104–114 *Nucleic Acids Research, 2015, Vol. 43, No. 1* *Published online 12 December 2014*
doi: 10.1093/nar/gku1316

A common set of distinct features that characterize noncoding RNAs across multiple species

Long Hu[1,2], Chao Di[2], Mingxuan Kai[2], Yu-Cheng T. Yang[2], Yang Li[2], Yunjiang Qiu[2], Xihao Hu[3], Kevin Y. Yip[3], Michael Q. Zhang[4,5] and Zhi John Lu[2,*]

[1]PKU-Tsinghua-NIBS Graduate Program, School of Life Sciences, Peking University, Beijing 100871, China, [2]MOE Key Laboratory of Bioinformatics, Center for Synthetic and Systems Biology and Center for Plant Biology, School of Life Sciences, Tsinghua University, Beijing 100084, China, [3]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, [4]Department of Molecular and Cell Biology, Center for Systems Biology, The University of Texas, Dallas 800 West Campbell Road, RL11 Richardson, TX 75080-3021, USA and [5]MOE Key Laboratory of Bioinformatics and Bioinformatics Division, Center for Synthetic and Systems Biology, TNLIST and School of Medicine, Tsinghua University, Beijing 100084, China

Received October 16, 2014; Revised December 03, 2014; Accepted December 05, 2014

ABSTRACT

To find signature features shared by various ncRNA sub-types and characterize novel ncRNAs, we have developed a method, *RNAfeature*, to investigate >600 sets of genomic and epigenomic data with various evolutionary and biophysical scores. *RNAfeature* utilizes a fine-tuned intra-species wrapper algorithm that is followed by a novel feature selection strategy across species. It considers long distance effect of certain features (e.g. histone modification at the promoter region). We finally narrow down on 10 informative features (including sequences, structures, expression profiles and epigenetic signals). These features are complementary to each other and as a whole can accurately distinguish canonical ncRNAs from CDSs and UTRs (accuracies: >92% in human, mouse, worm and fly). Moreover, the feature pattern is conserved across multiple species. For instance, the supervised 10-feature model derived from animal species can predict ncRNAs in Arabidopsis (accuracy: 82%). Subsequently, we integrate the 10 features to define a set of noncoding potential scores, which can identify, evaluate and characterize novel noncoding RNAs. The score covers all transcribed regions (including unconserved ncRNAs), without requiring assembly of the full-length transcripts. Importantly, the noncoding potential allows us to identify and characterize potential functional domains with feature patterns similar to canonical ncRNAs (e.g. tRNA, snRNA, miRNA, etc) on ∼70% of human long ncRNAs (lncRNAs).

INTRODUCTION

The advent of high-throughput sequencing technologies has facilitated the identification of a large number of previously unannotated transcripts, many of which correspond to novel noncoding RNAs (ncRNAs). But only ∼1% of the human genome have been annotated as canonical ncRNAs (e.g. rRNA, tRNA, miRNA, snoRNA, etc.) and long ncRNAs (lncRNAs) by ENCODE (1). It has been difficult to comprehensively identify all noncoding transcripts from a species since there are many ncRNA sub-types, some ncRNAs are expressed only in particular cell types or conditions, and some have properties similar to messenger RNAs (mRNAs). Among the novel ncRNAs, lncRNAs with >200 nucleotides each represent a large class of ncRNAs that have attracted a lot of attention due to their diverse functional roles recently discovered (2,3). However, only a small portion of the lncRNAs has been well validated and characterized. Many transcripts detected by RNA-seq and array data were directly annotated as lncRNAs based on coding potential filters (4–6).

To evaluate and characterize novel ncRNAs, it is crucial to comprehensively understand the signature features shared by various known ncRNA types. Unlike coding genes with clear organizational structures utilized by the transcriptional and translational machineries, ncRNAs in general have less obvious sequence characteristics. Some ncRNAs, lncRNAs in particular, are not strongly conserved evolutionarily, and do not have clear secondary structures. Many ncRNAs express only in certain cell types or conditions, and would thus be missed if expression data of these cell types or conditions are not available. Some lncRNAs also share common properties with mRNAs, containing introns and poly-A tails. Due to all these reasons, it is difficult to identify and characterize novel ncRNAs compre-

*To whom correspondence should be addressed. Tel: +86 10 62789217; Fax: +86 10 62789217; Email: zhilu@tsinghua.edu.cn

hensively using any single type of information alone (7,8). Consequently, the idea of combining multiple lines of evidence for novel ncRNA identification has been proposed. By combining features derived from sequences, structures, evolutionary conservation and expression profiles, the models produced by these integrative methods were shown to have much higher sensitivity and specificity than methods based on single features (1,9–12).

Having high accuracy notwithstanding, these models are in general hard to interpret due to the large number (tens or even hundreds) of features involved. In addition, most of these previous studies focused on ncRNAs from one particular species. It is not clear whether the features useful for identifying ncRNAs from these species are generally useful in identifying ncRNAs from other species. These complex models also suffer from the drawback that they cannot be applied to a certain species if some of the involved features are not available in this species.

Here, we propose a new strategy called *RNAfeature* for determining a succinct set of essential and informative features that can accurately identify various ncRNA types from multiple species. It involves an intra-species part for identifying characteristic ncRNA features from each species by combining existing biophysical knowledge of ncRNAs with high-throughput experimental data, and an inter-species part for forming a common set of general ncRNA features across different species. The method used in the intra-species part of *RNAfeature* is improved upon our previous supervised model (*incRNA*) for identifying ncRNAs from a single species (1,9), with new feature types and substantially more data supplied as inputs. The inter-species part of *RNAfeature* is a novel component newly introduced in this work. In addition, we also introduce a *context influence score* (*CIS*) to consider the long distance effect of certain features (e.g. some histone modification signals are enriched at the promoter region).

RNAfeature investigated >600 sets of genomic and epigenomic data with various evolutionary and biophysical scores to look for conserved signature features of ncRNAs in multiple species. It narrowed down on 10 essential features (including sequences, structures, expression profiles and histone modification signals) that are sufficient for identifying all known types of canonical ncRNAs (miRNA, rRNA, snRNA, Y RNA, etc.) and accurately distinguish them from other genomic elements such as CDSs and UTRs in human, mouse, fly and worm (accuracies: >92%). Furthermore, the 10-feature model derived from animal species is conserved and predicts ncRNAs in Arabidopsis (accuracy: 82%). We also show that this feature set is robust in that even when we omit some canonical ncRNA types from the input of *RNAfeature*, the resulting model can still accurately identify ncRNAs of the omitted types. This indicates the generality of the identified features and the ability of *RNAfeature* in finding novel ncRNA types.

