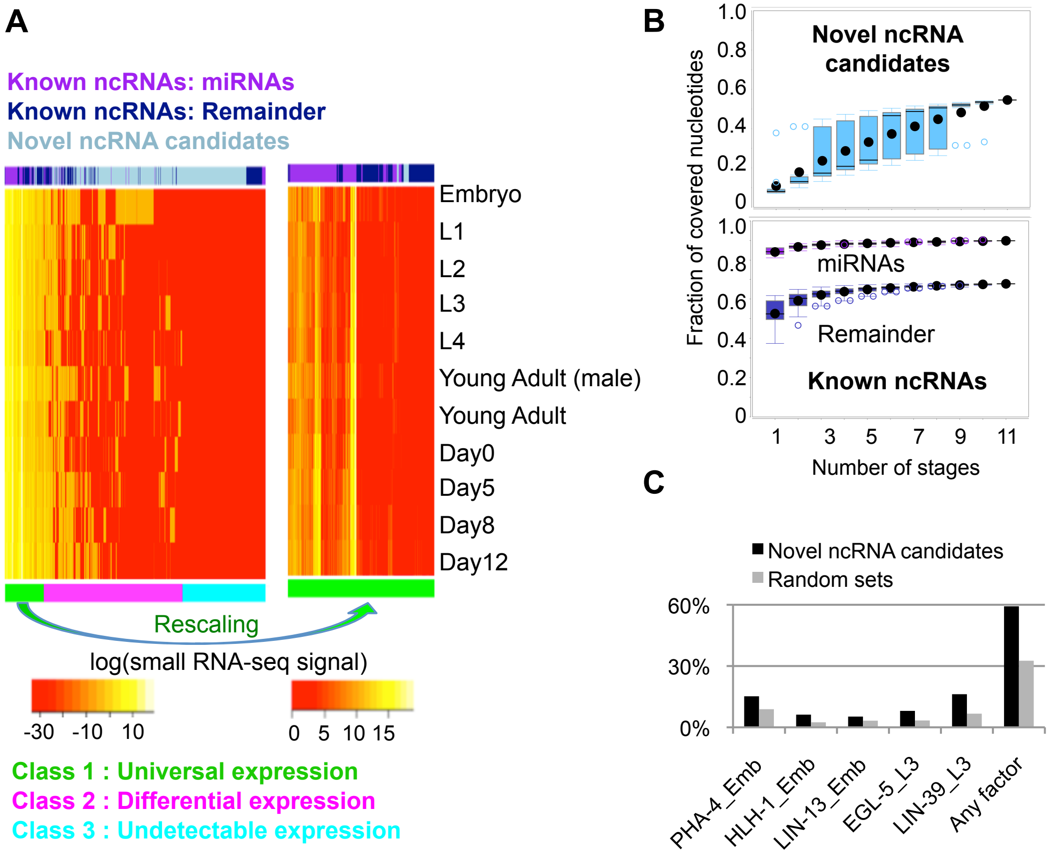


	2° Struc.	Exp. Class	POL II	TFs
ncRNA 1	…	…	…	…
ncRNA 2	…	…	…	…
ncRNA 3	…	…	…	…
… …	…	…	…	…


A
1.0
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1.0
incRNA (all features)
Rfam
RNAz
B
Sensitivity
1.0
0.8
0.6
0.4
0.2
0
0
0.05
0.1
0.15
0.2
All features
Expression only
Sequence and structure only
C
1.0
0.8
0.6
0.4
0.2
0
0
0.05
0.1
0.15
0.2
False Positive Rate
All features
*+ small RNA-seq only
*+ tiling array only
*+ poly-A+ RNA-seq only
*:Sequence and structure
D
Gold-standard Set
CDS score
1.0
0.8
0.6
0.4
0.2
0
P_{low}
P_{high}
0
0.2
0.4
0.6
0.8
1.0
ncRNA score
■ Known ncRNAs
■ CDSs
■ UTRs
■ Unexpressed intergenic
E
Predictions
CDS score
1.0
0.8
0.6
0.4
0.2
0
P_{low}
P_{high}
0
0.2
0.4
0.6
0.8
1.0
ncRNA score
■ High-conf. novel ncRNA candidates
■ Medium-conf. novel ncRNA candidates


A
15 ncRNA Candidates
+ Positive Control
Negative
Controls
M 1 2 3 4 5 6 7 8 M 1 2 3 4 5 6 7 8 M
M 9 10 11 12 13 14 15 P M 9 10 11 12 13 14 15 16 M
+RT
-RT
B
Candidate
ncRNA (lane 7)
PHA-4
POL II
Input
RNA Tiling Array
TARs
Small RNA-seq
Poly-A RNA-seq
Refseq (+)
F25B5.2
F25B5.1
Coordinate
5,960,000
5,965,000,
5,969,000


Known ncRNAs
32.3%
67.7%
56.6%
11.1%
Not highly structured
Highly structured
With matching structures in Rfam
Without matching structures in Rfam
High-confidence novel ncRNA candidates
32.4%
67.6%
14.2 %
53.4%
Unexpressed intergenic regions (predicted)
86.7%
13.3%
9.7%
3.5%
Medium-confidence novel ncRNA candidates
75.6%
24.4%
20.2%
4.3%
Low confidence ambiguous regions
93.5%
6.5%
6.2%
0.3%


A
Known ncRNAs: miRNAs
Known ncRNAs: Remainder
Novel ncRNA candidates
Embryo
L1
L2
L3
L4
Young Adult (male)
Young Adult
Day0
Day5
Day8
Day12
Rescaling
log(small RNA-seq signal)
-30
-10
10
0
5
10
15
Class 1 : Universal expression
Class 2 : Differential expression
Class 3 : Undetectable expression
B
Fraction of covered nucleotides
1.0
0.8
0.6
0.4
0.2
0
Novel ncRNA candidates
miRNAs
Remainder
Known ncRNAs
1
3
5
7
9
11
Number of stages
C
Novel ncRNA candidates
Random sets
60%
30%
0%
PHA-4_Emb
HLH-1_Emb
LIN-13_Emb
EGL-5_L3
LIN-39_L3
Any factor