Computational Molecular Biology

2013, Vol.3, No.4, 24-30 http://cmb.sophiapublisher.com

BioPublisher©

Research Report Open Access

Predicting Long Non-coding RNAs Based on Genomic Sequence Information

Jie Lv[1], Hongbo Liu[1], Hui Liu[1], Qiong Wu[1], Yan Zhang[2]

1. School of Life Science and Technology, Harbin Institute of Technology, Harbin, 150001, China
2. College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China

Corresponding Author email: kigo@hit.edu.cn/yanyou1225@gmail.com; Author

Computational Molecular Biology, 2013, Vol.3, No.4 doi: 10.5376/cmb.2013.03.0004

Abstract The binary classification of coding and non-coding genes is simplified near to 50 years. Genome-wide transcriptome studies have revealed that there exist tens of thousands of long non-coding RNAs (lncRNAs), while the functions are being uncovered slowly. Accurate identification of lncRNAs is the initial step to the systematic characterization of lncRNAs. The diversity of transcription patterns for lncRNAs challenges the available non-coding RNA prediction algorithms. Until now, prediction of lncRNAs mostly relies on genomic sequence and cross-species alignment information. Here, we review the main strategies that can discriminate lncRNA from protein-coding transcripts. Especially, recently available machine learning algorithms are shown efficient to the rapid and accurate identification of lncRNAs from a large number of putative lncRNAs based on transcriptome assembled transcripts, which would provide the basis of understanding of lncRNA biology.

Keywords Next-Generation sequencing; Prediction; Computational approaches; Machine learning; RNA-Seq

Background

Many studies have shown that the transcriptomes of mammalian genomes are more pervasive and complex than previously anticipated (Kapranov et al., 2007b; Djebali et al., 2012). It has become known that most of the mammalian genomes are transcribed, once referred to as “dark matter” (Johnson et al., 2005).

Surprisingly, the noncoding transcriptomes are receiving more and more attentions until recent few years (Maher, 2012). In last few years, the discovery and functional analysis of a large number of small RNAs (length <200 nt) have dominated the non-coding RNA field. These small RNAs can further be grouped into distinct categories (e.g., miRNAs, piRNAs, and endogenous siRNAs) based on molecular biology features such as genomic, structural and translational features (Dinger et al., 2008). In contrast, the number of lncRNAs (>200 nt) appears to be even larger than small RNAs, which is also revealed by tiling array studies (Kapranov et al., 2007a). Though some of lncRNAs may be precursors of small RNAs, it is believed that many lncRNAs are transcribed as independent transcripts either polyadenylated or non-polyadenylated (Kiyosawa et al., 2005), however non-polyadenylated RNAs have not been well studied to date. Non-polyadenylated transcripts may harbor a large number of lncRNAs, which may be underrepresented in early ESTs and cDNA data. Currently, no tool in principle allows the reliable identification of both of long (>200 nt) and short (<200 nt) transcripts, as the biology is quite different (Solda et al., 2009). To date, many tools and algorithms are available for lncRNA prediction.

The biological importance of the lncRNAs may still be underestimated, though lncRNAs habour regulatory functions of different kinds (Prasanth and Spector, 2007). This is partly due to that lncRNAs are similar to protein-coding mRNAs in genomic sequences and lack obvious features to distinguish from other categories of non-coding RNAs. Few lncRNAs

Preferred citation for this article:

Zhang et al., 2013, Predicting Long Non-coding RNAs Based on Genomic Sequence Information, Computational Molecular Biology, Vol.3, No.4 24-30 (doi: 10.5376/cmb.2013.03.0004)

Received: 24 Nov., 2013 | Accepted: 10 Dec., 2013 | Published: 27 Dec., 2013

are characterized functionally and the regulatory importance of lncRNAs is still under debate. Large-scale identifications of lncRNAs have met dilemma that the overlap rate between different projects is generally poor, though using similar cDNA library construction (Carninci et al., 2005; Imanishi et al., 2004), highlighting the difficulties to discriminate lncRNAs from protein-coding RNAs. Computational approaches and metrics are candidate approaches that are efficient to identify lncRNAs from genomic sequences and distinguish from protein-coding transcripts are discussed here. We list the available software tools that are easy to use in Table 1.