Finally, we used the 10 features to define a set of noncoding potential scores, which can identify, evaluate and characterize novel noncoding RNAs. The potential score covers all transcribed regions (including unconserved ncRNAs) without the full-length transcripts being assembled, because it is derived from complementary features at bin level (100 nt). The noncoding potential has identified the potential

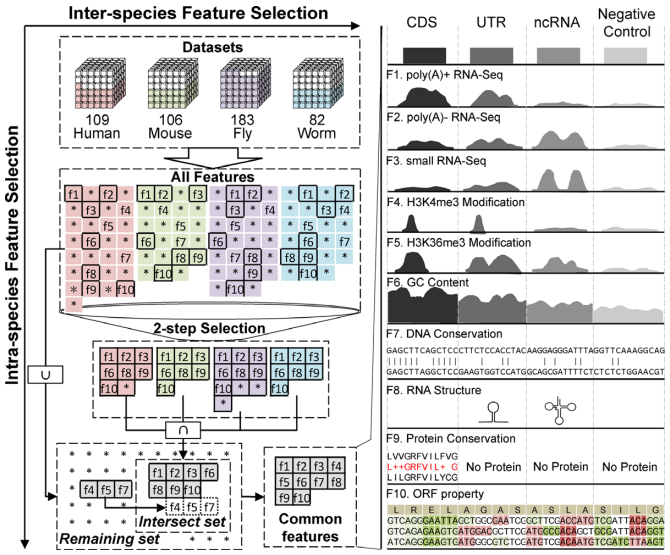


Figure 1. Workflow of *RNAfeature:* an inter-species feature selection method. We collected high-throughput datasets, sequence and structural features from four species and classified them according to the feature type as indicated by f1, f2, …, f10, and others. Features from different species are color-coded by red for human, green for mouse, purple for fly and blue for worm. We used all those features to classify the four genomic elements, i.e. coding sequence (CDS), untranslated region (UTR), noncoding RNA (ncRNA), and the negative control. The feature selection process includes intra-species feature selection part that eliminates features with low prediction power for genomic regions within the same species, and inter-species feature selection part that checks the remaining features shared by multiple species. After obtaining the four sets of candidate features from the intra-species feature selection, we took the intersection of them to obtain a common set which was then combined with three selected features that are important in at least three species. The final feature set contains 10 features that are listed with toy data in the right panel.

functional regions with feature patterns similar to canonical ncRNAs on ∼70% of human long ncRNAs (lncRNAs). For instance, ∼10% of the human lncRNAs (e.g. MALAT1) contain local domains with canonical structures (e.g. tRNA-like). Overall, our work provides a whole-genome resource to support further biological discoveries and mechanism studies of novel ncRNAs in model organisms.

MATERIALS AND METHODS

Overview of *RNAfeature*

RNAfeature has two pre-process parts (assigning feature values and assigning annotations to genomic regions) and two feature selection parts (intra-species and inter-species feature selection) (Figure 1).

In the first pre-processing part of *RNAfeature*, we curated 622 high-throughput data sets in five species, i.e. *Homo sapiens*, *Mus musculus*, *Caenorhabditis elegans, Drosophila melanogaster* and *Arabidopsis thaliana*. For every 100nt, called genomic bin, in the genome, we calculated not only the expression, histone modifications and TRF (transcription and regulation factors) binding signals from high-throughput data, but also various sequence and structural features. We further considered the upstream and down-

stream signals of genomic bins for each features, especially the histone modifications and TRF binding.

The second pre-processing part is to label genomic bins by their genomic elements (i.e. CDS, UTR, rRNA, miRNA, lncRNA, etc.) based on the gold standard annotations. In each species, we carefully prepared a gold-standard set of annotations for four types of genomic locations: canonical ncRNAs, confirmed coding sequences (CDSs), 5′ and 3′ untranslated regions (UTRs) and negative controls (defined as intergenic regions with weak expression signals). The bins overlapped with these annotations were selected as the gold-standard bins for training and testing.

The accuracies of classifying these four types of genomic regions were used for feature selection. We utilized a supervised machine-learning framework with cross-validations. When constructing the training and test sets, we optimized the sampling method. Different machine learning classifiers were also compared and optimized for the feature selection.

The intra-species feature selection had a pre-filter step and a rigorous search step. In the pre-filter step, we used a recursive feature elimination (RFE) algorithm (13) to filter inessential features. Then, we used a greedy hill climbing method, called greedy backward algorithm (GBA) (14), to rigorously eliminate features.

Subsequently, the inter-species feature selection had two steps to optimize the feature set shared by multiple species. We first intersected the feature sets from different species and retained other features in the union set as *remaining features*. To save essential features incorrectly filtered at previous steps, we added an optimized subset of the *remaining features* into the intersected set to obtain the final set.

The common features were selected from four animal species (i.e. human, mouse, fly and worm) and validated in a plant species, Arabidopsis.

Feature scores from high-throughput data

In total, we curated and processed 622 high-throughput datasets in five species: *H. sapiens, M. musculus, C. elegans, D. melanogaster* and *A. thaliana* (Supplemental Table S1 and Supplemental Dataset 1). The data sets belong to three types that are expression levels (16 billion RNA-seq reads and 400 million array probes), histone modifications (18 types of histone modifications) and TRF binding (two TRFs). The reference genome sequence of each species was split into 100nt bins, with a step size of 50nt. The feature values were calculated at bin level.

Expression data. Various expression values were obtained from raw data, including 98 poly(A)+ RNA, 41 poly(A)− RNA, 48 total RNA, 70 small RNA sequencing data sets, and 101 poly(A)+ or total RNA tiling array data sets (see details in Supplemental Notes, Supplemental Table S1 and Supplemental Dataset 1). We analyzed approximately 16 billion mapped reads from RNA-seq data and 400 million probes from tiling array data. For RNA-seq, we calculated the reads per kilobase per million (RPKM) of each bin using the DEGseq software package (http://www.bioconductor.org/packages/release/bioc/html/DEGseq.html). For tiling array data, the expression values were calculated using an R package AffyTiling (http://www.bioconductor.org/packages/release/bioc/html/AffyTiling.html), and the maximum intensity of overlapped probes was assigned to each bin. The signals from different replicates were averaged. Recent studies have shown that many novel ncRNAs tend to be specifically expressed in certain conditions or tissues (sometimes only up-regulated in one specific condition) (15,16). In order to make our model more sensitive to novel ncRNA detection, we used the maximum expression values (of RPKM or the probe intensity) among different conditions (i.e. tissues, cell lines, etc.) for each type of expression feature such as poly(A)+, poly(A)− and small RNA.

Histone modification and transcriptional regulatory factor binding data. We curated ChIP-seq data from 18 types of histone modifications and two TRFs (see details in Supplemental Note and Supplemental Table S1). The peaks from the human ChIP-seq data were downloaded from Encode (Supplemental Dataset 1). For other ChIP-seq data, we used MACS14 (17) to convert raw reads. The values of each sample were transformed into Z-score by smoothing 0.01% of the outlier values. The Z-scores from different conditions (i.e. tissues, developing time, cell lines, etc.) were averaged to produce a single value for each type of data.

Sequence and structural scores

We curated and calculated various computational scores for each bin of 100nt in genomes of five species (Supplemental Table S2).

GC content was calculated by counting the proportion of G and C bases in the sequence.

DNA sequence conservation was measured using BLASTn with default parameters for worm, fly and Arabidopsis. The library databases for worm, fly and Arabidopsis were downloaded from Wormbase (with 19 other nematode species), Flybase (with 11 other *Drosophila* species) and EnsemblePlants (with 31 other plant species), respectively. The PhastCons scores for humans are phastCons46way, and the scores for mice are phastCons30way as provided by the UCSC genome browser.

Protein conservation was measured by BLAST based on the same library sources as the DNA sequence conservation, except that we extracted all the vertebrate protein sequences from NCBI nr database to build the BLAST library for human and mouse. We used tBLASTx for worm, and BLASTx for other species to calculate the conservation scores.

RNA secondary structure stability was represented by the p-value of the free energy calculated by the Randfold program (18). The dinucleotide frequency was retained and 1000 shuffles were used to generate a random background of RNA structures.

RNA secondary structure homologs were searched by the *INFERNAL* program with default parameters (19).

RNA secondary structure conservation was denoted by structure conservation index (SCI) scores calculated using *RNAz* (20) based on multiple sequence alignments with default parameters.

