Bioinformatics, 31(24), 2015, 3897–3905
doi: 10.1093/bioinformatics/btv480
Advance Access Publication Date: 26 August 2015
Original Paper

OXFORD

Sequence analysis

LncRNA-ID: Long non-coding RNA IDentification using balanced random forests

Rujira Achawanantakun, Jiao Chen, Yanni Sun* and Yuan Zhang

Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on November 17, 2014; revised on July 8, 2015; accepted on August 7, 2015

Abstract

Motivation: Long non-coding RNAs (lncRNAs), which are non-coding RNAs of length above 200 nucleotides, play important biological functions such as gene expression regulation. To fully reveal the functions of lncRNAs, a fundamental step is to annotate them in various species. However, as lncRNAs tend to encode one or multiple open reading frames, it is not trivial to distinguish these long non-coding transcripts from protein-coding genes in transcriptomic data.

Results: In this work, we design a new tool that calculates the coding potential of a transcript using a machine learning model (random forest) based on multiple features including sequence characteristics of putative open reading frames, translation scores based on ribosomal coverage, and conservation against characterized protein families. The experimental results show that our tool competes favorably with existing coding potential computation tools in lncRNA identification.

Availability and implementation: The scripts and data can be downloaded at https://github.com/zhangy72/LncRNA-ID

Contact: yannisun@msu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

It has been suggested that less than 2% of the human genome codes for proteins (Pennisi, 2012). The majority of the transcriptome contains non-coding RNA (ncRNA) genes (Djebali *et al.*, 2012), which function directly as RNAs instead of coding for proteins (Pauli *et al.*, 2011; Wilusz *et al.*, 2009). The most recently discovered class of ncRNAs is long non-coding RNAs (lncRNAs), which are generally defined as non-coding transcripts of length above 200 nucleotides (Bu *et al.*, 2012; Derrien *et al.*, 2012; Hung and Chang, 2010). Increasing evidence has shown that lncRNAs play important and diverse biological functions. For example, lncRNAs ANRIL and HOTAIR bind to chromatin-remodelling complexes PRC1 and PRC2 to alter chromatin and transcription. GAS5 lncRNA acts as a decoy for the GR transcription factor and prevents GR from binding to DNA and transcriptional activation. MALAT1 RNA binds to SR proteins to regulate mRNA alternative splicing, whereas BACE-1AS RNA binds to the complementary BACE-1 mRNA to regulate BACE-1 translation. As a result, the dysfunctions of lncRNAs are associated with a wide range of diseases ranging from neurodegeneration to cancer (Wapinski and Chang, 2011).

LncRNAs exist in many species such as Arabidopsis (Liu *et al.*, 2012), Zea mays (Boerner and McGinnis, 2012), honey bee (Humann *et al.*, 2013), chicken (Arriaga-Canon *et al.*, 2014) and zebrafish (Liu *et al.*, 2013). In recent years, a large number of lncRNAs have been identified. GENCODE (Derrien *et al.*, 2012) comprises 9277 manually annotated lncRNA genes in the human genome. The LncRNADisease database (Chen *et al.*, 2013) contains 1564 human lncRNAs that are likely to be associated with diseases. Thus, given the functional importance and ubiquity of lncRNAs, it is important to annotate them on a genome scale in various species.

With the advances of the next-generation sequencing technologies, the transcriptomes of a large number of organisms have been sequenced, providing us a unique opportunity to mine lncRNAs. The assembled transcripts contain different types of functional elements such as small ncRNAs, lncRNAs and protein-coding genes. LncRNAs can be effectively distinguished from most small ncRNAs, such as

miRNAs and snoRNAs, using size as the main criterion. However, lncRNAs share similarities with protein-coding transcripts in terms of transcript length and splicing structure (Guttman *et al.*, 2013). In addition, lncRNAs tend to encode putative open reading frames (ORFs). For instance, H19, Xist, Mirg, Gtl2 and KcnqOT1 all have putative ORFs greater than 100 amino acids, but have been characterized as functional ncRNAs (Prasanth and Spector, 2007). Thus, a major challenge for lncRNA identification is to distinguish lncRNAs from protein-coding genes, especially in non-model organisms lacking comprehensive protein-coding gene annotation.

Many efforts have been made to distinguish between lncRNA and protein-coding transcripts, ranging from applying a threshold for a single feature to more complicated supervised machine learning methods. One commonly used feature is the length of the ORF. For example, a simple approach is to classify a transcript containing an ORF of length above 100 amino acids as a protein-coding gene (Okazaki *et al.*, 2002). This criterion is arbitrary and is not always correct (Dinger *et al.*, 2008). Using this simple criterion, the mouse Xist RNA gene (Brockdorff *et al.*, 1992), which encodes a putative ORF of 298 amino acids (aa), was mis-classified as a protein-coding gene when it was first discovered (Borsani *et al.*, 1991). A probability model constructed from multiple criteria has also been used to identify protein-coding and non-coding regions such as the one used in the Phylogenetic Codon Substitution Frequencies (PhyloCSF). PhyloCSF is based on nucleotide substitutions of multispecies sequences. It uses multiple sequence alignments to calculate the phylogenetic conservation score. As an alignment-based method, it usually requires high quality alignments (Schloss, 2010), which are not trivial to produce and can incur high computational cost.

More accurate approaches for identifying lncRNAs use supervised machine learning methods. These approaches can be further divided into two types. One relies on sequence alignments and the other is alignment-free. Representative examples of alignment-based methods include Coding-Potential Calculator (CPC) (Kong *et al.*, 2007). CPC aligns transcripts against known protein databases. As homologous protein-coding genes tend to share higher sequence conservation than lncRNAs, the alignment score or its statistical significance provides useful information to differentiate these two types of transcripts (Chodroff *et al.*, 2010; Marchler-Bauer *et al.*, 2013; Marchler-Bauer and Bryant, 2004).

An alternative and faster approach is alignment-free methods such as CPAT (Wang *et al.*, 2013) and PLEK (Li *et al.*, 2014). CPAT integrates linguistic features of transcript sequences into a logistic regression model for lncRNA prediction. PLEK integrates k-mer based features into a support vector machine to distinguish lncRNAs from mRNAs. In addition, CPAT and PLEK allow users to create a model with their own data. This option, which is not present in CPC and PhyloCSF, is very useful for lncRNA identification in different species.

