Li et al. BMC Bioinformatics 2014, 15:311
http://www.biomedcentral.com/1471-2105/15/311

BMC Bioinformatics

SOFTWARE Open Access

PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved *k*-mer scheme

Aimin Li[1,2], Junying Zhang[1*] and Zhongyin Zhou[3,4]

Abstract

Background: High-throughput transcriptome sequencing (RNA-seq) technology promises to discover novel protein-coding and non-coding transcripts, particularly the identification of long non-coding RNAs (lncRNAs) from *de novo* sequencing data. This requires tools that are not restricted by prior gene annotations, genomic sequences and high-quality sequencing.

Results: We present an alignment-free tool called PLEK (***p***redictor of ***l***ong non-coding RNAs and m***e***ssenger RNAs based on an improved ***k***-mer scheme), which uses a computational pipeline based on an improved *k*-mer scheme and a support vector machine (SVM) algorithm to distinguish lncRNAs from messenger RNAs (mRNAs), in the absence of genomic sequences or annotations. The performance of PLEK was evaluated on well-annotated mRNA and lncRNA transcripts. 10-fold cross-validation tests on human RefSeq mRNAs and GENCODE lncRNAs indicated that our tool could achieve accuracy of up to 95.6%. We demonstrated the utility of PLEK on transcripts from other vertebrates using the model built from human datasets. PLEK attained >90% accuracy on most of these datasets. PLEK also performed well using a simulated dataset and two real *de novo* assembled transcriptome datasets (sequenced by PacBio and 454 platforms) with relatively high indel sequencing errors. In addition, PLEK is approximately eightfold faster than a newly developed alignment-free tool, named Coding-Non-Coding Index (CNCI), and 244 times faster than the most popular alignment-based tool, Coding Potential Calculator (CPC), in a single-threading running manner.

Conclusions: PLEK is an efficient alignment-free computational tool to distinguish lncRNAs from mRNAs in RNA-seq transcriptomes of species lacking reference genomes. PLEK is especially suitable for PacBio or 454 sequencing data and large-scale transcriptome data. Its open-source software can be freely downloaded from https://sourceforge.net/projects/plek/files/.

Keywords: RNA-seq, lncRNA, *k*-mer, Prediction, *de novo* sequencing, *de novo* assemble

Background

With the development of high-throughput transcriptome sequencing techniques (RNA-seq) [1], numerous transcripts have been identified in many species [2-4]. A new class of transcripts, long non-coding RNAs (lncRNAs, typically >200 nt), are of particular interest because they contribute to many important biological processes, such as dosage compensation [5], regulation of gene expression [6] and cell cycle regulation [7,8]. Moreover, a number of studies showed that mutations and dysregulations in lncRNA genes were associated with human diseases [9,10], such as cancers [11-13]. It remains a challenge to distinguish mRNAs from lncRNAs, especially for *de novo* sequencing without highly confident reference sequences and comprehensive annotations, because lncRNAs show many features similar to mRNAs, such as poly(A) tails, splicing and approximate sequence length [14]. Until now, several tools, such as CPC [15] and PhyloCSF [16], have been developed based on known protein databases, intrinsic sequence features and sequence conservation properties. These tools show varied efficiencies in distinguishing lncRNAs from mRNAs for different datasets. Most

* Correspondence: jyzhang@mail.xidian.edu.cn
[1]School of Computer Science and Technology, Xidian University, Xi'an, PR China
Full list of author information is available at the end of the article

BioMed Central

© 2014 Li et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

of them rely heavily on sequence alignment, and are therefore time-consuming and restricted by prior annotations. Recently, a software tool named Coding-Non-Coding Index (CNCI) was developed [17]. It discriminates coding from non-coding transcripts using intrinsic sequence features. CNCI performs better than previous tools for poorly annotated species or those without whole-genome sequence information, but it may misclassify transcripts when there are insertion or deletion (indel) sequencing errors among them. Such errors are common in the current Roche (454) and Pacific Biosciences (PacBio) sequencing platforms [18-21].

In this study, we developed a characteristic *k*-mer based alignment-free tool named PLEK, to solve the above-mentioned problems. PLEK takes calibrated *k*-mer frequencies of a transcript sequence as its computational features. With these features, the support vector machine (SVM) algorithm was used to build a binary classification model to separate lncRNAs from mRNAs. The classification model achieved high accuracy (95.6%) on training data with 10-fold cross-validation. PLEK also performed well on data from other vertebrates, using the classification model built from human training data. For transcripts containing indel sequencing errors, PLEK also attained high accuracy (>94%) in simulated and real transcriptome datasets. Moreover, PLEK is 8 times faster than CNCI and 244 times faster than CPC on the same test data. Therefore, PLEK is an accurate, robust and fast tool. It is suitable for vertebrates lacking high-quality genome sequences and annotation information, and is especially effective for the *de novo* assembled transcriptome data generated by PacBio or 454 sequencing platforms.

Implementation

Careful selection of high-quality training data and their appropriate computational features is crucial to build an accurate, robust and fast binary classifier. In this section, we describe the data used to build the classification model and to test its performance. We then describe the distinct computational pipeline of *k*-mer usage. This is followed by the construction of a binary classifier using these data, *k*-mer usage features and SVM algorithm. Finally, we introduce simulation of indel sequencing errors on human protein-coding transcripts and transcriptome sequencing data from PacBio and 454 platforms.

Data description

The RefSeq [22,23] and GENCODE [24-26] projects provide comprehensive, non-redundant and well-annotated set of sequences, which can be used to build high-quality training and test datasets. Human protein-coding transcripts were downloaded from the RefSeq database (release 60) and human long non-coding transcripts were collected from GENCODE v17. There were 34,691 protein-coding transcripts with the length of >200 nt in the human RefSeq dataset, and 22,389 long (>200 nt) non-coding transcripts in the human GENCODE dataset. For performance assessment of cross-species prediction, we gathered transcripts from other vertebrates from the Ensembl [27] database (v72) (Table 1). To compare PLEK against CNCI, CPC and PhyloCSF, 6,015 mouse lncRNAs were gathered from GENCODE database (vM2). Mouse mRNAs were collected from RefSeq (release 60) and those with 'putative','predicted' or 'pseudogene' annotations were excluded. All the sequences used were longer than 200 nt.