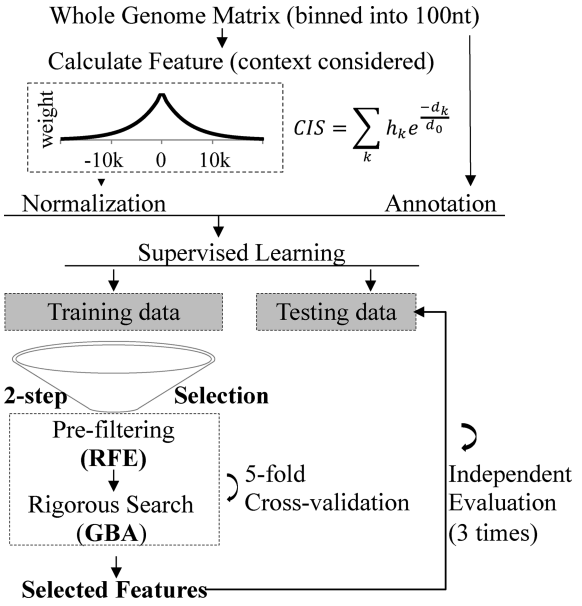


Figure 2. Details of intra-species feature selection and *context influence score*. The whole genome is split into 100-nt bins. Then, the bins are classified by genomic annotations (i.e. CDS, UTR, negative control and canonical ncRNA). Multiple feature values are calculated for each bin, considering both local score and a *CIS*. Subsequently, the features are pre-filtered using RFE algorithm and rigorously searched with a GBA. They are evaluated by a 5-fold cross validation inside the training set. Finally, the selected features are further validated with an independent test set. The accuracy of classifying the four genomic elements (CDS, canonical ncRNA, UTR and negative control) is the criteria of evaluating the selection performance.

ORF property was measured by *RNAcode* using the same multiple sequence alignments. It was also called coding potential in *RNAcode* paper (6).

Multiple species alignments. We downloaded 46-way alignments for human, 30-way alignments for mouse, 7-way alignments for worm and 15-way alignments for fly from the UCSC Genome Browser (http://hgdownload.soe.ucsc.edu/goldenPath). The 5-way alignments for Arabidopsis were downloaded from VISTA (http://pipeline.lbl.gov/downloads.shtml).

Definition of *context influence score*—downstream and upstream influence

Because of the long distance effects of features (e.g. some histone modification signals enriched at the promoter region), we considered the upstream and downstream values for each local bin (upper panel in Figure 2). Considering that the influence of context declines with distance, we calculated a *CIS* of a feature in a bin using an exponential weight function (21):

$$CIS = \sum_k h_k e^{\frac{-d_k}{d_0}}$$

h_k = feature score

d_k = distance to the current bin

d_0 = context range(assigned by species and feature type)

The up-stream *CIS*, local value, and down-stream *CIS* were always considered as a group when providing the feature for each bin. The distance parameter d_0 was optimized in each species (Supplemental Note). For histone modification and TRF features, we used aggregating plots of the signals around TSS (transcription start site) regions to determine the d_0 (Supplemental Figures S1 and S2). It is 2000nt for fly, worm and Arabidopsis genomes, and 5000nt for human and mouse genomes. For expression, sequence, and structural features, d_0 was determined by the exon size (Supplemental Figure S3), as these features might be highly correlated for bins within the same exon. It is 1500nt for fly, worm and Arabidopsis genomes, and 3000nt for human and mouse genomes. For each species, we compared model performances with and without *CIS*. Overall, considering the long distance effect has improved our model (Supplemental Figure S4).

Feature score normalization

Initially, the feature values have very diverse range since they were derived from genomic loci of multiple species. To reduce the data heterogeneity, we first smoothed the top and bottom 0.005% values of each feature. Then, the features with large range (>1000) in value were log-transformed by $y = \log_{10}(x - \text{minimum} + 1)$. Subsequently, all the feature values were scaled between 0 and 1 with the linear transformation:

$$y = \frac{x - \text{minimum}}{\text{maximum} - \text{minimum}}.$$

Gold standard annotations

We used the genome annotations downloaded from Gencode (version 10), MGI (version GRCm38), Wormbase (version ws220), Flybase (version r5.45), and TAIR (version 10) to annotate each genomic bin. The summary of the annotations is shown in Supplemental Table S3. In all five species, four basic genomic elements were used as gold standard annotations for the feature selection calculations: CDSs, untranslated regions (5′ and 3′ UTRs), canonical ncRNA and negative controls. The canonical ncRNAs are well-classified ncRNAs, including rRNA, tRNA, snoRNA, snRNA, miRNA, Y RNA and 7SK RNA. The negative controls were obtained using three criteria: (1) bins located in the intergenic regions (Supplemental Figure S5), (2) with expression levels lower than the average values of all intergenic bins across all expression data, (3) containing no ambiguous nucleotides (i.e. *N*). The intergenic regions were defined as regions located at least *N*-nt away from any annotated elements (i.e. coding genes, ncRNAs, TEs and pseudogenes). *N* is 2000 for human and mouse, and 500 for worm, fly and Arabidopsis.

Each genomic bin (100nt) is annotated by the following order (Supplemental Figure S6): (1) if 50% of a bin overlaps with a known ncRNA region (such as canonical ncRNAs and long ncRNAs), it is labeled as the corresponding ncRNA type; (2) if 90% of a bin overlaps with a CDS, UTR, ancestral repeat or intergenic region, it is labeled correspondingly; (3) if 50% of a bin overlaps with a pseudogene, intronic, TE or ambiguous region, it is labeled correspondingly. The summary of the annotated bins is shown in Supplemental Table S4.

Construction of training and test sets

We tested various methods and strategies to determine the appropriate training and test sets. When optimizing the classifier and feature selection algorithm, we adapted the training strategy from our previous study (9), used 5-fold cross-validation in the training set (2/3 of the entire data), followed by further validation in an independent test set (1/3 of the data) (Figure 2).

Genomic distance. Because the feature values (e.g. expression level) of close bins are highly correlated, we need to ensure that bins sampled in the test set are distant from those used in the training set. Thus, nearby bins were packed into one block, and the bins from the same block were put in the same set. The minimum genomic distance between adjacent blocks was chosen for each genome (Supplemental Figures S7 and S8). To avoid exons from the same gene being packed into two blocks, we used a distance that was longer than 90% of introns, which was 15 000nt for human and mouse, and 5000nt for worm, fly and Arabidopsis (Supplemental Table S5).

Balanced classes. In most genomes, the CDS bins outnumber the ncRNA bins. To prevent the prediction bias (Supplemental Figure S9), we sampled same numbers of bins from each genomic class (Supplemental Figures S10 and S11) for the following feature selection steps.

Classifier optimization

We tested various machine-learning classifiers to optimize the classification performance, including the Naïve Bayes, Logistic Regression, Support Vector Machine (SVM) and Random Forest (Supplemental Note). We compared their performances for the 4-class prediction (CDS, UTR, canonical ncRNA and negative control) in human. The Random Forest method was finally chosen for the downstream analysis because it had the best performance (Supplemental Figure S12).

Construction of the initial feature set

We curated 25 types of expression, histone and TRF features (Supplemental Table S1). We selected 15 features that were available in at least three species as the initial set for feature selection. Previously studies have shown that gene-rich regions tend to be more conserved and have higher GC content (22,23). It was also suggested that many ncRNAs have conserved and stable secondary structures (9,24). Therefore, we also added seven sequence and structure scores (GC content, DNA sequence conservation, protein conservation, RNA secondary structure stability, RNA secondary structure homologs, RNA secondary structure conservation and ORF property).