Despite the promising progress for lncRNA identification, there is still a need for better approaches and tools. In particular, existing machine learning-based tools do not carefully handle the imbalanced training data, in which one class has far more instances than the other. The issue of imbalanced training data is particularly pronounced for lncRNA identification when it is formulated as a binary classification problem in existing tools. For example, due to poor annotation of lncRNAs, many species have far less characterized lncRNAs than protein-coding genes. As a result, a classifier tends to over-predict query transcripts as the protein-coding transcript (major class) (Probost, 2000). In addition, many existing tools need users to provide a score threshold for lncRNA identification, which is not always obvious from users' perspective. For example, PhyloCSF and CPAT do not suggest the specific type of an input transcript, but only output a coding potential score. The predefined score cutoffs of PhyloCSF and CPAT vary from species to species. PhyloCSF's score cutoffs of 50 and 300 were used for mouse (Guttman *et al.*, 2010) and Zebrafish (Pauli *et al.*, 2012), respectively. CPAT suggests the score cutoffs of 0.364 and 0.44 for human and mouse, respectively. These specific score cutoffs cannot be immediately applied to other species. Even worse, not every tool can be trained on different species to provide users with necessary information to choose an appropriate score cutoff.

In this article, we present LncRNA-ID, an lncRNA identification tool, which applies random forest (RF) classification (Breiman, 2001) to distinguish lncRNAs from protein-coding genes. RF is a classification model aggregating multiple classification trees generated from boot-strap samples and has been successfully applied in bioinformatics (Chen and Ishwaran, 2012; Leung *et al.*, 2013; Wu *et al.*, 2009). LncRNA-ID has several advantages over existing tools. First, it still takes advantage of alignment-based features, which have strong discriminative power. However, instead of using genome-scale multiple sequence alignments or pairwise alignments against all existing protein sequences, LncRNA-ID uses profile hidden Markov model (profile HMM)-based alignments, rendering more sensitive homology search and shorter running time than existing alignment-based lncRNA identification tools. Second, LncRNA-ID is easy to use as it does not require users to provide a score cutoff. It automatically determines the type of a query transcript as well as provides a coding potential score. Third, LncRNA-ID can be applied to various species by providing an option to train the classifier for different data. Fourth, LncRNA-ID does not require a large number of training data of neither protein-coding transcripts nor lncRNAs to construct a classifier and can handle imbalanced classes in the training data. In our experiments, we evaluated the performance of LncRNA-ID on two different species, human and mouse. In addition, we benchmarked LncRNA-ID with CPC, CPAT, PhyloCSF and PLEK on both species. The experimental results show that lncRNA-ID has both good sensitivity and specificity.

2 Materials and methods

In this section, we first talk about the features used in LncRNA-ID and then describe the construction of our classification model. The features used in LncRNA-ID are derived from three different groups: open reading frame (ORF), ribosome interaction and protein conservation. Each feature is selected either based on the literature or the empirical observations. We will show that using multiple features can significantly improve the performance of classification.

2.1 ORF features

ORF is one of the most commonly used criteria to distinguish an lncRNA from a coding transcript. A true protein-coding transcript tends to have longer ORFs than those in lncRNAs. We derive two ORF-related features: ORF length and ORF coverage. The ORF length is defined as the length of the longest reading frame identified in three forward frames. The ORF coverage is defined as the ratio of the length of the chosen ORF to the length of the transcript. LncRNAs tend to have shorter ORF and lower ORF coverage than protein-coding transcripts. Supplementary Figure S1 compares the distribution of ORF-related features between lncRNAs and coding transcripts.

2.2 Ribosome interaction features

These features are based on the interaction mechanism between the ribosome and mRNAs during protein translation (Shaw, 2008).

Ribosomes consist of two parts, a large subunit where two tRNA binding sites are located and a small subunit where the mRNA binding site is located. The translation is initiated when the small ribosomal subunit attaches to the mRNA at a start codon. The ribosome starts to translate the mRNA towards the 3′ end until it encounters a stop codon. At the end of the protein translation, termination factors release the synthesized protein for use in the cell and the ribosome is split back into large and small subunits. Many studies have successfully applied ribosome footprint to identify functional proteins (Arava *et al.*, 2003; Guttman *et al.*, 2013; Ingolia *et al.*, 2011; Xing *et al.*, 2009). In particular, ribosome profiling (Ingolia *et al.*, 2011), which sequences mRNA fragments bound to ribosomes, provides a quantitative snapshot of protein translation. However, the availability of ribosome profiling data is still limited. Thus, in this work, we design computational features to quantify the main attributes related to ribosome interaction with mRNAs. We define the features according to these interaction states: initiation, translation and termination.

2.2.1 Initiation

The initiation interaction features are derived from the Kozak motif. The Kozak consensus is a favorable motif for a ribosome scanning pattern and initiates translation. It greatly impacts protein translation efficiency (De Angioletti *et al.*, 2004; Kozak, 1989; Xu *et al.*, 2010). Kozak motif has the consensus GCCRCCAUGG (R represents purine) and is located in the region around the initiator codon of an ORF. In the Kozak determining experiment, single base mutants are performed on mRNAs and the protein productions of the mutant sequences are measured. It has been demonstrated that nearly all ribosomes will initiate at the start codon (Kozak, 1999), AUG. The highly conserved nucleotides at positions −3 and +4 (the A of AUG is +1) and −2 and −1 play a major role in the initiation of the translation process.

We thus derive two features from Kazak motif: the nucleotides at the positions {−3, +4} and {−2, −1}. The Kozak features determine the potency of a starting site. A strong starting site, which enhances the translation efficiency, occurs when nucleotides at these positions are conserved, whereas a less conservation indicates a weak starting site (Kozak, 1997, 1999).

2.2.2 Translation

The interaction between the 3′ end of rRNAs and mRNA transcripts exhibits changes of binding energy along the transcript. The binding energy consists of the free energy needed to open the binding site and the energy gained from hybridization. We use RNAup (Muckstein *et al.*, 2006) to compute the thermodynamics of the interaction between the 3′ end of 18S rRNA and a transcript.

To capture the change of the binding energy, we calculated a series of binding energies between the 3′ end of the 18S rRNA and a transcript by moving the 3′ end of 18S rRNA toward the 3′ direction of a transcript one nucleotide at a time. Let δ_i be the free energy at position i. Let N_i denote the number of Watson-Crick base pairs of binding starting at position i. The ribosome coverage is thus defined as:

$$\text{ribosome coverage} = \sum_{i=1}^{L} \{N_i | \delta_i < 0\},$$

where L is the sequence length.