Improved *k*-mer scheme

To characterize lncRNA and mRNA transcript sequences, we used *k*-mer usage and sliding-windows with a one-nucleotide step-length to analyze each transcript. A *k*-mer pattern is a specific string with *k* nucleotides, each can be *A*, *C*, *G* or *T*. For $k = 1$ to 5, we had $4 + 16 + 64 + 256 + 1024 = 1{,}364$ patterns: 4 *one*-mer patterns, 16 *two*-mer patterns, 64 *three*-mer patterns, 256 *four*-mer patterns, and 1,024 *five*-mer patterns.

In designing a sliding-window of length *k*, $k = 1,2,\ldots,5$, which slides along the transcript of length *l* by a step-length of one nucleotide, if the string in the transcript within the window matched with some pattern *i*, the occurrence number of the pattern in the transcript, denoted by c_i, was increased by one. We did not simply use usage frequency c_i/s_k, $i = 1$ to 1364 (where s_k was the total number of times that the sliding-window of size *k*-nt could slide along the transcript, $s_k = l\text{-}k + 1$); however, we calibrated it as f_i by a factor relating to the length of the pattern, w_k, as the features of the transcript for prediction. The features used for prediction were given in formula (1).

$$f_i = \frac{c_i}{s_k} w_k, \quad k = 1,2,3,4,5.\ i = 1,2,\ldots,1364 \tag{1}$$

$$s_k = l\text{–}k + 1, \quad k = 1,2,3,4,5 \tag{2}$$

$$w_k = \frac{1}{4^{5-k}}, \quad k = 1,2,3,4,5 \tag{3}$$

Construction of classification model

To produce a balanced training dataset, we collected all the 22,389 long non-coding transcripts from the GENCODE v17 dataset (labelled as the "negative" class) and randomly selected 22,389 protein-coding transcripts from the human RefSeq dataset (labelled as the "positive" class). The 1,364 calibrated *k*-mer usage frequencies of each transcript were regarded as computation features. First, these calibrated frequencies were normalized to the range from 0 to 1 using the *svm-scale* program from the LIBSVM package (version 3.17) [28]. Second, a

Table 1 Data sources and performance of cross-species prediction

Species	Data source	Number of transcripts	Accuracy of CNCI	Accuracy of PLEK
Mus musculus	RefSeq mRNA	26062	**93.9%**	88.1%
	Ensembl ncRNA	2963	**97.1%**	89.9%
Danio rerio	RefSeq mRNA	14493	**95.3%**	91.3%
	Ensembl ncRNA	419	89.3%	**90.9%**
Xenopus tropicalis	RefSeq mRNA	8874	92.9%	**94.5%**
	Ensembl ncRNA	279*	99.7%	**100.0%**
Bos taurus	RefSeq mRNA	13190	94.3%	**94.8%**
	Ensembl ncRNA	182	**100.0%**	99.5%
Pan troglodytes	RefSeq mRNA	1906	**90.2%**	87.1%
	Ensembl ncRNA	1166	**100.0%**	99.9%
Sus scrofa	RefSeq mRNA	3978	**93.4%**	85.1%
	Ensembl ncRNA	241	95.9%	**98.3%**
Macaca mulatta	RefSeq mRNA	5709	**92.0%**	85.0%
	Ensembl ncRNA	359	99.7%	**100.0%**
Gorilla gorilla	RefSeq mRNA	33025	**87.4%**	83.8%
	Ensembl ncRNA	367	**99.7%**	**99.7%**
Pongo abelii	RefSeq mRNA	3401	93.4%	**98.0%**
	Ensembl ncRNA	392	99.8%	**100.0%**

PLEK and CNCI were tested on the same data; better accuracies are shown in bold face type. For RefSeq mRNAs, those with 'putative', 'predicted' or 'pseudogene' annotations were excluded (except for *Gorilla gorilla*).
*279 non-coding transcripts with lengths of more than 150 nt.

support vector machine (SVM) with a radial basis functional kernel, whose variance is *gamma*, was selected as the binary classifier. The optimal *C* of the SVM and *gamma* of the kernel were obtained using the *grid.py* script of the LIBSVM package. During the process of parameter searching, 10-fold cross-validation was carried out to assess the performance of the classification model for each *C* and *gamma* parameter. Finally, we built an SVM binary classifier with the optimal *C* and *gamma*.

Simulation of indel sequencing errors

Assembly of transcripts is made difficult by short read length sequences, which are typically generated by next-generation sequencing technologies [29,30], especially by Illumina sequencing platforms. In contrast, PacBio and 454 platforms generate longer reads, which tend to be more easily assembled than short reads [31]. A large number of studies have been performed on these two kinds of sequencers [32-38]. However, the indel sequencing error rates are relatively high in PacBio and 454 sequencing data [19,39]. A tool robust to such errors is desirable to distinguish lncRNAs and mRNAs, and facilitates annotation of lncRNAs and mRNAs of a species without whole-genome sequences.

We simulated single-base homopolymer-associated indel sequencing errors in protein-coding transcripts to evaluate the robustness of our tool, because they are the most typical indel sequencing errors in PacBio and 454 sequencing platforms. Without loss of generality, we simulated 0 to 3 single-base indel sequencing errors per 100 bases (the error rate p was 0% to 3%). For a transcript with a length of l bases, it had $n = lp$ indel errors. We first counted the number of homopolymers of various lengths. Suppose the corresponding number of different homopolymers with lengths of l_1 , l_2, ..., l_t was m_1, m_2, ..., m_t, respectively, where $l_1 > l_2 > \ldots > l_t$, and t is the number of different lengths of homopolymers. A biological fact is that the likelihood of an indel error increases with the length of a homopolymer [19], even with no possibility that the indel error is in place between homopolymers. Thus, the indel errors start with the longest homopolymers with the length of l_1. If $m_1 < n$, the homopolymers with the length of l_2 are also inserted or deleted with bases, and so on, until n relatively longer homopolymers are processed. For these n homopolymers, we randomly inserted or deleted an identical base. If there are many homopolymers and few indels, the positions of indels will be evenly distributed in the transcripts (see Additional file 1 for an example).