Table 1 A subset of softwares available for lncRNA identification

Software tool	Web address
CRITICA (Badger and Olsen, 1999)	http://www.ttaxus.com/software.html
ESTScan (Lottaz et al., 2003)	http://myhits.isb-sib.ch/cgi-bin/estscan
CPC (Kong et al., 2007)	http://cpc.cbi.pku.edu.cn/
PORTRAIT (Arrial et al., 2009)	http://bioinformatics.cenargen.embrapa.br/portrait/
RNAcode (Washietl et al., 2011)	http://wash.github.io/rnacode/
CNCI (Sun et al., 2013b)	http://www.bioinfo.org/software/cnci/
CPAT (Wang et al., 2013)	http://lilab.research.bcm.edu/cpat/index.php
iSeeRNA (Sun et al., 2013a)	http://sunlab.lihs.cuhk.edu.hk/iSeeRNA/

1 Basic Strategy to Discriminate between lncRNAs from Protein-coding RNAs

There are mainly two approaches that are commonly used for distinguishing lncRNA from protein-coding RNA sequences, open reading frame (ORF)-based approaches and comparative genomic analysis-based approaches.

1.1 Open reading frame (ORFs) length

ORF length is the mostly commonly used approach to distinguish lncRNAs from protein-coding RNAs and is still widely employed in recent algorithms. By chance, putative ORFs in non-coding RNAs are expected to be significantly shorter than protein-coding RNAs (Dinger et al., 2008; Solda et al., 2009). The threshold of 300 nt (putative 100 codons) is often used to screen for protein-coding RNAs. In accordance with this, >95% of protein sequences in Swiss-Prot database have >100 aa in length. The threshold seems somewhat arbitrary as for a few well-characterized lncRNAs, *H19*, *Xist*, *Gtl2* and *Kcnq1ot1*, because all have putative ORFs >100 codons, violating the rule based on this threshold (Dinger et al., 2008; Solda et al., 2009). Therefore, the ORF length metric appears problematic under the given cutoff value. In addition, some relatively short proteins (<100 aa) may by chance be misclassified as lncRNAs.

1.2 ORF conservation

An alternative approach to overcome the problems of ORF length is to assess the similarity to known proteins or protein domains for the putative ORFs of potential lncRNA transcripts, as occurrence of ORF conservation in given transcripts may be indicative of *bona fide* lncRNAs, which would differ from those without cross-species orthologs that may be evolved randomly. Many studies treat transcripts lacking of ORF conservation as lncRNAs. BLASTX (Gish and States, 1993), rsCDS (Furuno et al., 2003), Pfam (Punta et al., 2012) and SUPERFAMILY (Gough et al., 2001) are programs based on ORF conservation information. More useful, CSTminer (Castrignano et al., 2004) has the ability to screen for lncRNAs from transcriptomes. The predictions based on ORF conservation are problematic as this approach is restricted by current protein annotations and some lncRNAs such as pseudogenes have evolved from protein-coding RNAs (Duret et al., 2006).

1.3 Comparative sequence analysis

Comparative sequence analysis identify lncRNAs based on conservation of amino acid sequences in multiple genome alignments (Dinger et al., 2008; Solda et al., 2009). One applicable metric is the codon substitution frequency (CSF), which has been widely used in large-scale lncRNA identifications. This

approach is effective based on the expected probability of nucleotide substitutions in a codon between a candidate sequence and probable homologous sequences (Lin et al., 2007). However, further information that is inherent within multiple genomic sequence alignments can be exploited. phyloCSF (Lin et al., 2011), a newly developed algorithm, exploits a statistical framework to compare model based on protein-coding genes and another model with non-coding genes. Unfortunately, no automated software is available to implement the algorithm, making it hard for newcomers to follow. Similarly, RNAcode method (Washietl et al., 2011) integrates information based on nucleotide substitution frequency into a framework while without machine learning component to predict non-coding RNAs. While the comparative methods are useful to identify conserved lncRNAs, further approaches that can achieve fast identifications of lncRNA transcripts are still needed.