Intra-species feature selection

The accuracy (ACC) of classifying four genomic elements (CDS, canonical ncRNA, UTR and negative control) was used as the criteria to select the best feature set within each species (Figure 2). The accuracy, called ACC in statistics, is the ratio of correctly predicted instances over all instances belonging to the four genomic elements. Here, the correctly predicted instances are calculated from four classes: correctly predicted CDS bins + correctly predicted UTR bins + correctly predicted ncRNA bins + correctly predicted negative control bins. We started from the feature set mentioned above, followed by a two-step wrapper method to remove the worst features. One was a pre-filtering step that used the RFE method (13). The other was a rigorous search step using the GBA (14).

Performance variance. We first estimated the performance variance of a classifier (i.e. Random Forest) when the same training set was used repeatedly (Supplemental Figure S13). The distribution of classifier accuracies for 100 repeats was approximately a normal distribution (Supplemental Figure S13A). And, the randomness was quantified as 4-fold of the standard deviation of the accuracy. Thus, during the pre-filtering step, a smaller set was selected by sacrificing the accuracy within the randomness allowance, instead of a feature set with the highest accuracy value (Supplemental Figure S13B and C).

Pre-filtering step used the RFE method. RFE irreversibly eliminated the feature with the lowest rank one by one, until finding the smallest feature set (Supplemental Figure S14). We tested two ranking criteria for RFE: Spearman's correlation coefficient and Random Forest Importance Score. We then chose the latter one because fewer features were selected with a higher accuracy (Supplemental Table S6A).

Rigorous search step used the GBA. It iteratively evaluated a subset of features and removes useless ones (Supplemental Figure S15).

Other selection strategies such as embedded method were also tested and compared (Supplemental Note and Supplemental Table S6B).

We summarized the features used and selected by intra-species feature selection in Supplemental Figure S16.

Inter-species feature selection. Inter-species feature selection chose features with balanced performances among multiple species. After the intra-species feature selection, seven features shared by all four species were retained and called the *intersecting feature set*: $I = (I_1, ..., I_7)$. The remaining 15 features were called the *remaining feature set*: $L = (L_1, ..., L_{15})$ (Supplemental Figure S17).

Because some essential features shared by multiple species could be filtered out during the intra-species feature selection, we tried to add some features from the *remaining set L* back to the *intersecting set I*. Picking n features from the *remaining feature set* has $m (m = C_{15}^{n})$ different possibilities. G represents a subset of L. It can be denoted as $G_{n,j}$, where $j = 1, ..., m$ (Supplemental Figure S18). We ranked the classifier accuracies for different combination set ($C_{n,j} = G_{n,j} + I$) within every species, s(s = human, mouse, worm or fly) (Supplemental Figure S19A). We found that DNA sequence conservation, H3K36me3, and H3K4me3 were frequently selected in highly ranked combinations in at least three species.

To test the robustness of the selection method, we tried different combination sizes, n, from 2 to 4, and it always outputted the three features from the remaining set, L (Supple-

mental Figure S19B). Therefore, ten features ($I_{1-7} + L_{1-3}$) were selected as the final feature set (Supplemental Figure S20).

Feature weights in multiple species

To compare models across species, we did quantile normalization for the feature values.

Feature weights using Softmax. The Softmax algorithm (http://cs229.stanford.edu), a multinomial logistic regression algorithm, was used to evaluate the weights of the 10 selected features for CDS, ncRNA, UTR, and negative control.

ncRNA types used for cross-type validations

The performance of ten-feature model for the 4-class (CDS, UTR, negative control and canonical ncRNA) prediction was validated on different types of canonical ncRNAs (i.e. transfer RNA, tRNA; ribosomal RNA, rRNA, microRNA, miRNA; small nuclear RNA, snRNA; small nucleolar RNA, snoRNA; Y RNA; 7SK RNA). For instance, when predicting a specific type of ncRNA (e.g. rRNA in human), we used 1202 CDS bins, 1202 UTR bins, 1202 negative control bins and 1202 rRNA bins in the test set. In the training set, we used 10 000 CDS bins, 10 000 UTR bins, 10 000 negative control bins and 10 000 all other six types of ncRNAs bins.

In addition, we also treated different ncRNA subtypes as different classes and trained multi-class models using the selected 10 features (Supplemental Figure S21). Some subtypes (e.g. rRNA and miRNA) could be better separated from other ncRNAs. In general, different subtypes of ncRNAs share similar feature patterns for the selected common features, which make it hard to separate them from each other. The common properties and feature patterns would enable the model to have the potential of finding novel ncRNA types.

Noncoding potential calculation

We calculated noncoding potential as the probability of being canonical ncRNAs for every bin (100nt) of human genome. We built models based on the selected 10 features and four classes (canonical ncRNA, CDS, UTR and negative control). The training set contains all bins in the gold standard set, i.e. 349 390 CDS bins, 617 150 UTR bins, 15 784 canonical ncRNA bins and 260 054 negative control bins. We used the following bagging method to build the models. Every time, 10 000 CDS bins, 10 000 canonical ncRNA bins, 10 000 UTR bins and 10 000 negative control bins were randomly sampled from the training set to build a model. Then the sampling method was repeated 100 times and 100 models were built. On average, every bin in the training set was used for approximately two times. The predicted values to be ncRNA of the 100 models were averaged as the noncoding potential score for each bin.

We calculate the noncoding potential scores for the other species (mouse, worm and fly) as well, using the same strategy.

RESULTS

Identifying common features of canonical ncRNAs across multiple species using *RNAfeature*

We developed an integrated feature selection method, *RNAfeature* (Figure 1, see 'Materials and Methods' section), to identify the essential features of various ncRNA types across multiple species, from >600 sets of high-throughput experimental data and various computational scores (Supplemental Table S1 and Supplemental Dataset 1). These datasets and computational scores constitute the raw features, and they were chosen based on their potential capability of distinguishing ncRNAs from three other types of genomic elements, namely CDSs, UTRs and inactive intergenic regions (negative control). Because most of the novel long ncRNAs are not well confirmed, we only used the well-classified ncRNAs (called canonical ncRNAs, including rRNA, tRNA, snRNA, miRNA, Y RNA, etc; see 'Materials and Methods' section) in the training set for the feature selection calculations. We supplied as inputs of *RNAfeature* the annotated sequence elements at each genomic locus, and the values of the raw features at each locus and its flanking upstream and downstream regions, which are useful in defining the genomic context of the locus (upper panel in Figure 2). *RNAfeature* then performed an intra-species round of feature selection to identify key features that distinguish ncRNAs from other sequence elements in each species (Figure 2), followed by an inter-species round to determine features generally useful in identifying ncRNAs in human, mouse, worm and fly.

Based on the accuracy of classifying the four types of genomic elements, 10 common features were finally selected for the four animal species, namely GC content, DNA sequence conservation, protein sequence conservation, RNA secondary structure homologs, reading frame property, small RNA-seq, poly(A)+ RNA-seq and poly(A)− RNA-seq signals, and the histone modification signals H3K36me3 and H3K4me3 (Figure 1, Supplemental Figure S20).