Construction of a real sequencing dataset

We used following two transcriptome datasets, sequenced by PacBio and 454 platforms, to test the performance of PLEK on *de novo* assembled transcripts without reference genomes. The first dataset was recently released by PacBio

(Pacific Biosciences, Menlo Park, CA, USA). After full-length cDNA sequencing of the human MCF-7 transcriptome by the PacBio Isoform Sequencing (Iso-Seq) technology, PacBio used an iterative clustering algorithm and Quiver [40] to assemble *de novo* 44,531 polished, full-length and non-redundant transcripts. We aligned these transcripts with human RefSeq mRNAs (release 60) and GENCODE lncRNAs (v17) using NCBI Blastn with the parameters: *-task megablast -evalue 1e-10 -perc_identity 80 -max_target_seqs 1*. Then, filtering these transcripts by query coverage >80%, subject coverage >80% and gaps >0, we obtained 3,306 transcripts (3,185 mRNAs and 121 lncRNAs) that may have had indel sequencing errors.

The second dataset, a HelaS3 cell line transcriptome, was sequenced by a 454 GS FLX Titanium platform and is available at the Sequence Read Archive (SRA) under accession no. SRA063146 [41]. LUCY (1.20p) [42] and SeqClean (ftp://occams.dfci.harvard.edu/pub/bio/tgi/software/) were used to discard the adapter sequences, low quality reads and sequences of less than 50 bp, resulting in 4,222,133 high-quality sequences. Trinity r20131110 [43] with default parameters was used to assemble these *de novo*. Trinity assembled 65,583 transcripts, which were subsequently aligned with human RefSeq mRNAs and GENCODE lncRNAs, and were filtered in the same manner as the first dataset. Finally, 3,098 transcripts with possible indel sequencing errors were retained (3,045 mRNAs, 53 lncRNAs).

Running PhyloCSF on mouse datasets

In order to compare PLEK against PhyloCSF [16] on mouse datasets, we set up a local instance of Galaxy [44]. Multiple alignments of 59 assemblies to the mouse genome (mm10/GRCm38) were downloaded from the UCSC Genome Browser (http://hgdownload-test.sdsc.edu/goldenPath/mm10/multiz60way/maf/). BED files describing mouse lncRNA and mRNA transcripts were loaded onto the local Galaxy webserver and the tool 'Stitch Gene Blocks' was used to retrieve multiple alignment files with sequence entries for the following genome builds based on the 60-way Multiz alignment to *mm10*: *mm10*, *rn5*, *dipOrd1*, *cavPor3*, *speTri2*, *oryCun2* and *ochPri2*. Genome build names were converted to common names and PhyloCSF was run using the options: *–frames = 3, –orf = StopStop3* and *–removeRefGaps*.

Results

Different usage frequencies of *k*-mer strings

To verify the difference between mRNA and lncRNA in *k*-mer string usage, we calculated the calibrated usage frequencies of all the 1,364 *k*-mer patterns in the positive training dataset (22,389 protein-coding transcripts) and negative training dataset (22,389 long non-coding transcripts). We used the *Wilcox* rank-sum test to determine which *k*-mer pattern usage was significantly different between mRNAs and lncRNAs. With a significance level of 10^{-6}, we found that 1,278 patterns were significantly different in their usage (see Additional files 2 and 3). This demonstrated that the differences between the usage frequencies of these *k*-mers could largely differentiate the two groups. Therefore, our improved *k*-mer scheme is a suitable algorithm to distinguish mRNAs and lncRNAs. 10-fold cross-validation on the human training datasets was performed. The accuracy was 95.6%. Although PLEK was not better than the state-of-the-art CNCI tool on the dataset (CNCI's accuracy was 97.3% on human RefSeq and GENCODE datasets), it worked better on transcriptome data from PacBio and 454 datasets (see section entitled 'Performance on PacBio and 454 datasets').

Performance in cross-species prediction

At present, genomic sequences and annotations of most organisms are of poor quality or are unavailable. To analyze the transcriptome data of these organisms, we could draw wide support from the well-annotated related organisms in a cross-species manner. For example, we could try using the models built by human training data to analyze data from other vertebrates that have not been deeply explored.

We tested PLEK on several other vertebrates to assess its performance in cross-species prediction of protein-coding and non-coding transcripts, and found that it worked well (Table 1). This result demonstrated that PLEK exhibits good performance in cross-species prediction, as the CNCI does, which performed uniformly on all the species of vertebrates [17]. We also found that the more similar the genome of a vertebrate was to that of human, the better the performance of the model. Therefore, for species without reference genomes or with poor annotation information, one could use transcripts and annotation of closely related organisms to build models to distinguish their protein-coding and non-coding transcripts.

Robustness to indel sequencing errors

We applied PLEK to human protein-coding transcripts with simulated indel sequencing errors to evaluate its robustness and compare its performance with that of CPC and CNCI. CPC is widely used to assess the protein-coding potential of a transcript based on alignment with a protein database [14,45,46]. CNCI effectively distinguishes protein-coding and non-coding sequences independent of reference genomes and known annotations by profiling adjoining nucleotide triplets (ANT). CNCI provides highly accurate prediction of transcripts assembled from RNA-seq data in a cross-species manner. In our study, human protein-coding transcripts were extracted from the RefSeq database and the overlapping

transcripts with the training set were removed. Using the indel error simulation approach, we examined robustness performance across different indel sequencing error rates from 0% to 3% (Figure 1). The results showed that the accuracies of PLEK and CPC were not affected by a small amount of indels, whereas the accuracy of CNCI decreased significantly with increasing indel error rates, indicating that PLEK is a robust tool for distinguishing protein-coding and non-coding transcripts with homopolymer-associated indel sequencing errors.