The ribosome coverages were computed on three regions: the whole transcript, ORF and 3′UTR. These three features illustrate the level of ribosome occupancy on a sequence. For protein-coding transcripts, we expect to see higher ribosome coverage on the whole transcript and the ORF region.

2.2.3 Termination

We define the ribosome release score (RRS) to capture a ribosome release signal. The RRS takes advantage of the fact that ribosomes are released when reaching a stop codon. As a result, a sharp drop in ribosome occupancy can be observed at the start of the 3′UTR of coding transcripts. In contrast, translational termination should not occur in non-coding transcripts (Guttman *et al.*, 2013; Vasquez *et al.*, 2014).

The RRS is laboratorially measured using the quantitative sequences from a deep sequencing of ribosome-protected mRNA fragments called ribosome profiling (Ingolia *et al.*, 2011). Although ribosome profiling is decently used as a quantitative snapshot of protein translation, it requires additional efforts on a laboratory experiment. Therefore, it is currently not widely available. However, it is expected to become more widely available with the demand from research communities and the progress in cost-effective sequencing technologies.

In the absence of ribosome profiling data, we estimate RRS as the ratio of ribosome coverage in the putative ORF to ribosome coverage in the corresponding 3′UTR.

$$\text{RRS} = \frac{\text{Ribosome coverage of ORF}/\text{length(ORF)}}{\text{Ribosome coverage of 3′UTR}/\text{length(3′UTR)}},$$

where RRS indicates the relative degree of ribosome occupancy bias at the terminal binding site in a sequence. True protein-coding transcripts are expected to have larger RRS than non-coding transcripts.

2.3 Protein conservation features

True protein-coding transcripts tend to show better conservation against characterized proteins. We measure the conservation using profile hidden Markov model (profile HMM)-based alignment scores. In particular, we chose HMMER (Finn *et al.*, 2011) to align a transcript against all available protein families, such as the ones in Pfam (Punta *et al.*, 2012). Applying profile-based homology search has several advantages, compared with pairwise alignment (Durbin *et al.*, 1998; Zhang *et al.*, 2013, 2014; Zhang and Sun, 2011). First, the number of gene families is significantly smaller than the number of sequences, rendering a much smaller search space. For example, there are only about 14 000 manually curated protein families in Pfam (Punta *et al.*, 2012). However, they cover nearly 80% of the UniProt Knowledgebase (Magrane and Consortium, 2011) and the coverage is increasing every year as enough information becomes available to form new families (Punta *et al.*, 2012). As the profile-based homology search tool HMMER is as fast as BLAST (Eddy, 2009), using profile-based search provides a shorter search time. In addition, alignments of query sequences against each protein family are independent from each other and thus can be naturally parallelized on high performance computing platforms. Second, previous work has shown that using family information can improve the sensitivity of remote protein homology search (Durbin *et al.*, 1998; Zhang and Sun, 2012). For the transcriptomes of non-model organisms, sensitive remote homology search is especially important for identifying possibly new homologs.

Specifically, each transcript is aligned to all protein families using HMMER. We use 0.1 as the E-value cutoff for HMMER. When more than one alignment is generated for a query sequence, the alignment with the best E-value is used. For the chosen alignment for a transcript, we derive the following three features: (i) the score, (ii) the length of aligned region in the query sequence and (iii) the length of the profile in the alignment. A true protein-coding transcript is likely to produce an alignment with higher score and longer alignments than lncRNAs.

In total, we extract 11 features: ORF length, ORF coverage, two Kozak motif-related features, ribosome coverage on three

regions: transcript, ORF and 3′UTR, ribosome release score, alignment score, alignment length with respect to profile HMM, and the transcript. Supplemental Figures S1–S7 show that although each feature exhibits different value distributions for the two types of transcripts, none of the features is able to fully distinguish lncRNAs from coding transcripts. The importance scores of features based on impurity reduction in decision trees (Tuv *et al.*, 2009) are also provided in Supplementary Table S8. Thus, it is important to combine multiple features to maximize the discriminative power. We formalize this problem as a binary classification problem where lncRNAs are defined as the positive class and protein-coding transcripts are defined as the negative class. All these features will be incorporated into the chosen classification model: balanced random forest, which we will describe below.

2.4 Balanced random forest

A decision tree is a commonly used classification model in machine learning. Random forest (RF) consists of multiple decision trees. Each decision tree is built from a bootstrap sample, which is a random sample drawn from the training data. During prediction, RF outputs the class agreed by most of the individual trees. We select the RF for the following reasons:

1. It is able to effectively handle missing data, which is common in lncRNA identification. For example, some lncRNA transcripts do not have protein conservation and the features such as alignment score or alignment length could be missing.
2. It natively supports categorical features without requiring any transformation. The typical conversion for categorical data is to create dummy binary variables to represent each category value. However, this may decrease the predictive power of the features and is time-consuming because of the potentially large number of dummy features. With RF, we are able to directly use Kozak motif features without the need for conversion.

Inspired by Chen et al. (2004), we extended RF to balanced random forest (BRF), which contains multiple RFs and each RF is built from a subset of the training data. BRF enables LncRNA-ID to learn from the imbalanced training data where the numbers of lncRNA and protein-coding samples are highly different. Imbalanced training data is common for lncRNA identification. A recent study found that lncRNAs are at least four times more than protein-coding genes in the human genome (Kapranov *et al.*, 2007). In practice, the majority class in the training data is protein-coding transcript because there are more protein-coding gene annotation than lncRNA annotation for most organisms. For example, in the GENCODE database (Derrien *et al.*, 2012), there are 12 526 annotated lncRNAs and 95 099 annotated coding transcripts in the human genome. For the mouse genome, there are 6053 annotated lncRNAs and 47 394 annotated coding transcripts in GENCODE. Thus, there is a need for a classification method that can effectively learn from imbalanced training data where one of the two classes has more samples (majority) than the other class (minority).