Canonical ncRNAs exhibit conserved patterns of the selected features across different species

To understand the significance of the 10 common features in identifying the different types of sequence elements, we used the multinomial logistic regression algorithm Softmax to illustrate their feature weights for the four genomic classes in each species (Figure 3A). The averaged feature values for the four classes are also shown in Supplemental Figures S22 and S23. In all four species, CDSs were generally more conserved and highly expressed, and the canonical ncRNAs tended to be more structured and had the strongest signals in the small RNA-seq data. H3K4me3 was more highly weighted in the UTRs, which could be caused by their enrichment in promoter regions (25). The inclusion of upstream and downstream features of each locus might help separate ncRNAs from UTRs, in that if a locus was close to a CDS, it would be more likely to be a UTR. We then mixed the data from all four species to build a single cross-species model. This model displays similar feature patterns as the four species-specific models (Figure 3A), further confirm-

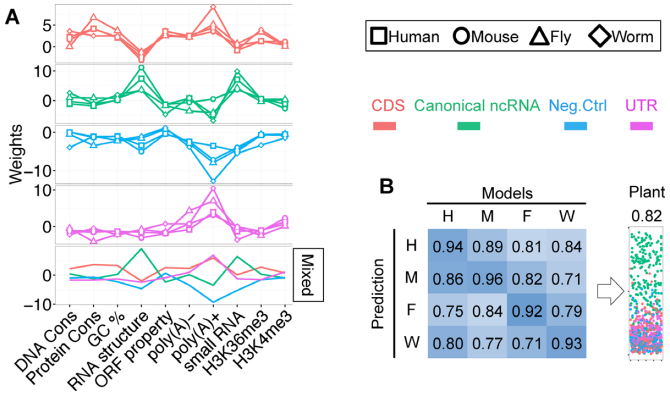


Figure 3. Conserved feature pattern of canonical ncRNAs in multiple species. (**A**) Weights of the 10 features in multinomial logistic regression models in distinguishing four kinds of genomic regions: coding sequence (CDS, red), canonical noncoding RNAs (ncRNAs, green), negative controls (Neg.Ctrl, blue) and untranslated regions (UTR, purple). In each plot, the weights learned from human data are marked by square, and similarly mouse by round, fly by triangle and worm by diamond. The plot at the bottom shows the weights from a mixed case that treats the genomic regions from four species the same. Quantile normalization was applied to compare and utilize data across species. Explanations on the feature labels are: DNA Cons for DNA sequence conservation; protein Cons for protein sequence conservation; GC% for GC content; RNA structure for RNA secondary structure homologs searched by *INFERNAL* (19); open reading frame (ORF) property for scores calculated by *RNAcode* (6). (**B**) The classification accuracies (i.e. ACC of distinguishing the four kinds of genomic regions) are measured across multiple species, which uses the data from one species as the training set and the data from a different species as the test set. The short names for the four cases are H for human, M for mouse, F for fly and W for worm. The mixed case trained the model on data from all four animal species, and tested on data for a plant, Arabidopsis, with results shown as a scatter plot at the right side in the same color code for the four kinds of genomic regions.

ing the consistency of the selected features across the four species.

For each of the four species, the 10 features consistently separated the four sequence element classes with above 90% accuracy (Figure 3B, diagonal cells). To test the robustness of the feature patterns, we then applied the model (a mathematical function that predicts the element class of a given locus based on its 10 features) obtained from each species to the other three species (data from different species were quantile normalized). The resulting accuracies were fairly high (70–90%, Figure 3B, off-diagonal cells), showing the generality of these models across the four species. We also applied the mixed model derived from the four animal species to the plant species *A. thaliana*. The accuracy (ACC) of classifying four genomic elements (CDS, canonical ncRNA, UTR and negative control) was 82% (Figure 3B, Supplemental Figure S24). We also calculated other statistics using canonical ncRNAs as positives and the other three classes as negatives (precision: 0.84, sensitivity: 0.85, specificity: 0.95). This high accuracy illustrates the potential of extending the use of the 10 common features beyond animal species, although in this work we focus on the four animal species.

The selected common features identify ncRNAs better than having all raw features of any single type

We next investigated whether the 10 features in the selected set were necessary and sufficient, by asking (i) whether these 10 diverse features could identify ncRNAs better than only one type of (expression, epigenomics or structural features) and (ii) whether having more features of these single types could improve their ability in identifying ncRNAs. Specifically, we compared the model constructed from the 10 selected features with models constructed from all raw features of any one of the three types. The results show that the 10 features identified ncRNAs better than these single-type features in terms of both sensitivity and specificity in all four species (Figure 4, Supplemental Table S7 and Supplemental Figure S25), even the single-type feature sets contained substantially larger number of raw features. These results show that it was the diversity and specific choices of the features, rather than the number of features, that contributed most to the ability of the 10 selected features in identifying ncRNAs.

The selected common features are capable of finding novel ncRNA types

Since most of the annotated ncRNAs supplied as inputs to *RNAfeature* were canonical ncRNAs, it is important to assess the robustness of the selected features against finding novel ncRNA types. We first performed this assessment using human data by omitting one type of canonical ncRNAs from the inputs of *RNAfeatue* (miRNA, rRNA, snRNA, snoRNA, tRNA, Y RNA or 7SK RNA), and asked *RNAfeature* to model feature patterns that can distinguish the remaining six ncRNA types from the other three element classes (CDS, UTR and negative control) (Supplemental Figure S26). The resulting models were able to separate the ncRNAs of the omitted sub-type from the other three element classes (average sensitivity: ∼0.89), even when the omitted ncRNA type was markedly different from the others (e.g. the sensitivity was ∼0.88 when rRNAs were the omitted sub-type) (Figure 5A).

We then performed this cross-type assessment in the other three species based on the annotated ncRNA types available in these species. Good accuracies were again observed (Figure 5B and Supplemental Figure S27). Comparing the four species, the accuracies of these cross-type assessments were slightly better in human and mouse, probably due to the larger number of annotated ncRNAs in these two species. In general, however, the accuracies in all four species were much higher than would be expected by random chance.

We also repeated these cross-type assessments by using SCI calculated by *RNAz* (20) to compute the RNA secondary structural feature instead of using *Rfam/INFERNAL* (26,27), to avoid having any of the 10 selected features also involving in the ncRNA annotations. The resulting accuracies were only slightly affected (average sensitivity ∼0.86 with SCI as compared to ∼0.89 with *Rfam/INFERNAL*) (Supplemental Table S8), showing that the models worked well with different definitions of structural scores.

Taken together, all these results show that the common features selected by *RNAfeature* capture general properties

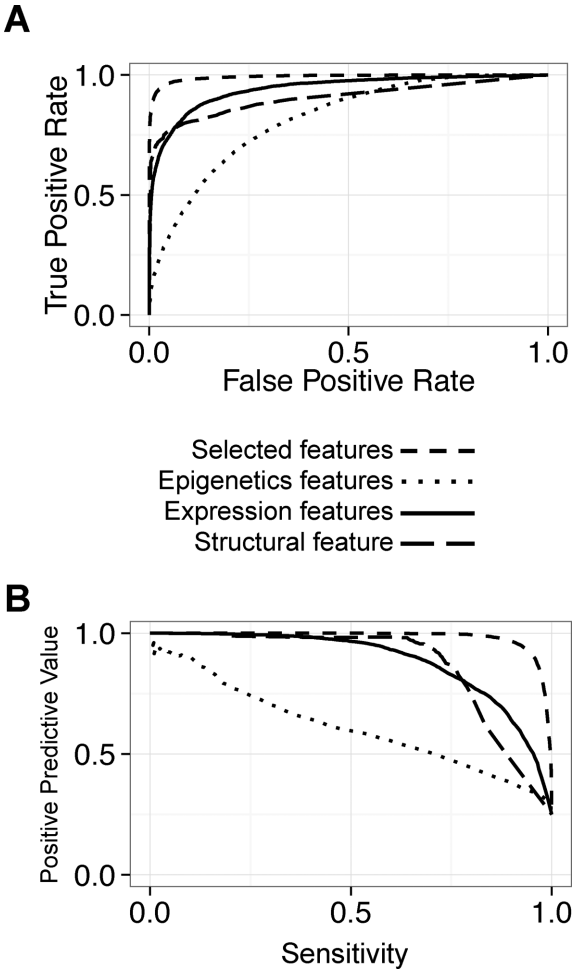


Figure 4. Performance of predicting canonical nRNAs based on different feature sets. (**A**) The performance of *RNAfeature* is evaluated by predicting known genomic regions with 100nt in size to belong to four classes (CDS, UTR, negative control, and canonical ncRNAs). After the 4-class prediction, ROC curve are drawn using canonical ncRNAs as positives, and the union of other three classes as negatives. We compared four cases of feature sets. Selected features refer to the 10 features chosen by *RNAfeature*. The epigenetic features include all histone modifications and TRF binding signatures. Expression features contain all expression profiles. Structural feature is the RNA secondary structure homologs searched by Rfam/INFERNAL. The number after each feature set gives the area under the corresponding ROC curve (AUC). (**B**) The performance of RNAfeature is showed by the sensitivity-precision curves.