Performance on PacBio and 454 datasets

We compared the performance of PLEK with that of CNCI and CPC using *de novo* assembled transcripts derived from PacBio and 454 platforms. 3,306 transcripts from the MCF-7 transcriptome sequenced by a PacBio platform and 3,098 transcripts from HelaS3 transcriptome generated by a 454 platform were fed into these three tools, respectively. PLEK maintained its high accuracy, 0.947 on the PacBio dataset and 0.954 on the 454 dataset, which was higher than that of CNCI (0.913 and 0.937, respectively) (Table 2). CPC achieved the highest accuracy (>0.970), but the lowest specificity (<0.472). CPC remained high positive and negative predictive values (PPV and NPV) (>0.926). PLEK and CNCI had relatively better PPV (>0.991) but poor NPV (<0.407). Sensitivity, specificity, PPV, NPV and accuracy were calculated using the following formulae:

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

$$PPV = \frac{TP}{TP + FP}; \quad NPV = \frac{TN}{TN + FN};$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

FN, false negative; *FP*, false positive; *TN*, true negative; *TP*, true positive.

Performance comparison on mouse datasets

We compared the performance of PLEK with that of CNCI, CPC and PhyloCSF on the mouse datasets which were composed of 6,015 lncRNAs and a random sample of 6,015 mRNAs. We evaluated these four tools and measured their accuracy on protein-coding and long non-coding transcripts, respectively. Figure 2 shows the fraction of transcripts that were classified as coding or non-coding by each tool. On protein-coding transcripts, all these tools performed well. Only a small fraction of protein-coding transcripts were misclassified as non-coding (Figure 2A). On non-coding transcripts, PLEK and CNCI outperformed PhyloCSF and CPC. At least 22% non-coding transcripts were misclassified as coding by PhyloCSF and CPC (Figure 2B). These results indicate that PLEK is a reasonably efficient tool.

Computational performance

Advances in sequencing technologies have produced huge amount of transcriptome data. To date (Mar 17, 2014), there are approximately 4,158 studies of transcriptome analysis in the Sequence Read Archive (SRA). 1,606 of these studies generated more than 10 Gb bases, and 234 of them produced more than 100 Gb bases. The scale of such data will become much larger when we comprehensively analyze the RNA-seq data from different studies. Thus it is necessary to develop a high-speed tool to separate lncRNAs from mRNAs in large-scale transcriptome data.

We measured computation time of PLEK, CNCI, CPC and PhyloCSF on a sample of 1,000 protein-coding

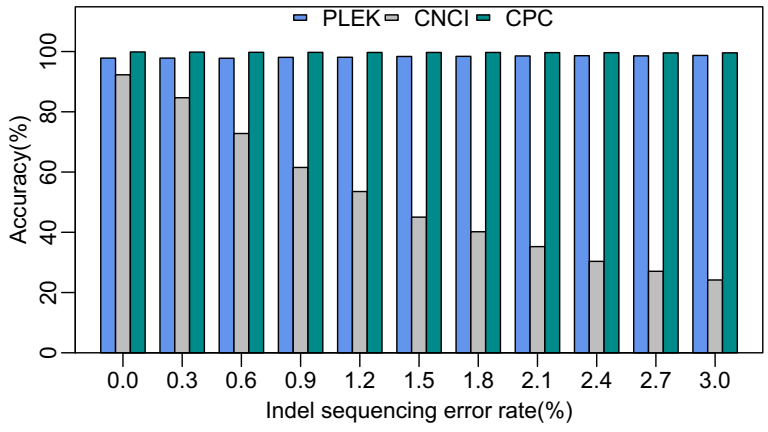

Figure 1 Comparison of robustness towards indel sequencing errors. The x-axis is the indel numbers per 100 bases (indel sequencing error rates). Performance (accuracy) of CNCI declines significantly as the indel error rate increases.

Table 2 Performances on transcripts derived from PacBio and 454

Dataset	Tool	Sensitivity	Specificity	PPV	NPV	Accuracy
MCF-7 (PacBio)	PLEK	0.947	**0.958**	**0.998**	0.407	0.947
	CPC	**0.999**	*0.190*	0.970	**0.958**	**0.970**
	CNCI	0.918	0.787	0.991	*0.269*	0.913
HelaS3 (454)	PLEK	0.955	**0.925**	**0.999**	0.262	0.954
	CPC	**0.999**	*0.472*	0.991	**0.926**	**0.990**
	CNCI	0.939	0.811	0.997	*0.189*	0.937

Bold face type indicates the best performances (sensitivity, specificity, PPV, NPV, accuracy) among PLEK, CPC and CNCI. Italic face type indicates the worst specificity and NPV among these tools.

sequences and 1,000 long non-coding sequences randomly selected from human RefSeq and GENCODE v17 databases, respectively. All these tools were run in a single-threading manner on an IBM X3650 M4 computer equipped with two E5-2640 CPUs and 64 GB of RAM. PLEK took 128 seconds to process the data, which was approximately eightfold faster than CNCI (1,048 seconds), 244 times than CPC (31,247 seconds) and 1,421 times than PhyloCSF (181,925 seconds) (Table 3). Additionally, PLEK could be easily configured for multi-threading parallel computing, which will further save computation time. Thus, PLEK is especially suitable for classifying a large number of transcripts conducted by RNA-seq technologies.