2 Integrative Algorithms to Discriminate lncRNAs from Protein-coding RNAs

Despite different approaches vary in principle, these approaches show broad concordance in performance (Frith et al., 2006). Yet different approaches can be combined to achieve better effects, as previous studies have demonstrated. For example, CRITICA algorithm (Badger and Olsen, 1999) that employs statistical model and comparative approach was shown to best-perform among the selected ten bioinformatic methods for the FANTOM cDNA set (Frith et al., 2006). Other algorithms use statistical approaches to integrate distinct categories of signatures, for instance, polyadenylation sites, splice sites, and sequence homology. As an example, DIANA-EST uses artificial neural network method and statistical model to discriminate coding regions (Hatzigeorgiou et al., 2001), and ESTScan employs hidden Markov model (Lottaz et al., 2003).

Recent tools, CONC (Liu et al., 2006), CPC (Kong et al., 2007), iSeeRNA(Sun et al., 2013a), CPAT (Wang et al., 2013) and CNCI (Sun et al., 2013b) use machine learning algorithms to distinguish protein-coding mRNAs from lncRNAs. These algorithms distinguish lncRNAs from protein-coding RNAs based on multiple genome-derived and other features, for instance, putative peptide length, putative amino acid composition, protein homologs, RNA secondary structure, and multi-species protein alignments.

First, CONC is an algorithm and software that can classify input transcripts as protein-coding RNAs or non-coding RNAs based on a machine learning algorithm (Liu et al., 2006). The CONC algorithm uses protein related features including RNA secondary structure, RNA solvent accessible surface area, in addition to sequence compositional entropy, peptide length, protein homology and amino acid frequency. Though CONC works well based on high-quality full-length cDNAs (Maeda et al., 2006), it is in practice slow to run for large datasets and lacks of a web interface. In addition, CONC only reports ‘coding’ or ‘non-coding’ while does not provide results with detailed explanations and other related information. CPC (Kong et al., 2007) uses three ORF related features and three BLASTX-derived features and incorporates them into Support Vector Machine (SVM) algorithm. The authors (Kong et al., 2007) used same data set (5610 protein-coding and 2670 non-coding RNAs) as CONC to obtain a trained SVM model. Though CPC used fewer signatures than CONC (6 versus 180) but comparable yet even better performance was observed. Easy-to-use web tool and standalone version of CPC are both available to use (Table 1).

Though protein homologs are very useful to improve prediction accuracy, these programs using this information may be inappropriate for prediction from neglected species such as fungus et al.. PORTAIT (Arrial et al., 2009), a software based on SVM, was specifically designed to overcome this obstacle, which takes into consideration EST sequencing errors, frameshifts and truncations information.

iSeeRNA is a recently published SVM-based standalone tool. It was demonstrated to have high accuracy, balanced specificity and sensitivity for lncRNAs. iSeeRNA is fast to run, which is an alternative tool for filtering candidate lincRNAs from transcriptome assembled data.

Coding-Non-Coding Index (CNCI) is another recent tool for lincRNA identification, by using genomic sequence derived information of adjoining nucleotide triplets (ANT). CNCI can effectively distinguish lncRNAs from protein-coding transcripts, which are especially useful for lncRNAs with incomplete ends and cis-antisense pairs. CNCI is appropriate for transcriptome assembled data from less-studied species, as CNCI can efficiently predict non-coding transcripts based solely on nucleotide frequency of transcript sequences.

Wang et al. (2013) found ORF related features and hexamer usage bias features to be efficient features for distinguishing between protein-coding and lncRNA transcript prediction and integrated them into a logistic regression model (Wang et al., 2013). Based on trained model, they developed CPAT, which had both high accuracy and speed, Typically, CPAT is efficient in running time with four orders of magnitude in speed faster than CPC and CSF algorithms, suitable to transcriptome assembled data for the ever growing RNA-seq community.