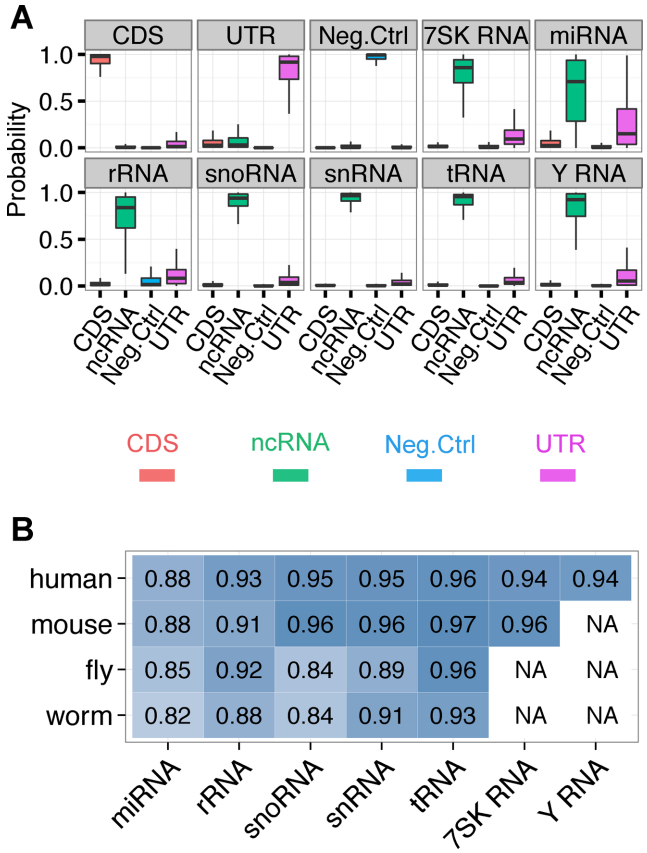


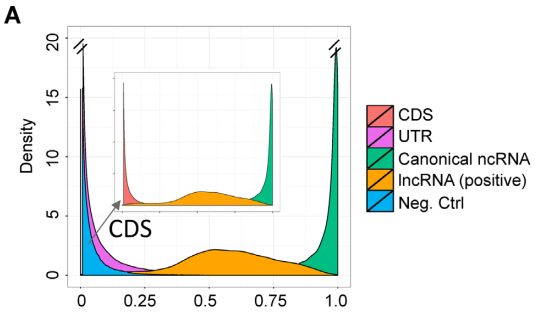
Figure 5. Performance of cross-type validations. (**A**) Boxplots show the probabilities (in y-axis, range: 0–1) of a certain type of genomic bins (as shown in the title of each window) being predicted to be CDS, UTR, ncRNA, or negative control (labeled in x-axis) in human genome using *RNAfeature*. For each specific type of ncRNA (e.g. rRNA) panel, we used different ncRNAs in training and test sets. The test set consists of one specific nRNA type (e.g. rRNA), and the training set consists of all other types of canonical ncRNAs (e.g. tRNA, miRNA, snRNA, etc.). The other three classes are the same (i.e. CDS, UTR and negative control) in each panel. (**B**) The accuracies (ACC) of cross-type predictions for the four classes are calculated for all four species.

of diverse ncRNA types even if some types are absent in its input data. The integrative model has the potential to find novel ncRNA types.

The selected common features define a noncoding potential score to identity confident regions on lncRNAs

Although the genomic regions are pervasively transcribed, only small portions (e.g. ∼1% in human) are annotated as canonical ncRNAs and long ncRNAs (lncRNAs) (1). Even for the annotated lncRNAs, most of them are not well confirmed. Many of them are simply filtered from the newly assembled transcripts based on coding potential scores (4–6). Moreover, because a lncRNA transcript tends to be flexible and long, its local motifs and structure may be essential for its biological and biophysical roles in a cell (28). Cases have been found in which a single transcript contains multiple domains that function differently (e.g. *MALAT1*) (29). While the 10 common features were identified based on canonical ncRNAs most of which are short (expect for rRNAs), we argued that that they could also help identify local regions on lncRNAs that share similar properties as the canonical ncRNAs such as secondary structures. These regions should have better confidence and have the potential possessing a noncoding function.

We took all bins (100nt) annotated in the genome, built 4-class models (canonical ncRNA, CDS, UTR and negative control) based on the selected 10 features, and predicted the probability of a bin to be canonical ncRNA in human, mouse, worm and fly. We call this probability noncoding potential. By integrating complementary features, the potential score is able to cover all transcribed regions (including unconserved genomic regions). Since the calculation of noncoding potential score is based on local bins, it does not require accurate assembly of the full-length transcripts,



B

Noncoding potential	RNA structure	small RNA	poly(A)-	poly(A)+	H3K4me3	H3K36me3	DNA Cons	Protein Cons	ORF property	GC %
0~0.1	0.00	0.02	2.03	4.60	0.08	0.07	0.14	0.17	0.01	0.53
0.1~0.2	0.00	0.02	1.69	2.22	0.08	0.06	0.10	0.14	0.00	0.53
0.2~0.3	0.00	0.03	1.12	1.30	0.06	0.05	0.09	0.12	0.00	0.52
0.3~0.4	0.00	0.05	3.98	1.84	0.05	0.05	0.09	0.11	0.00	0.51
0.4~0.5	0.01	0.07	1.37	0.93	0.05	0.05	0.09	0.11	0.00	0.52
0.5~0.6	0.02	0.10	0.95	0.75	0.04	0.05	0.09	0.10	0.00	0.51
0.6~0.7	0.02	0.39	0.94	0.52	0.04	0.05	0.09	0.08	0.00	0.50
0.7~0.8	0.03	0.81	0.15	0.31	0.03	0.05	0.08	0.08	0.00	0.50
0.8~0.9	0.08	0.44	0.37	0.28	0.03	0.05	0.08	0.08	0.00	0.49
0.9~1	0.33	0.76	0.06	0.04	0.03	0.06	0.07	0.15	0.00	0.50

Figure 6. Noncoding potential defines confident regions in human lncRNAs. (**A**) Noncoding potential scores are calculated for each type of genomic element at bin level (100nt), their densities are plotted. For each lncRNA bin, we predict its probabilities to be canonical ncRNA, CDS, UTR or negative control. If a bin has higher probability to be canonical ncRNA than other three classes, it is defined as positive (confident bin). (**B**) The feature values are averaged from all the lncRNA bins with different noncoding potential scores. Each expression score is the maximum RPKM across five cell-lines (GM12878, K562, H1-hESC, HeLa-S3 and HepG2). The other feature values are normalized from 0 to 1. We use the upstream *CIS* for H3K4me3 because it known as a promoter marker.

thereby avoiding the need for very deep RNA-seq, CAGE (30) or TIF-seq (31).