Discussion

Several studies have identified numerous lncRNAs from RNA-seq data [43,45,47-51]. However, the transcriptomes of many species, with partial or missing reference genomes, have been studied using PacBio or 454 sequencing techniques [32-38]. PacBio and 454 platforms generate longer read lengths than Illumina, which make it easier to assemble *de novo* transcripts without reference genomes [31,38]; however, transcripts generated by these platforms have relatively higher indel sequencing error rates [19,39]. Furthermore, an increasing number of RNA-seq data have been generated and the data scale is expanding rapidly with advances in high-throughput sequencing technologies. Therefore, it is necessary to develop a tool independent of known annotations and suitable for cross-species prediction that is robust to indel sequencing errors and fast enough to be affordable for large-scale data.

Appropriate computational features are very important for classification. Although conventional *k*-mer feature has been employed in several studies, such as gene identification [52], piRNA prediction [53], class-specific motif detection [54] and miRNA classification [55], we found that the proposed improved *k*-mer usage frequencies were good features to identify lncRNAs.

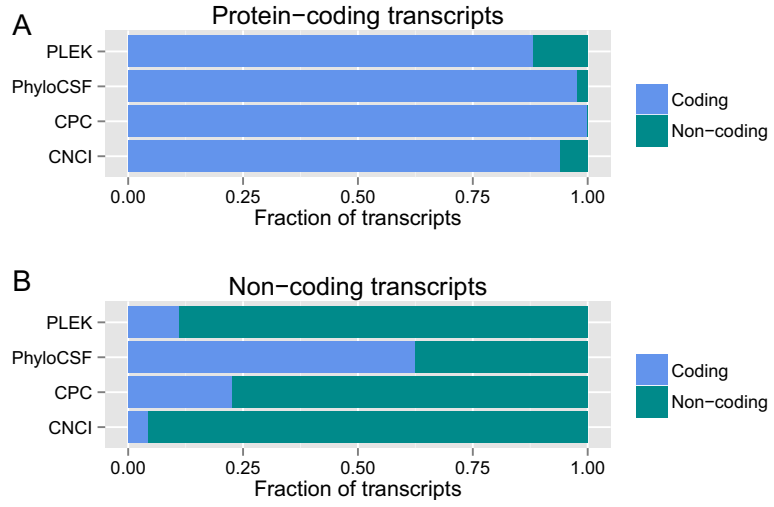

Figure 2 Results of PLEK, CPC, CNCI and PhyloCSF on mouse datasets. (A) The fraction of protein-coding transcripts classified as coding or non-coding. **(B)** The fraction of non-coding transcripts classified as coding or non-coding. Data were collected from RefSeq mouse protein-coding transcripts (release 60) and GENCODE mouse long non-coding transcripts (vM2). Shown is the fraction of transcripts classified as coding or non-coding by each tool. All these tools performed well on protein-coding transcripts. PLEK and CNCI outperformed CPC and PhyloCSF on long non-coding transcripts.

Table 3 Comparison of computational performances of PLEK, CNCI, CPC and PhyloCSF

Performance	PLEK	CNCI	CPC	PhyloCSF
Run time[a] (seconds)	128	1048	31247	181925[e]
Multi-threading[b]	Yes	Yes	No[d]	No
Online running[c]	No	No	Yes	No

Computational time was tested on 1,000 human mRNA transcripts and 1,000 human lncRNA transcripts.
[a]Computation time consumed when run in a single-threading manner.
[b]Can the software tool run in a multi-threading manner?
[c]Does the software tool provide a website for users to run online?
[d]CPC improves its computational performance using a page-cache method on its website.
[e]Bed files were load onto the Galaxy webserver (http://galaxy.nbic.nl/) and the tool 'Stitch Gene Blocks' was used to retrieve multiple alignment files with sequence entries for the following genome builds based on the 10-way Multiz alignment to *hg19*: *hg19*, *panTro2*, *tarSyr1*, *micMur1*, *otoGar1* and *rheMac2*. PhyloCSF was run using the options: *–removeRefGaps*.

More features would be used for prediction with increasing *k*. For example, 340 features are used when *k* ranges from 1 to 4, 1,364 features when *k* ranges from 1 to 5, and so on. Prediction accuracy increases with the increasing *k*; however, this is accompanied by an increasing computation load (Figure 3). As we increased *k*, and thus added more sequence features, we were able to better discriminate between lncRNA and mRNA sequences. *K*-mers with lengths greater than five do not significantly change the discrimination power of the SVM. Thus, for a trade-off between computational time and accuracy, we determined the range of *k* as 1 to 5. The model, PLEK, which uses these features, attained high prediction accuracy in 10-fold cross-validation on a human training dataset. PLEK runs faster than previous tools, CPC and CNCI, because CPC is an alignment-based tool and CNCI is obliged to calculate Hexamer ($k = 6$) usage.

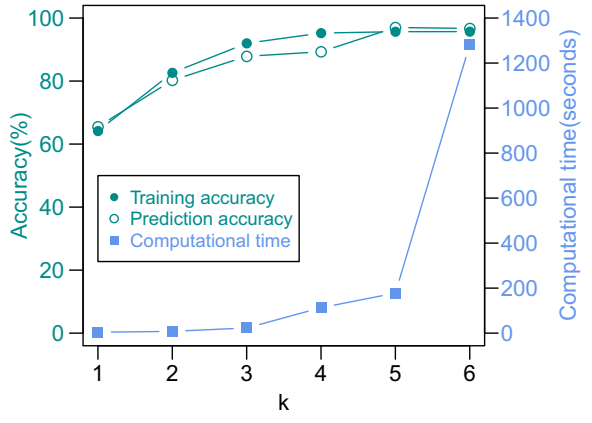

Figure 3 Performance comparison of various ranges of *k*. On the x-axis, '5' means that *k* ranged from 1 to 5. Training data comprised 22,389 human RefSeq mRNA transcripts and 22,389 GENCODE lncRNA transcripts. SVM classifiers were trained using 10-fold cross-validation on the training datasets. The figure indicates that the computation load rises and the accuracy increases along as *k* increases.