When learning from imbalanced training data using RF, there is a high possibility that a bootstrap sample contains very few or even none of the entities in the minority class, resulting in a classification tree with poor performance for predicting the minority class. A naive solution is to either conduct prior over-sampling of the minority class or prior down-sampling of the majority class. Down-sampling usually has a better performance over over-sampling (Huang *et al.*, 2008). However, a prior down-sampling of the majority class may result in loss of information, as a large part of the majority class is not used. In contrast to prior down-sampling, LncRNA-ID ensembles multiple RF classifiers induced from multiple balanced down-sampled data. This allows us to achieve better classification performance by maximizing the benefits of using abundant protein-coding data.

Our BRF is different from (Chen *et al.*, 2004) in that instead of creating balanced training subsets using random drawings, we divide the majority class into equal subsets according to the imbalanced ratio, which is the ratio of the size of the majority class to the size of the minority class. The purpose is to maximize the predictive power by ensuring that all training data are incorporated in constructing the classification model. The balanced random forest learning algorithm is shown in Supplementary Algorithm S11.

To create a balanced training data, down-sampling is performed on protein-coding transcripts, creating approximately an equal number of protein-coding and lncRNA transcripts in each subset n_i. Each training subset is then used to create an individual RF. Finally, we integrate all constructed RF classifiers into the BRF. The BRF classifier is then used to predict the type of a query transcript by aggregating the prediction results of ensemble classifiers.

Integrated with balanced random forest methodology using different types of features, LncRNA-ID has the following advantages: (i) can effectively handle limited or imbalanced learning data, which are commonly found in most species; (ii) incorporates different types of features, minimizing bias from a particular group of features. We employ the random forest classification implemented in Weka (Hall *et al.*, 2009) software package to construct our classification model. The optimal number of trees used in the random forest classification is determined based on the best performance obtained by 10-fold cross validation.

3 Results

LncRNA-ID can be applied to different species. We evaluated the performance of LncRNA-ID on both human and mouse. Specifically, we used four datasets. The first human dataset (H1) and the mouse dataset (M1) were generated from GENCODE consortium (Derrien *et al.*, 2012) within the framework of the ENCODE project. GENCODE is known to have the most comprehensive annotation of long non-coding RNAs available to date. The second human dataset (H2) is CPAT's original dataset generated from multiple resources: RefSeq (Pruitt *et al.*, 2007), a human lncRNA catalog (Cabili *et al.*, 2011) and GENCODE. We further conducted four additional experiments simulating training data with different imbalanced ratios to demonstrate that LncRNA-ID was able to maintain robust performance with imbalanced training data. Finally, we tested lncRNA-ID on a small but experimentally verified lncRNA dataset from mouse (M2).

To quantify the classification performance, we used five standard metrics: sensitivity, specificity, accuracy, false positive rate (FPR), positive predictive value (PPV) and F-score, which are defined as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}},\ \text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}}$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{P}+\text{N}},\ \text{FPR} = \frac{\text{FP}}{\text{FP}+\text{TN}},\ \text{PPV} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{F}-\text{score} = \frac{2\cdot\text{Sensitivity}\cdot\text{PPV}}{\text{Sensitivity}+\text{PPV}}$$

LncRNAs are regraded as the positive class and protein-coding transcripts are regarded as the negative class. TP is the number of

correctly classified lncRNAs and TN is the number of correctly classified protein-coding transcripts. Sensitivity is the proportion of correctly classified lncRNAs in the set of all lncRNAs. Specificity is the proportion of correctly classified protein-coding transcripts in the set of all protein-coding transcripts. Accuracy is the ratio of correctly classified transcripts in all predictions. False positive rate (FPR) refers to the portion of falsely classified lnRNAs among all protein-coding transcripts. PPV is a ratio of true positives to combined true and false positives. F-score is the harmonic mean of sensitivity and PPV and hence can be used as a single measure for the overall classification performance.

3.1 Performance of different groups of features on the human data (H1)

Supplemental Figures S1–S7 show that a single feature is not able to distinguish lncRNAs from protein-coding transcripts. We further evaluate the discriminative power of combined feature sets. We constructed this experiment using the human dataset (H1). This training dataset contains 15 308 protein-coding transcripts and 4586 lncRNAs randomly selected from GENCODE (Derrien *et al.*, 2012). The testing data contains 4000 coding transcript and 4000 lncRNAs. In the training set, only one transcript from each gene is used. Moreover, the training and testing datasets do not have transcripts from the same genes.

To directly test the performance of feature combination, we created the classification models using each group of features only or their combinations. The performance of each classification model was then evaluated on the testing data. The overall performance was measured by the area under ROC curve (AUC). AUC is a commonly used method to evaluate performances at all cutoff points, giving better insight into how well the classifier is able to separate the two classes. The greater the AUC is, the better overall classification performance the classifier achieves. The optimal performance is defined as the optimal operating point (FPR and sensitivity) on ROC curve that makes the result of a binary classification as close to a perfect predictor, where FPR = 0 and sensitivity = 1, as possible (Gonen, 2007). We employed the commonly used function *perfcurve* in Matlab (MATLAB, 2010) to compute the optimal operating point of a classifier.

Figure 1 shows the performance of LncRNA-ID using a single group of features versus multiple groups of features. The three groups of features exhibit highly different performance. The ribosome interaction features have the best discriminative power. According to Figure 1, combination of multiple groups of features leads to better performance than using a single group of features. In particular, combining these groups of features leads to the best performance and thus will be used in all our experiments.

We compared the performance of LncRNA-ID with several popular coding-potential or lncRNA prediction tools: CPC, CPAT, PhyloCSF and PLEK. These tools output the coding-potential of a transcript and can be used to classify a query transcript into coding or non-coding sequences. Below we present the experimental results of applying LncRNA-ID and the benchmark tools on three datasets. For each dataset, we introduce the important parameters used for each tool. We re-train the classification model in CPAT for different datasets to optimize its performance. Because CPC and PhyloCSF do not provide the re-training option, we use pre-built models. As stated in the users' manual, model training in PLEK is very time-consuming. We were not able to train a new model using the scripts in the PLEK software package on a machine with two 2.4 GHz quad-core Intel Xeon processors, 8 GB memory and Linux operating system. Therefore, we used the pre-built model in the PLEK software package with multi-thread configuration as suggested in its README file.