We then applied the noncoding potential on the annotated human lncRNAs (Gencode V10). A local region (100-nt bin) will be called confident region (positive) if its canonical ncRNA probability is higher than the probabilities of the other three classes. We found ~70% of the lncRNAs contains at least one confident bin having properties similar to canonical ncRNAs (Supplemental Table S9). The noncoding potential scores of confident lncRNA bins (positives) are usually between those of CDSs and canonical ncRNAs (Figure 6A), suggesting that these novel lncRNA regions have unique feature patterns that are different from any known genomic elements (e.g. CDS, UTR and canonical ncRNAs).

We further investigated the feature patterns of all the human lncRNA bins with different levels of noncoding potential (Figure 6B, Supplemental Figure S28). As expected, the lncRNA bins have lower conservation scores than CDSs (Supplemental Figures S22–S23). The high noncoding potential is mainly contributed by structural feature and/or small RNA signals. For instance, two lncRNAs with high noncoding potential scores (*MALAT1* and *NEAT1*/*Men β*) contain nonpoly(A) rich tract at their 3′-ends (29,32). They were processed into structured RNAs that generate small RNAs. This is also consistent with previous studies showing that some lncRNAs may be post-processed into smaller RNAs (28,33). In total, we identified 735 lncRNA transcripts (~10% of the total) containing conserved canonical structures (e.g. tRNA-like). These regions may function as regulatory structural motifs. The ratio of Poly(A)− RNA signal over Poly(A)+ RNA signal also tend to be higher when the noncoding potential increases. This is supported by a previous study showing that many lncRNAs (>24%) might lack poly(A) tails (34).

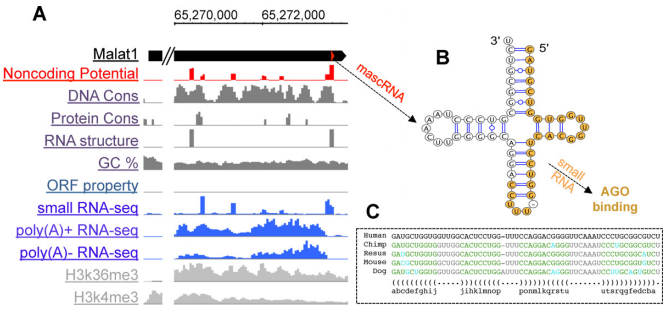


Figure 7. A lncRNA example showing the usage of 10 features, coding and noncoding potential. (**A**) The noncoding potential score, 10 features' signals/scores are displayed for a lncRNA (3′ end), *MATAT1*. (**B**) A structural domain, mascRNA, with high noncoding potential score is highlighted at the right panel. Argonaute proteins bind its small RNA product (yellow nucleotides). (**C**) The DNA sequence and RNA secondary structure of mascRNA are conserved across multiple species.

Application of the 10 selected features and noncoding potential scores

We use a well-studied lncRNA, *MALAT1* (metastasis associated lung adenocarcinoma transcript 1), to illustrate the biological and biophysical significance of the 10 features and noncoding potential scores (Figure 7). We first observe a strong H3K4me3 signal at 5′ end of *MALAT1*. This indicates that it is potentially regulated by certain histone modifications (Supplemental Figure S29). We further identify some locations with high noncoding potential scores at its 3′ end (Figure 7A). The 3′ end of *MALAT1* has been reported to be essential for promoting cell proliferation and invasion (35), and contains a functional motif, mascRNA (29). Notably, the mascRNA was reported broadly expressed in both human and mouse cell lines and normal tissues. Here, we rediscovered the mascRNA region in human *MALAT1* with a high noncoding potential score. The structural motif is a tRNA-like 61-nt cytoplasmic RNA (Figure 7B), which is conserved at both the DNA sequence and RNA secondary structure levels (Figure 7C). We also detected strong small RNA signals from this structural motif. Based on eight published CLIP-seq datasets, we found that the small RNAs were bound by the Argonaute proteins (AGO2 and AGO3) (36,37), which suggests that tRNA-like cytoplasmic mascRNA may generate some microRNA-like small RNAs. More supportively, tRNAs has been reported to be cleaved into small RNAs (tRNA-derived small RNAs, tsRNAs) by Dicer in cytoplasm (38), confirming the biological functions of Argonaute protein footprint on the small RNA derived from mascRNA.

This example illustrates that the classical structural conformation of canonical ncRNAs could be adopted by the novel long ncRNAs (e.g. *MALAT1*) to define novel functional domains. It is consistent with our *RNAfeature* model including RNA secondary structure conservation and homologous score.

DISCUSSION

We have developed a comprehensive computational method, *RNAfeature*, to evaluate and characterize novel ncRNAs with high resolution at the whole genome level. It revealed a set of distinct features commonly shared by different types of canonical ncRNAs in multiple species. This set includes 10 features from the sequence, structure, genome and epigenome merged from data obtained in various tissues, cell lines and development stages.

RNAfeature is an integrative method that is capable of distinguishing diverse ncRNA types from protein coding sequences and UTRs. The rigorous feature selection strategy of *RNAfeature* enabled us to select from >600 datasets a succinct set of informative features with clear evolutionary, biophysical and biological meanings. By integrating complementary features together and considering context influence (*CIS*), *RNAfeature* showed better sensitivity than ncRNA prediction methods based on single feature types (e.g. expression level and structure conservation; Figure 3 and Supplemental Figure S25). The patterns of these features are conserved in multiple species. We could use a model derived from animal species to predict plant ncRNAs. And the resulting ncRNA models are robust against missing ncRNA sub-types in the input data.

We further provided a set of noncoding potential scores, which can be used to identify, evaluate and characterize novel ncRNAs. By integrating complementary features, our method is able to cover all transcribed regions (including unconserved genomic regions). It also does not require the full-length transcripts being assembled, because the score calculation is based on bin level (100nt) (Supplemental Figure S30 and Supplemental Table S10). We used only the well-classified ncRNAs (canonical ncRNAs) to define the noncoding potential, because many annotated lncRNAs are not well validated. Ribosome was also observed on many lncRNAs (39), which indicates that some lncRNA regions may have coding potential. Therefore, our noncoding potential score might not predict some novel types of ncRNAs, because it was trained on the feature pattern of certain types of ncRNAs. Still, we have demonstrated that we could identify local regions with certain noncoding potential scores on ∼70% of the annotated human lncRNA transcripts (Figure 6, Supplemental Table S9).

One drawback of *RNAfeature* is that it takes a large amount of input data, which require a fairly large amount of pre-processing work. We therefore provide the 10 selected features and their computed values along the 4 transcriptomes (human, mouse, worm and fly) as a resource for future studies of lncRNAs, eliminating the need for repeating all the data processing work we have performed.

Based on our results in model organisms, we also suggest the 10-feature data to be curated and sequenced for the novel ncRNA identification, evaluation and characterization in other species. In the future, as more data that provide novel information about ncRNAs are produced, such as ribosome profiling (39) and whole genome structure profiling (40,41), we will keep updating the set of common features accordingly and explore other potential use of them.

AVAILABILITY

All the data (i.e. noncoding potential scores) are integrated as a unified resource at http://Rnafeature.ncrnalab.org/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

We thank Jiapei Yuan, Yang Wu, Boqin Hu, Yang Yang, Yuchuan Wang, Le Xu, Weiyi Li, Zhengyu Liang, Peipei Yin, Zhihang Fan, Hui Zheng, Junpei Umetsu and Xinqiang Ding for downloading and processing the raw data. We thank Jingyi Li, Yang Yang, Yu Liu, Luyun Wu for their help on editing the manuscript.