We multiplied the *k*-mer usage frequencies by various calibration factors (formula 3), which gave more weightage to those *k*-mer strings that are longer in length. Intuitively, *k*-mer usage frequencies generally decrease with the increase of the size of *k*-mer strings. For instance, the frequency of nucleotide *G* is theoretically around 0.25, while that of *GGGGG* is about 0.001. However, longer *k*-mer strings may contain more information than shorter ones: several studies indirectly use Hexamer and ANT ($k = 6$) usage as features to classify coding from non-coding RNAs [17,56]. We multiplied usage frequencies of longer *k*-mer strings by relatively larger factors for calibration, hence outweighing longer *k*-mer strings, and thereby improving the weights of the corresponding features.

The step-lengths of sliding-windows in a previous tool, CNCI, are fixed at three [17,53]. A single-base indel sequencing error in a protein-coding sequence may result in a false shift in reading frame [57], which can dramatically affect the performance of CNCI (Figure 1). In contrast, PLEK uses a sliding-window of size *k*, where *k* ranges from 1 to 5, to slide along a nucleotide sequence from its 5′ to 3′ end to count the occurrence number of *k*-mer strings. Frameshifts do not exert a strong influence on the calculation of *k*-mer usage when using a sliding-window of one-nucleotide step-length. This is why the results showed that our approach deals with indel sequencing errors robustly. This feature was also confirmed by the test on the real PacBio and 454 datasets (Table 2). On these datasets, the specificity of PLEK was the highest among these tools. Moreover, PLEK achieved an optimal balance between high specificity and high sensitivity (0.946 and 0.942 on PacBio, 0.955 and 0.925 on 454). PLEK thoroughly outperformed CNCI. Although CPC produced the highest sensitivity, it suffered from poor specificity (0.190 on PacBio, 0.472 on 454). CPC was more likely to misclassify lncRNAs as mRNAs, with high false-positive rates. There were large amount of mRNAs (96.3% in PacBio and 98.3% in 454) and few lncRNAs (3.7% in PacBio and 1.7% in 454) in these datasets, in conjunction with CPC's high sensitivity, which led to the high accuracy of CPC on the test data.

Compared with CPC, the sensitivity and specificity of PLEK is well balanced (Table 2). On the real mouse datasets including the same number of mRNA and lncRNA transcripts, PLEK could obtain high PPV and NPV (Figure 2). However, NPV of PLEK is probably low on highly imbalanced datasets or organisms with compact genomes. The number of coding RNAs is at least two orders of magnitude greater than that of non-coding RNAs on the real PacBio and 454 datasets we used in this study. In this situation, as most of the transcripts are coding, the prediction value over the transcripts called as non-coding is very low. Even quite a small portion, say 1%, of misclassification of mRNAs is likely to

result in remarkable decrease of NPV (Table 2). PLEK produces a decision value for each transcript and then labels it according to this criterion: the cut-off of discriminating transcripts is 0 by default, those with >0 decision values are labelled as mRNAs, and <0 as lncRNAs. Generally, greater absolute decision values indicate greater confidence in the prediction. Therefore, we can apply various criteria in different conditions to achieve satisfactory performance (high PPV or NPV). Similar criteria were also used or established in several researches using CPC and PhyloCSF [49,51,58]. Another method of transcending this limitation of PLEK is to perform additional analyses, such as orthology analyses of the ORFs contained in transcripts, scanning for known protein domains, etc. [45,49,58].

The classifier we currently released on Sourceforge was built on human training data. Although it obtained well performance on most vertebrates (Table 1), it might not be directly suitable to other species with very different sequence composition to human, such as *Drosophila melanogaster*, *Caenorhabditis elegans* and *Arabidopsis thaliana*. Therefore, we provided a Python script, named 'PLEKModelling.py', to assist users to build new classifiers. For the species that have not been well-explored, no sufficient reliable lncRNA transcripts are now available to build classifiers in most cases. We encourage users to use as many reliable transcripts of their relatives as possible to build classifiers. For instance, we trained a classifier using *Zea mays* Ensembl mRNA transcripts and lncRNA transcripts identified by Li *et al.* [59], and this classification model performed well on *Arabidopsis thaliana* and *Oryza sativa* datasets (Additional file 4).

Therefore, PLEK is a valuable alignment-free tool, which is accurate, robust and fast, to identify lncRNAs in *de novo* assembled transcripts without reference genomes.

Conclusions

Long non-coding RNAs are receiving increasing attention, and distinguishing lncRNAs from mRNAs in *de novo* sequencing data without reference genomes represents a challenge. To solve this problem, we designed an alignment-free tool, PLEK, which is based on an improved *k*-mer scheme. The computation of *k*-mer usage is different from that used in previous studies [53,55]: the step-lengths of the sliding-windows are one nucleotide, and the *k*-mer usage is calibrated according to the size of *k*-mer strings. PLEK worked well on human training data and in a cross-species manner on other vertebrates using the model built from human training data. It also performed well on simulated transcripts and real *de novo* assembled PacBio and 454 transcriptomes, all of which include relatively high levels of indel sequencing errors than data generated by Illumina platforms. PLEK struck a better balance between high specificity and high sensitivity than CPC on PacBio and 454 sequencing data. In addition, PLEK runs at least eightfold faster than previous available tools, CPC and CNCI. The results demonstrated that PLEK is particularly suited to transcriptome data with indel sequencing errors and growing large-scale transcriptome datasets. Thus, PLEK is a useful tool for distinguishing protein-coding and non-coding sequences from high-throughput sequencing data of many species without reference genomes.

Availability and requirements

Project name: PLEK.
Project home page: https://sourceforge.net/projects/plek/ website.
Operating systems(s): Linux.
Programming language: C, Python.
Other requirements: gcc, g++, Python.
License: GNU Public License version 3 (GPLv3).
Any restrictions to use by non-academics: none.

Additional files

Additional file 1: Demonstration of indel sequencing error simulation.

Additional file 2: Features and their P-values of rank-sum test.