In conclusion, SVM framework seems to outperform previous non-integrative approaches by combining multiple discriminating features and currently represents the pioneering tools for non-coding RNA prediction.

However, one important problem should be noticed that the incompleteness of full-length transcript sequences which is caused by incomplete reverse transcription, genomic contamination and internal priming of pre-mRNAs in large-scale sequencing can strongly affect the accuracy of these tools. Given the low expression levels of most lncRNAs, the putative lncRNAs may not be assembled efficiently by transcriptome assembly softwares.

3 Discussion

Many years ago, genome annotation has been a challenging task when a new genome is sequenced. In recent years, the task is even more urgent due to deep transcriptome sequencing. Identifying non-coding RNA sequences is now one of the most important steps in genomic element annotations. Considering most novel lncRNAs are less conserved and species-specific than protein-coding RNAs, detecting lncRNAs via alignment-based algorithms seems impractical.

In this review, we have demonstrated that the genomic sequence and sequence derived features are the foundations of algorithms for differentiating lncRNAs from protein-coding RNAs and can efficiently reflect intrinsic properties of protein-coding and lncRNA transcript. Though different studies used different genomic sequence features, only a few discriminating genomic sequence based features are efficient to improve prediction power, and can also significantly reduced computing cost. Because the tools mentioned in this article are solely based on sequence intrinsic composition, they are potentially applicable to species with only poorly annotated information.

It should also be noted that methods that classify a given transcript into protein-coding or lncRNA category are under the assumption that the RNA functions as protein-coding or non-coding. However, in real RNA world, a large number of RNAs may be bifunctional, that is to say, they can act as proteins or as regulatory long non-coding RNAs (Dinger et al., 2008). Though the tools mentioned there are powerful in discriminating protein-coding and lncRNAs, a large number of putative lncRNAs may be falsely classified as protein-coding RNAs based on a long ORF (putative ORFs >100 codons for *H19*), which is widely employed in recent algorithms.

Given the increasing next-generation data are generated by large-scale RNA-seq technology, there is a growing interest in prediction of lncRNAs. Many tools for predicting lncRNAs are available. However, software tools with higher reliability and faster speed are also needed, which would be useful to filter biology relevant lncRNA candidates from assembled transcripts based on RNA-seq data.

Authors' contributions

JL drafted the manuscript. HBL and HL collected materials. QW and YZ conceived of the study, and participated in its design and coordination. All authors read and approved the final manuscript.

Acknowledgments

The authors thank National Natural Science Foundation of China for funding. This work is supported by the National Natural Science Foundation of China [31171383, 31271558, 31371478, 31371334].

References

Arrial R.T., Togawa R.C., and Brigido Mde M., 2009, Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis, BMC Bioinformatics, 10: 239
http://dx.doi.org/10.1186/1471-2105-10-239
PMid:19653905 PMCid:PMC2731755

Badger J.H., and Olsen G.J., 1999, CRITICA: coding region identification tool invoking comparative analysis, Mol. Biol. Evol., 16: 512-524
http://dx.doi.org/10.1093/oxfordjournals.molbev.a026133
PMid:10331277