FUNDING

National Key Basic Research Program [2012CB316503]; National High-tech Research and Development Program of China [2014AA021103]; National Natural Science Foundation of China [31271402, 31100601, 91019016, 31361163004]; National Institutes of Health [HG001696, ES017166 to M.Q.Z.]; Hong Kong Research Grants Council Early Career Scheme [419612 to K.Y.]. Funding for open access: National Natural Science Foundation of China [31271402].

Conflict of interest statement. None declared.

REFERENCES

1. Gerstein,M.B., Rozowsky,J., Yan,K.-K., Wang,D., Cheng,C., Brown,J.B., Davis,C.A., Hillier,L., Sisu,C., Li,J.J. *et al.* (2014) Comparative analysis of the transcriptome across distant species. *Nature*, **512**, 445–448.
2. Fatica,A. and Bozzoni,I. (2014) Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.*, **15**, 7–21.
3. Lee,C. and Kikyo,N. (2012) Strategies to identify long noncoding RNAs involved in gene regulation. *Cell Biosci.*, **2**, 37.
4. Kong,L., Zhang,Y., Ye,Z.Q., Liu,X.Q., Zhao,S.Q., Wei,L. and Gao,G. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.*, **35**, W345–349.
5. Lin,M.F., Jungreis,I. and Kellis,M. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**, i275–282.
6. Washietl,S., Findeiss,S., Muller,S.A., Kalkhof,S., von Bergen,M., Hofacker,I.L., Stadler,P.F. and Goldman,N. (2011) RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA*, **17**, 578–594.
7. Guttman,M., Amit,I., Garber,M., French,C., Lin,M.F., Feldser,D., Huarte,M., Zuk,O., Carey,B.W., Cassady,J.P. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.
8. Guttman,M. and Rinn,J.L. (2012) Modular regulatory principles of large non-coding RNAs. *Nature*, **482**, 339–346.

9. Lu,Z.J., Yip,K.Y., Wang,G., Shou,C., Hillier,L.W., Khurana,E., Agarwal,A., Auerbach,R., Rozowsky,J., Cheng,C. *et al.* (2011) Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res.*, **21**, 276–285.
10. Lv,J., Liu,H., Huang,Z., Su,J., He,H., Xiu,Y., Zhang,Y. and Wu,Q. (2013) Long non-coding RNA identification over mouse brain development by integrative modeling of chromatin and genomic features. *Nucleic Acids Res.*, **41**, 10044–10061.
11. Ramos,A.D., Diaz,A., Nellore,A., Delgado,R.N., Park,K.Y., Gonzales-Roybal,G., Oldham,M.C., Song,J.S. and Lim,D.A. (2013) Integration of genome-wide approaches identifies lncRNAs of adult neural stem cells and their progeny in vivo. *Cell Stem Cell*, **12**, 616–628.
12. Gerstein,M.B., Lu,Z.J., Van Nostrand,E.L., Cheng,C., Arshinoff,B.I., Liu,T., Yip,K.Y., Robilotto,R., Rechtsteiner,A., Ikegami,K. *et al.* (2010) Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*, **330**, 1775–1787.
13. Granitto,P.M., Furlanello,C., Biasioli,F. and Gasperi,F. (2006) Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemom. Intell. Lab. Syst.*, **83**, 83–90.
14. Harikumar,G., Couvreur,C. and Bresler,Y. (1998) Acoustics, speech and signal processing. *Proceedings of the 1998 IEEE International Conference on IEEE*, Vol. **3**, pp. 1877–1880.
15. Cabili,M.N., Trapnell,C., Goff,L., Koziol,M., Tazon-Vega,B., Regev,A. and Rinn,J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.
16. Di,C., Yuan,J., Wu,Y., Li,J., Lin,H., Hu,L., Zhang,T., Qi,Y., Gerstein,M.B., Guo,Y. *et al.* (2014) Characterization of stress-responsive lncRNAs in Arabidopsis thaliana by integrating expression, epigenetic and structural features. *The PlantJ ournal*, **80**, 848–861.
17. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
18. Bonnet,E., Wuyts,J., Rouze,P. and Van de Peer,Y. (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, **20**, 2911–2917.
19. Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.
20. Gruber,A.R., Findeiss,S., Washietl,S., Hofacker,I.L. and Stadler,P.F. (2010) Rnaz 2.0: improved noncoding RNA detection. *Pac. Symp. Biocomput.*, **15**, 69–79.
21. Giannopoulou,E.G. and Elemento,O. (2013) Inferring chromatin-bound protein complexes from genome-wide binding assays. *Genome Res.*, **23**, 1295–1306.
22. Duret,L. and Galtier,N. (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet.*, **10**, 285–311.
23. Wang,J., Zhuang,J., Iyer,S., Lin,X., Whitfield,T.W., Greven,M.C., Pierce,B.G., Dong,X., Kundaje,A., Cheng,Y. *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798–1812.
24. Johnsson,P., Lipovich,L., Grander,D. and Morris,K.V. (2014) Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim. Biophys. Acta*, **1840**, 1063–1071.
25. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.K., Koche,R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
26. Gardner,P.P., Daub,J., Tate,J.G., Nawrocki,E.P., Kolbe,D.L., Lindgreen,S., Wilkinson,A.C., Finn,R.D., Griffiths-Jones,S., Eddy,S.R. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**, D136–D140.
27. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
28. Derrien,T., Johnson,R., Bussotti,G., Tanzer,A., Djebali,S., Tilgner,H., Guernec,G., Martin,D., Merkel,A., Knowles,D.G. *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.*, **22**, 1775–1789.
29. Wilusz,J.E., Freier,S.M. and Spector,D.L. (2008) 3′ end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell*, **135**, 919–932.
30. Shiraki,T., Kondo,S., Katayama,S., Waki,K., Kasukawa,T., Kawaji,H., Kodzius,R., Watahiki,A., Nakamura,M., Arakawa,T. *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 15776–15781.
31. Pelechano,V., Wei,W. and Steinmetz,L.M. (2013) Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*, **497**, 127–131.
32. Sunwoo,H., Dinger,M.E., Wilusz,J.E., Amaral,P.P., Mattick,J.S. and Spector,D.L. (2009) MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res.*, **19**, 347–359.
33. Nam,J.W. and Bartel,D.P. (2012) Long noncoding RNAs in C. elegans. *Genome Res.*, **22**, 2529–2540.
34. Yang,L., Duff,M.O., Graveley,B.R., Carmichael,G.G. and Chen,L.L. (2011) Genomewide characterization of non-polyadenylated RNAs. *Genome Biol.*, **12**, R16.
35. Xu,C., Yang,M., Tian,J., Wang,X. and Li,Z. (2011) MALAT-1: a long non-coding RNA and its important 3′ end functional motif in colorectal cancer metastasis. *Int. J. Oncol.*, **39**, 169–175.
36. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M. Jr, Jungkamp,A.C., Munschauer,M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
37. Kishore,S., Jaskiewicz,L., Burger,L., Hausser,J., Khorshid,M. and Zavolan,M. (2011) A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods*, **8**, 559–564.
38. Haussecker,D., Huang,Y., Lau,A., Parameswaran,P., Fire,A.Z. and Kay,M.A. (2010) Human tRNA-derived small RNAs in the global regulation of RNA silencing. *RNA*, **16**, 673–695.
39. Ingolia,N.T. (2014) Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.*, **15**, 205–213.
40. Ding,Y., Tang,Y., Kwok,C.K., Zhang,Y., Bevilacqua,P.C. and Assmann,S.M. (2014) In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*, **505**, 696–700.
41. Yang,Y., Umetsu,J. and Lu,Z.J. (2014) Global signatures of protein binding on structured RNAs in Saccharomyces cerevisiae. *Sci. China. Life Sci.*, **57**, 22–35.