Additional file 3: Visualization of the *k*-mer usage differences between mRNAs and lncRNAs.

Additional file 4: Building a new classifier using PLEKModelling.py.

Abbreviations

454: Roche 454; ANT: Adjoining nucleotide triplets; AUC: Area under the curve; bp: Base pair; CNCI: Coding-Non-Coding Index; CPC: Coding Potential Calculator; lncRNA: Long non-coding RNA; mRNA: Messenger RNA; NPV: Negative predictive value; nt: Nucleotide; PacBio: Pacific Biosciences; PLEK: Predictor of long non-coding RNAs and messenger RNAs based on an improved *k*-mer scheme; PPV: Positive predictive value; SVM: Support vector machine.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JYZ and AML participated in the design of the study. AML carried out the experiments and performed the statistical analysis. ZYZ participated in comparing PLEK against other exiting tools on mouse datasets. JYZ and AML drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We thank Professor Junying Zhang's team members for technical discussions and helpful comments. We also thank the anonymous reviewers of the manuscript for their helpful comments and suggestions. This work was supported by the Natural Science Foundation of China under Grants 61070137, 91130006, 61201312 and 61303122; and the Research Fund for the Doctoral Program of Higher Education of China (20130203110017). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author details

[1]School of Computer Science and Technology, Xidian University, Xi'an, PR China. [2]School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, PR China. [3]Department of Molecular and Cell Biology, School of Life Sciences, University of Science and Technology of China, Hefei, PR China. [4]State Key Laboratory of Genetic Resources and Evolution,

Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, PR China.

Received: 18 November 2013 Accepted: 1 September 2014
Published: 19 September 2014

References

1. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: **The transcriptional landscape of the yeast genome defined by RNA sequencing.** *Science* 2008, **320**(5881):1344–1349.
2. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10**(1):57–63.
3. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol* 2010, **28**(5):511–515.
4. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**(7):621–628.
5. Flintoft L: **Non-coding RNA: Structure and function for lncRNAs.** *Nat Rev Genet* 2013, **14**(9):598.
6. Mercer TR, Dinger ME, Mattick JS: **Long non-coding RNAs: insights into functions.** *Nat Rev Genet* 2009, **10**(3):155–159.
7. Tripathi V, Shen Z, Chakraborty A, Giri S, Freier SM, Wu X, Zhang Y, Gorospe M, Prasanth SG, Lal A, Prasanth KV: **Long noncoding RNA MALAT1 controls cell cycle progression by regulating the expression of oncogenic transcription factor B-MYB.** *PLoS Genet* 2013, **9**(3):e1003368.
8. Wang X, Arai S, Song X, Reichart D, Du K, Pascual G, Tempst P, Rosenfeld MG, Glass CK, Kurokawa R: **Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription.** *Nature* 2008, **454**(7200):126–130.
9. Batista PJ, Chang HY: **Long noncoding RNAs: cellular address codes in development and disease.** *Cell* 2013, **152**(6):1298–1307.
10. Wapinski O, Chang HY: **Long noncoding RNAs and human disease.** *Trends Cell Biol* 2011, **21**(6):354–361.
11. Yang L, Lin C, Jin C, Yang JC, Tanasa B, Li W, Merkurjev D, Ohgi KA, Meng D, Zhang J, Evans CP, Rosenfeld MG: **lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs.** *Nature* 2013, **500**(7464):598–602.
12. Schmitt AM, Chang HY: **Gene regulation: Long RNAs wire up cancer growth.** *Nature* 2013, **500**(7464):536–537.
13. Qi P, Du X: **The long non-coding RNAs, a new cancer diagnostic and therapeutic gold mine.** *Mod Pathol* 2013, **26**(2):155–165.
14. Ulitsky I, Bartel David P: **LincRNAs: genomics, evolution, and mechanisms.** *Cell* 2013, **154**(1):26–46.
15. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G: **CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine.** *Nucleic Acids Res* 2007, **35**(Web Server issue):W345–W349.
16. Lin MF, Jungreis I, Kellis M: **PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions.** *Bioinformatics* 2011, **27**(13):i275–i282.
17. Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, Liu Y, Chen R, Zhao Y: **Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts.** *Nucleic Acids Res* 2013, **41**(17):e166.
18. Meyer M, Stenzel U, Hofreiter M: **Parallel tagged sequencing on the 454 platform.** *Nat Protoc* 2008, **3**(2):267–278.
19. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ: **Performance comparison of benchtop high-throughput sequencing platforms.** *Nat Biotechnol* 2012, **30**(5):434–439.
20. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y: **A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.** *BMC Genomics* 2012, **13**:341.
21. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, Gibbs RA: **Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology.** *PLoS One* 2012, **7**(11):e47768.
22. Pruitt KD, Tatusova T, Brown GR, Maglott DR: **NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.** *Nucleic Acids Res* 2012, **40**(Database issue):D130–D135.
23. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2007, **35**(Database issue):D61–D65.
24. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R: **GENCODE: producing a reference annotation for ENCODE.** *Genome Biol* 2006, **7 Suppl 1**:S4. 1-9.
25. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigo R: **The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression.** *Genome Res* 2012, **22**(9):1775–1789.
26. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, *et al*: **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome Res* 2012, **22**(9):1760–1774.
27. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Garcia-Giron C, Gordon L, Hourlier T, Hunt S, Juettemann T, Kahari AK, Keenan S, Komorowska M, Kulesha E, Longden I, Maurel T, McLaren WM, Muffato M, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, *et al*: **Ensembl 2013.** *Nucleic Acids Res* 2013, **41**(Database issue):D48–D55.
28. Chang C-C, Lin C-J: **LIBSVM: A library for support vector machines.** *ACM Trans Intell Syst Technol* 2011, **2**(3):1–27.
29. Martin JA, Wang Z: **Next-generation transcriptome assembly.** *Nat Rev Genet* 2011, **12**(10):671–682.
30. Schuster SC: **Next-generation sequencing transforms today's biology.** *Nat Methods* 2008, **5**(1):16–18.
31. Mason CE, Elemento O: **Faster sequencers, larger datasets, new challenges.** *Genome Biol* 2012, **13**(3):314.
32. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH: **Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing.** *Mol Ecol* 2008, **17**(7):1636–1647.
33. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, *et al*: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437**(7057):376–380.
34. Hale MC, McCormick CR, Jackson JR, Dewoody JA: **Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (Acipenser fulvescens): the relative merits of normalization and rarefaction in gene discovery.** *BMC Genomics* 2009, **10:**203.
35. Adamidi C, Wang Y, Gruen D, Mastrobuoni G, You X, Tolle D, Dodt M, Mackowiak SD, Gogol-Doering A, Oenal P, Rybak A, Ross E, Sanchez Alvarado A, Kempa S, Dieterich C, Rajewsky N, Chen W: **De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics.** *Genome Res* 2011, **21**(7):1193–1200.
36. Zeng S, Xiao G, Guo J, Fei Z, Xu Y, Roe BA, Wang Y: **Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, Epimedium sagittatum (Sieb. Et Zucc.) Maxim.** *BMC Genomics* 2010, **11:**94.
37. Renaut S, Nolte AW, Bernatchez L: **Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (Coregonus spp. Salmonidae).** *Mol Ecol* 2010, **19 Suppl 1:**115–131.
38. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Adam MP: **Hybrid error correction and de novo assembly of single-molecule sequencing reads.** *Nat Biotechnol* 2012, **30**(7):693–700.
39. Luciani F, Bull RA, Lloyd AR: **Next generation deep sequencing and vaccine design: today and tomorrow.** *Trends Biotechnol* 2012, **30**(9):443–452.
40. **PacBio blog, data release, human MCF-7 transcriptome.** [http://blog.pacificbiosciences.com/2013/12/data-release-human-mcf-7-transcriptome.html]
41. Tilgner H, Raha D, Habegger L, Mohiuddin M, Gerstein M, Snyder M: **Accurate identification and analysis of human mRNA isoforms using deep long read sequencing.** *Genes Genome Genet* 2013, **3**(3):387–397.