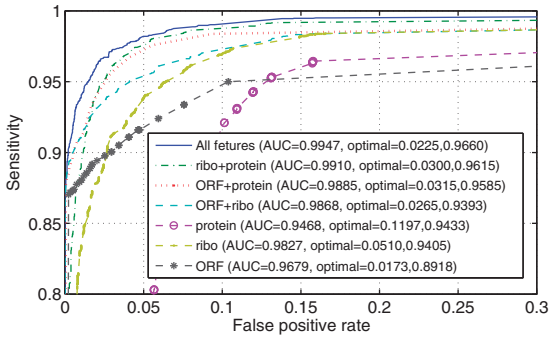

Fig. 1. Performance comparison among feature groups: ORF features (ORF), ribosome interaction features (ribo), protein conservation features (protein) and the combined feature sets

Table 1. Performance comparison on H1

Metric	LncRNA-ID	CPC	CPAT	PhyloCSF	PLEK
Sensitivity	96.28	66.48	86.95	77.08	**99.52**
Specificity	95.28	**99.97**	99.55	85.08	89.18
F-score	**95.80**	79.85	92.80	79.45	94.57
Accuracy	**95.78**	83.22	93.25	81.34	94.32

CPC, CPAT, PhyloCSF and PLEK were evaluated using their default score cutoffs. Bold numbers indicate the highest value of the metrics.

3.2 Performance evaluation on the human data (H1)

We ran CPC using UniRef90 (Suzek *et al.*, 2007) as the reference protein database, which is a relatively comprehensive protein database suggested by CPC. We created the classification model of CPAT from the training data using the script provided in CPAT's software package. The created classifier was then used to predict the transcripts in the testing data and the performance was evaluated using the CPAT's suggested optimal score cutoff.

A multiple sequence alignment of 45 vertebrate genomes, including the human genome, was downloaded from the UCSC Genome Browser and was used as the input alignment to PhyloCSF. We specified the option that allowed PhyloCSF to search all three reading frames and report the best result as suggested in the PhyloCSF website (Lin *et al.*, 2011). We used the default score cutoff of PhyloCSF to generate the classification results. We ran PLEK using its default settings.

Table 1 shows the comparison of classification performance of all tools on H1. LncRNA-ID had the best F-score and accuracy. PLEK had the highest sensitivity. Although CPC had the highest specificity, its sensitivity and accuracy were much lower than those of LncRNA-ID. CPC's classifier is based on six features. Three of them are ORF-related features and the others are derived from the alignments of a query sequence against existing protein sequences. These features could cause a bias toward protein-coding transcripts if an lncRNA contains an ORF sharing similarity with existing protein sequences. This might be a major reason behind the low sensitivity of CPC. Overall, LncRNA-ID shows the best tradeoff between sensitivity and accuracy.

We further evaluated how the classification performance of different tools changed under difference score cutoffs. Figure 2 shows

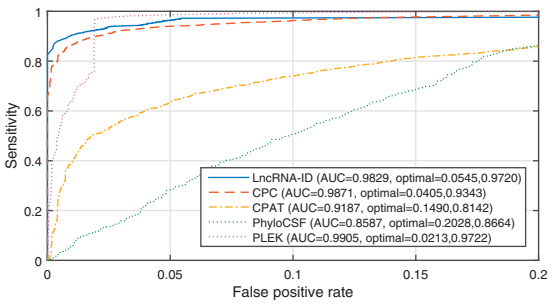

Fig. 2. ROC curves of different tools on H1. The AUCs and the sensitivity and FPR corresponding to the optimal operating point were given in the legend

false positive rate versus sensitivity (ROC curves) under different score cutoffs. The sensitivity and FPR corresponding to the optimal operating point for each tool are also shown in Figure 2. LncRNA-ID, CPC and PLEK have comparable AUC. However, PLEK has much lower sensitivity than LncRNA-ID and CPC when the FP rate is smaller than 0.019.

The optimal performance of CPAT and CPC was much better than that based on their default score cutoffs, showing that their default score cutoffs are data dependent and may not provide users with satisfactory classification performance. The optimal performance and AUC of PhyloCSF was much worse than LncRNA-ID, CPC and PLEK.

3.3 Performance evaluation on the mouse dataset (M1)

LncRNA-ID can be trained for any species with some characterized protein-coding and lncRNA genes. If no such training data is available, pre-built model can be used. In this experiment, we applied LncRNA-ID to the mouse dataset to show its application to a different species. The training dataset consists of 22 033 protein-coding transcripts and 2457 lncRNAs, which are randomly selected from GENCODE. The testing data contains 2000 coding transcript and 2000 lncRNAs. In both training and testing datasets, only one transcript from each gene is used. Moreover, the training and testing datasets do not have transcripts that are from the same genes. The number of lncRNAs in this dataset is around half of that contained in H1 because of limited lncRNA annotation in the mouse genome.

A multiple sequence alignment of 30 genomes, including the mouse genome, was downloaded from the UCSC Genome Browser and used as the input alignment to PhyloCSF. The score cutoff of 50, which was shown to accurately separate known protein-coding genes from known non-coding sequences (Guttman *et al.*, 2010, 2013), was used to generate the classification results for PhyloCSF.

Table 2 shows the performance comparison of different tools on the mouse dataset under their default parameters. LncRNA-ID had the best sensitivity and accuracy among all tools. Although CPAT and CPC had higher specificity than LncRNA-ID, their sensitivity and accuracy were lower than those of LncRNA-ID. LncRNA-ID also had the highest F-score among all tools, showing its best overall classification performance.

Except CPC, the sensitivity, F-score and accuracy of all tools on the mouse dataset were lower than those on H1, largely due to the smaller training dataset. Note that as we used CPC's pre-built classifier to evaluate the testing data, there might be some overlapping samples between CPC's training data and this testing data, giving an advantage to CPC's classifier over other tools.

Figure 3 shows the ROC curves of different tools. When the false positive rate was higher than 0.041, CPC had higher sensitivity than the other tools. However, its optimal sensitivity was much lower than that of LncRNA-ID. LncRNA-ID and CPC had much better AUC than the other tools. At the point with the optimal performance, CPCs sensitivity was 2.85% lower than that of LncRNA-ID and its false positive rate was 9.35% higher than that of LncRNA-ID. The performance of CPAT under its default score cutoff was much worse than its optimal performance when its score cutoff was changed. PhyloCSF had significantly poorer performance than the other tools. Although PLEK has been tested on mouse and several other species (Li *et al.*, 2014), its performance on our testing data is worse than other tools except PhyloCSF.