42. Chou H-H, Holmes MH: **DNA sequence quality trimming and vector removal.** *Bioinformatics* 2001, **17**(12):1093–1104.
43. Tan MH, Au KF, Yablonovitch AL, Wills AE, Chuang J, Baker JC, Wong WH, Li JB: **RNA sequencing reveals a diverse and dynamic repertoire of the Xenopus tropicalis transcriptome over development.** *Genome Res* 2013, **23**(1):201–216.
44. Goecks J, Nekrutenko A, Taylor J: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome Biol* 2010, **11**(8):R86.
45. Sigova AA, Mullen AC, Molinie B, Gupta S, Orlando DA, Guenther MG, Almada AE, Lin C, Sharp PA, Giallourakis CC, Young RA: **Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells.** *Proc Natl Acad Sci U S A* 2013, **110**(8):2876–2881.
46. Gao G, Vibranovski MD, Zhang L, Li Z, Liu M, Zhang YE, Li X, Zhang W, Fan Q, Vankuren NW, Long M, Wei L: **A long-term demasculinization of X-linked intergenic noncoding RNAs in Drosophila melanogaster.** *Genome Res* 2014, **24**(4):629–638.
47. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL: **Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.** *Genes Dev* 2011, **25**(18):1915–1927.
48. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A: **Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.** *Nat Biotechnol* 2010, **28**(5):503–510.
49. Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn JL, Regev A, Schier AF: **Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis.** *Genome Res* 2012, **22**(3):577–591.
50. Young RS, Marques AC, Tibbit C, Haerty W, Bassett AR, Liu JL, Ponting CP: **Identification and properties of 1,119 candidate lincRNA loci in the Drosophila melanogaster genome.** *Genome Biol Evol* 2012, **4**(4):427–442.
51. Zhou Z-Y, Li A-M, Adeola AC, Liu Y-H, Irwin DM, Xie H-B, Zhang Y-P: **Genome-wide identification of long intergenic noncoding RNA genes and their potential association with domestication in pigs.** *Genome Biol Evol* 2014, **6**(6):1387–1392.
52. Liu Y, Guo J, Hu G, Zhu H: **Gene prediction in metagenomic fragments based on the SVM algorithm.** *BMC Bioinformatics* 2013, **14 Suppl 5**:S12.
53. Zhang Y, Wang X, Kang L: **A k-mer scheme to predict piRNAs and characterize locust piRNAs.** *Bioinformatics* 2011, **27**(6):771–776.
54. Srinivasan SM, Vural S, King BR, Guda C: **Mining for class-specific motifs in protein sequence classification.** *BMC Bioinformatics* 2013, **14**:96.
55. Ding J, Zhou S, Guan J: **miRFam: an effective automatic miRNA classification method based on n-grams and a multiclass SVM.** *BMC Bioinformatics* 2011, **12**:216.
56. Fickett JW, Tung CS: **Assessment of protein coding measures.** *Nucleic Acids Res* 1992, **20**(24):6441–6450.
57. Garcia-Diaz M, Kunkel TA: **Mechanism of a genetic glissando: structural biology of indel mutations.** *Trends Biochem Sci* 2006, **31**(4):206–214.
58. Nam J-W, Bartel DP: **Long noncoding RNAs in C. elegans.** *Genome Res* 2012, **22**(12):2529–2540.
59. Li L, Eichten SR, Shimizu R, Petsch K, Yeh C-T, Wu W, Chettoor AM, Givan SA, Cole RA, Fowler JE: **Genome-wide discovery and characterization of maize long non-coding RNAs.** *Genome Biol* 2014, **15**(2):R40.

doi:10.1186/1471-2105-15-311
Cite this article as: Li *et al.*: PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved *k*-mer scheme. *BMC Bioinformatics* 2014 **15**:311.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central