Table 2. Performance comparison on the mouse dataset (M1)

Metric	LncRNA-ID	CPC	CPAT	PhyloCSF	PLEK
Sensitivity	**94.45**	76.55	38.80	24.50	88.11
Specificity	92.10	**98.75**	98.95	55.70	70.94
F-score	**93.36**	86.11	55.49	31.76	81.07
Accuracy	**93.28**	87.65	68.88	41.43	79.49

CPC, CPAT and PhyloCSF were evaluated using default score cutoffs. Bold numbers indicate the highest values of the metrics.

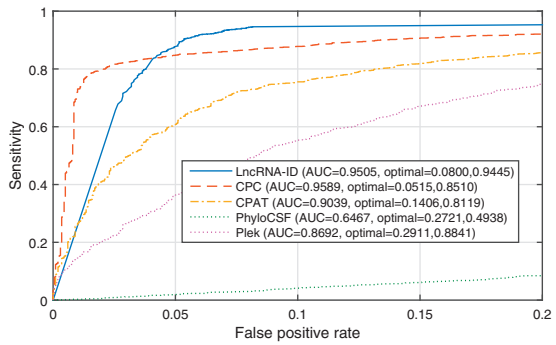

Fig. 3. ROC curves of different tools on the mouse dataset. The AUCs, and the sensitivity and FPR corresponding to the optimal operating point were indicated in the legend

3.4 Performance evaluation on CPAT's human dataset (H2)

In this experiment, we evaluated the performance of LncRNA-ID on the human dataset used by CPAT (Wang *et al.*, 2013). The training set was originally claimed to contain 10 000 coding transcripts selected from the RefSeq database and 10 000 randomly selected non-coding transcripts from GENCODE. However, some transcripts no longer exist in the databases. They might have been removed because of the duplication with the existing ones, non-qualification as new evidence has emerged, etc. As a result, the final training set contains 9929 coding transcripts and 9066 non-coding transcripts. The test set is the same as CPAT's original data. It contains 4000 coding transcripts from RefSeq database and 4000 lncRNAs from a human lncRNA catalog. The performance of CPC and PhyloCSF had been benchmarked on this dataset in (Wang *et al.*, 2013) and were used in our experiment. We created the classification model of CPAT from the training transcripts using the script provided in CPAT's software package. The created classifier was then used to predict the transcripts in the testing dataset and the performance was evaluated using the CPAT's suggested optimal score cutoff.

The first three columns of Table 3 show the performance comparison between LncRNA-ID and CPAT. LncRNA-ID

Table 3. Performance comparison on H2

	LncRNA-ID	CPAT	LncRNA-ID				RF[a]				CPAT			
	Original data		S2	S3	S6	S8	S2	S3	S6	S8	S2	S3	S6	S8
Sensitivity	**93.80**	87.58	93.43	93.54	92.72	92.73	90.03	87.97	82.39	80.48	79.51	73.01	60.32	54.46
Specificity	96.03	**97.32**	95.41	95.23	95.47	95.38	96.15	95.94	96.53	96.60	98.14	98.50	99.29	99.42
F-score	**94.82**	92.20	94.36	94.34	94.01	93.91	92.87	91.62	88.65	87.53	87.84	83.66	74.91	70.21
Accuracy	**94.89**	92.45	94.42	94.38	94.10	94.06	93.09	91.95	89.46	88.54	88.82	85.75	79.81	76.94

Bold numbers indicate the highest values of the metrics.
[a]Random forest classifiers created with our feature sets.

(sensitivity: 93.80%, specificity: 96.03%, F-score: 94.82%) achieved a better overall performance compared with CPAT (sensitivity: 87.58%, specificity: 97.32%, F-score: 92.20%), CPC (sensitivity: 73.75%, specificity: 99.9%, F-score: 84.84%) and PhyloCSF (sensitivity: 62.75%, specificity: 90.24%, F-score: 72.75%). It has comparable performance to PLEK (sensitivity: 95.68%, specificity: 93.58%, F-score: 94.68%). Please note that we used the performance of CPC and PhyloCSF that were benchmarked on the same testing data in (Wang *et al.*, 2013).

3.5 Imbalanced training data

Using the H2 dataset, we evaluated how imbalanced training data affected the classification performance of LncRNA-ID and other classification models including CPAT. We choose CPAT because: (i) it has good performance on both human and mouse datasets; (ii) it is the only tool whose model can be easily and efficiently re-trained using different training data. In addition, to evaluate whether using balanced random forest is more robust to imbalanced training data than other classifiers, we compared the performance of LncRNA-ID with a logistic regression model that used our feature set (LR), and a regular random forest model that used our feature set (RF). The performance of LR was shown in Supplementary Table S9.

We constructed four datasets, S2, S3, S6 and S8, from the original training set simulating the condition of imbalanced training data. In S2, lncRNAs in the training set were randomly divided into two subsets. Each lncRNA subset was combined with coding transcripts in the original training set, generating two training subsets in total. We created the classification models of LncRNA-ID and CPAT using each of the two training subsets and evaluated their performance using the same test set. The experiments on S3, S6 and S8 were conducted in the same manner except that lncRNAs were divided into three, six and eight subsets, respectively. We evaluated the performance of different tools on all the training subsets of each dataset and reported the average of each evaluation metric.

Table 3 shows the performance comparison of LncRNA-ID, CPAT and RF on the imbalanced datasets. LncRNA-ID had higher average sensitivity than CPAT on all the imbalanced datasets. The performance of CPAT's classifiers trained with the subsets significantly decreased compared with that trained with the full training set. This shows that limited learning data led to less discriminative power of CPAT's classifier. In contrast, LncRNA-ID, which uses BRF, was able to maintain stable performance in all datasets with different ratios of imbalance.

The average specificity of LncRNA-ID on all the imbalanced datasets was 3.46% lower than that of CPAT. However, the average sensitivity of LncRNA-ID was 26.28% higher than that of CPAT. The average accuracy and F-score of LncRNA-ID were also much higher than those of CPAT, showing better overall classification performance. The sensitivity of CPAT dramatically dropped by 9.21–37.81% while the sensitivity of LncRNA-ID decreased by less than 1% compared with those trained with the full training set. Moreover, with the increase in the imbalance ratio, the sensitivity of CPAT dramatically decreased from 79.51 to 54.46% while its specificity only increased from 98.14 to 99.42%. This shows that CPAT suffered not only from the limited learning data but also the impact of the imbalance in the training data.

RF showed much higher sensitivity with slightly lower specificity compared with CPAT. Similar to CPAT, the performance of RF also suffered from the increase of the imbalanced ratio in the training data. From S2 to S8, its sensitivity decreased about 10% while the sensitivity of LncRNA-ID only decreased by about 1%. This shows the better performance of balanced random forest compared with a regular random forest.

3.6 Performance evaluation on experimentally verified lncRNAs (M2)

One challenge for evaluating and comparing the performance of lncRNA identification tools is to obtain true and reliable testing data. In all previous experiments, we have tested the performance of various software on lncRNA datasets widely adopted by previously published tools. However, not all the lncRNAs in these databases have been experimentally verified. To comprehensively assess the performance of these tools, we evaluated the performance of LncRNA-ID on experimentally verified lncRNAs in the mouse embryonic stem cells (Guttman *et al.*, 2013).

These lncRNAs transcripts were experimentally identified using ribosome profiling and mRNA-seq data. Only those lincRNAs with significant expression level relative to the randomized genomic average were selected. As a result, this dataset contains 800 lncRNA transcripts.

As this dataset only has experimental verified lncRNAs, which are positive cases, the PPV value of a classifier is always equal to one. Therefore, we focus on evaluating the sensitivity of different tools to identify these lncRNAs because the specificity of these tools on mouse data has been tested in the M1 experiment. The classification model trained on M1 was used to evaluate the performance of LncRNA-ID on this dataset. We used CPAT's classification model created for mouse and the suggested score cutoff to evaluate its performance. PhyloCSF, CPC and PLEK were applied using the same settings as in M1.

LncRNA-ID achieved the highest sensitivity (96.25%) compared with CPC (95.80%), CPAT (92.99%), PhyloCSF (91.51%) and PLEK (69.50%). These results show that LncRNA-ID has consistently high sensitivity on commonly used lncRNA databases and on the smaller experimentally verified lncRNA dataset.

3.7 Running time

We measured the running time of all tested tools on the testing data of H1, which is the largest. All tools were ran on the same node with two 2.4 GHz quad-core Intel Xeon processors, 24 GB memory and Linux

operating system. It took CPC, CPAT, PhyloCSF and PLEK 86.51 h, 35.36 s, 15 097.60 h and 21.47 m to process the data. Note that the running time of PhyloCSF did not include the time used for preparing the input multiple sequence alignments, which can be computationally expensive. LncRNA-ID took 65.36 s to process the data. Its speed was comparable to CPAT, and much faster than CPC and PhyloCSF.

4 Discussion

In this work, we have proposed and implemented LncRNA-ID, an accurate lncRNA identification tool using balanced random forest classification. LncRNA-ID ensembles multiple random forest classifiers induced from balanced down-sampled data and thus is able to maintain robust performance with imbalanced and limited learning data. The experimental results in both human and mouse genomes demonstrate that the features used by LncRNA-ID have strong discriminative power in distinguishing lncRNAs from protein-coding transcripts. In addition, we conducted experiments using LR, which combines our feature set with the logistic regression model in CPAT, and compared its performance with LncRNA-ID. The experimental results showed the utility of balanced random forest model for highly imbalanced training data.

According to our empirical experimental results, the ribosome interaction features are the most discriminative features. The reason might be that ribosome is closely related to the protein translation mechanism, which is the major difference between coding and noncoding transcripts. In our experiments, we focused on the ribosome interaction in eukaryotes where annotated lncRNAs are more publicly available. In eukaryotes, a small 40S ribosomal subunit contains 18S rRNA whereas in prokaryotes a small 30S ribosomal subunit contains 16S rRNA. Further study is needed to investigate whether the similar ribosome signal could be captured with 16S rRNA as with 18S rRNA.

Among all classification tools, PhyloCSF had the worst performance. A closer look at this result shows that in the human dataset, PhyloCSF could not determine the coding status of a decent amount of lncRNAs (16.97%) and some coding transcripts (0.03%). This is because they are either non-conserved transcripts or do not have sufficient long ORFs. If both ribosome profiling data (Ribo-Seq) and mRNA-Seq data are available, a more accurate ribosome release signal could be measured using the number of mapped reads on ORF and 3′UTR. The RRS is then defined as the ratio of the two normalized ratios of mapped reads on these two regions, $\mathrm{RRS} = (r_{\mathrm{ORF}}/r_{3'\mathrm{UTR}})_{\mathrm{Ribo-Seq}}/(r_{\mathrm{ORF}}/r_{3'\mathrm{UTR}})_{\mathrm{mRNA-Seq}}$ (Ingolia *et al.*, 2011), where r is the number of mapped reads on a sequence.

PLEK has comparable performance to LncRNA-ID on human but did not perform well on mouse data. PLEK was run on the prebuilt model, which was trained on a very large number of coding transcripts and lncRNAs. Thus it is possible that their training data contains some of our testing data and leads to very good performance on human. If PLEK can be re-trained using our training data, we can conduct a more thorough comparison.

Funding

This work is supported by the National Science Foundation [DBI-0953738].

Conflict of Interest: none declared.

References

Arava,Y. *et al.* (2003) Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA*, **100**, 3889–3894.

Arriaga-Canon,C. *et al.* (2014) A long non-coding RNA promotes full activation of adult gene expression in the chicken globin domain. *Epigenetics*, **9**, 173–181.

Boerner,S. and McGinnis,K.M. (2012) Computational identification and functional predictions of long noncoding RNA in Zea mays. *PLoS ONE,* **7**, e43047.

Borsani,G. *et al.* (1991) Characterization of a murine gene expressed from the inactive X chromosome. *Nature*, **351**, 325–329.

Breiman,L. (2001) Random forests. *Mach. Learn.,* **45**, 5–32.

Brockdorff,N. *et al.* (1992) The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell,* **71**, 515–526.

Bu,D. *et al.* (2012) NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res.*, **40**, D210–D215.

Cabili,M.N. *et al.* (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.

Chen,C. *et al.* (2004) Using random forest to learn imbalanced data. *Technical report*, Department of Statistics, University of Berkeley.

Chen,G. *et al.* (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.*, **41**, D983–D986.

Chen,X. and Ishwaran,H. (2012) Random forests for genomic data analysis. *Genomics,* **99**, 323–329.

Chodroff,R.A. *et al.* (2010) Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome Biol.*, **11**, R72.

De Angioletti,M. *et al.* (2004) Beta + 45 G–C: a novel silent beta-thalassaemia mutation, the first in the Kozak sequence. *Br. J. Haematol.*, **124**, 224–231.

Derrien,T. *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.

Dinger,M.E. *et al.* (2008) Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput. Biol.*, **4**, e1000176.

Djebali,S. *et al.* (2012) Landscape of transcription in human cells. *Nature,* **489**, 101–108.

Durbin,R. *et al.* (1998) *Biological Sequence Analysis Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, UK.

Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inf.,* **23**, 205–211.

Finn,R.D. *et al.* (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, 29–37.

Gonen,M. (2007) *Analyzing Receiver Operating Characteristic Curves With SAS*. SAS Press Series, USA.

Guttman,M. *et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, **28**, 503–510.

Guttman,M. *et al.* (2013) Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell,* **154**(1), 240–251.

Hall,M. *et al.* (2009) The WEKA data mining software: An update. *SIGKDD Explorations,* **11**.

Huang,K.-Z. *et al.* (2008) *Machine Learning: Modeling Data Locally and Globally*. Springer Science and Business Media, Germany.

Humann,F.C. *et al.* (2013) Sequence and expression characteristics of long noncoding RNAs in honey bee caste development—potential novel regulators for transgressive ovary size. *PLoS ONE,* **8**, e78915.

Hung,T. and Chang,H.Y. (2010) Long noncoding RNA in genome regulation: prospects and mechanisms. *RNA Biol.,* **7**, 582–585.

Ingolia,N.T. *et al.* (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell,* **147**, 789–802.

Kapranov,P. *et al.* (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science,* **316**, 1484–1488.

Kong,L. *et al.* (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.*, **35**, W345–W349.

Kozak,M. (1989) Context effects and inefficient initiation at non-aug codons in eucaryotic cell-free translation systems. *Genome Res.,* **9**, 5073–5080.

Kozak,M. (1997) Recognition of aug and alternative initiator codons is augmented by g in position +4 but is not generally affected by the nucleotides in positions +5 and +6. *EMBO J.,* **16**, 2482–2492.

Kozak,M. (1999) Initiation of translation in prokaryotes and eukaryotes. *Gene,* **234**, 187–208.

Leung,Y.Y. *et al.* (2013) CoRAL: predicting non-coding RNAs from small RNA-sequencing data. *Nucleic Acids Res.*, **41**, e137.

Li,A. *et al.* (2014) PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics,* **15**, 311.

Lin,M.F. *et al.* (2011) Phylocsf: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics,* **27**, i275–i282.

Liu,J. *et al.* (2012) Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *Plant Cell,* **24**, 4333–4345.

Liu,Y. *et al.* (2013) Inheritable and precise large genomic deletions of non-coding RNA genes in zebrafish using TALENs. *PLoS One,* **8**, e76387.

Magrane,M. and Consortium,U. (2011) UniProt knowledgebase: a hub of integrated protein data. *Database,* **2011**, bar009.

Marchler-Bauer,A. and Bryant,S.H. (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.*, **32**, W327–W331.

Marchler-Bauer,A. *et al.* (2013) CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res.*, **41**, D348–D352.

MATLAB (2010) *version 7.10.0 (R2010a).* The MathWorks Inc., Natick, MA.

Muckstein,U. *et al.* (2006) Thermodynamics of RNA-RNA binding. *Bioinformatics,* **22**, 1177–1182.

Okazaki,Y. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60 770 full-length cDNAs. *Nature,* **420**, 563–573.

Pauli,A. *et al.* (2011) Non-coding RNAs as regulators of embryogenesis. *Nat. Rev. Genet.*, **12**, 136–149.

Pauli,A. *et al.* (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res.,* **22**, 577–591.

Pennisi,E. (2012) Genomics. ENCODE project writes eulogy for junk DNA. *Science,* **337**, 1159, 1161.

Prasanth,K.V. and Spector,D.L. (2007) Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. *Genes Dev.*, **21**, 11–42.

Probost,F. (2000) Machine learning from imbalanced data sets 101, *Proc. Learning from Imbalanced Data Sets: Papers from the Am. Assoc. for Artificial Intelligence Workshop*, 2000.

Pruitt,K.D. *et al.* (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.,* **35**, 61–65.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, 290–301.

Schloss,P.D. (2010) The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput. Biol.,* **6**.

Shaw,K. (2008) Biological applications of support vector machines. *Nat. Educ.,* **1**, 201.

Suzek,B.E. *et al.* (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics,* **23**, 1282–1288.

Tuv,E. *et al.* (2009) Feature selection with ensembles, artificial variables, and redundancy elimination. *J. Mach. Learn. Res.*, **10**, 1341–1366.

Vasquez,J. *et al.* (2014) Comparative ribosome profiling reveals extensive translational complexity in different *Trypanosoma brucei* life cycle stages. *Nucl. Acids Res.*, **42**, 3623–3637.

Wang,L. *et al.* (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.*, **41**, e74.

Wapinski,O. and Chang,H.Y. (2011) Long noncoding RNAs and human disease. *Trends Cell Biol.*, **21**, 354–361.

Wilusz,J.E. *et al.* (2009) Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.*, **23**, 1494–1504.

Wu,J. *et al.* (2009) Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics,* **25**, 30–35.

Xing,C. *et al.* (2009) Identification of protein-coding sequences using the hybridization of 18S rRNA and mRNA during translation. *Nucleic Acids Res.*, **37**, 591–601.

Xu,H. *et al.* (2010) Length of the ORF, position of the first AUG and the Kozak motif are important factors in potential dual-coding transcripts. *Cell Res.*, **20**, 445–457.

Zhang,Y. and Sun,Y. (2011) HMM-FRAME: accurate protein domain classification for metagenomic sequences containing frameshift errors. *BMC Bioinformatics,* **12**, 198.

Zhang,Y. and Sun,Y. (2012) MetaDomain: a profile HMM-based protein domain classification tool for short sequences. In: *Proceedings of Pacific Symposium on Biocomputing (PSB).*

Zhang,Y. *et al.* (2013) A Sensitive and Accurate protein domain cLassification Tool (SALT) for short reads. *Bioinformatics,* **29**, 2103–2111.

Zhang,Y. *et al.* (2014) A scalable and accurate targeted gene assembly tool (SAT-Assembler) for next-generation sequencing data. *PLoS Comput. Biol.*, **10**, e1003737.