Carninci P., Kasukawa T., Katayama S., Gough J., Frith M.C., Maeda N., Oyama R., Ravasi T., Lenhard B., Wells C., Kodzius R., Shimokawa K., Bajic V.B., Brenner S.E., Batalov S., Forrest A.R., Zavolan M., Davis M.J., Wilming L.G., Aidinis V., Allen J.E., Ambesi-Impiombato A., Apweiler R., Aturaliya R.N., Bailey T.L., Bansal M., Baxter L., Beisel K.W., Bersano T., Bono H., Chalk A.M., Chiu K.P., Choudhary V., Christoffels A., Clutterbuck D.R., Crowe M.L., Dalla E., Dalrymple B.P., De Bono B., Della Gatta G., Di Bernardo D., Down T., Engstrom P., Fagiolini M., Faulkner G., Fletcher C.F., Fukushima T., Furuno M., Futaki S., Gariboldi M., Georgii-Hemming P., Gingeras T.R., Gojobori T., Green R.E., Gustincich S., Harbers M., Hayashi Y., Hensch T.K., Hirokawa N., Hill D., Huminiecki L., Iacono M., Ikeo K., Iwama A., Ishikawa T., Jakt M., Kanapin A., Katoh M., Kawasawa Y., Kelso J., Kitamura H., Kitano H., Kollias G., Krishnan S.P., Kruger A., Kummerfeld S.K., Kurochkin I.V., Lareau L.F., Lazarevic D., Lipovich L., Liu J., Liuni S., Mcwilliam S., Madan Babu M., Madera M., Marchionni L., Matsuda H., Matsuzawa S., Miki H., Mignone F., Miyake S., Morris K., Mottagui-Tabar S., Mulder N., Nakano N., Nakauchi H., Ng P., Nilsson R., Nishiguchi S., Nishikawa S., et al., 2005, The transcriptional landscape of the mammalian genome, Science, 309: 1559-1563
http://dx.doi.org/10.1126/science.1112014
PMid:16141072

Castrignano T., Canali A., Grillo G., Liuni S., Mignone F., and Pesole G., 2004, CSTminer: a web tool for the identification of coding and noncoding conserved sequence tags through cross-species genome comparison, Nucleic Acids Res., 32: W624-627
http://dx.doi.org/10.1093/nar/gkh486
PMid:15215464 PMCid:PMC441624

Dinger M.E., Pang K.C., Mercer T.R., and Mattick J.S., 2008, Differentiating protein-coding and noncoding RNA: challenges and ambiguities, PLoS Comput Biol, 4: e1000176
http://dx.doi.org/10.1371/journal.pcbi.1000176
PMid:19043537 PMCid:PMC2518207

Djebali S., Davis C.A., Merkel A., Dobin A., Lassmann T., Mortazavi A., Tanzer A., Lagarde J., Lin W., Schlesinger F., Xue C., Marinov G.K., Khatun J., Williams B.A., Zaleski C., Rozowsky J., Roder M., Kokocinski F., Abdelhamid R.F., Alioto T., Antoshechkin I., Baer M.T., Bar N.S., Batut P., Bell K., Bell I., Chakrabortty S., Chen X., Chrast J., Curado J., Derrien T., Drenkow J., Dumais E., Dumais J., Duttagupta R., Falconnet E., Fastuca M., Fejes-Toth K., Ferreira P., Foissac S., Fullwood M.J., Gao H., Gonzalez D., Gordon A., Gunawardena H., Howald C., Jha S., Johnson R., Kapranov P., King B., Kingswood C., Luo O.J., Park E., Persaud K., Preall J.B., Ribeca P., Risk B., Robyr D., Sammeth M., Schaffer L., See L.H., Shahab A., Skancke J., Suzuki A.M., Takahashi H., Tilgner H., Trout D., Walters N., Wang H., Wrobel J., Yu Y., Ruan X., Hayashizaki Y., Harrow J., Gerstein M., Hubbard T., Reymond A., Antonarakis S.E., Hannon G., Giddings M.C., Ruan Y., Wold B., Carninci P., Guigo R., and Gingeras T.R., 2012, Landscape of transcription in human cells, Nature, 489: 101-108
http://dx.doi.org/10.1038/nature11233
PMid:22955620 PMCid:PMC3684276

Duret L., Chureau C., Samain S., Weissenbach J., and Avner P., 2006, The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene, Science, 312: 1653-1655
http://dx.doi.org/10.1126/science.1126316
PMid:16778056

Frith M.C., Bailey T.L., Kasukawa T., Mignone F., Kummerfeld S.K., Madera M., Sunkara S., Furuno M., Bult C.J., Quackenbush J., Kai C., Kawai J., Carninci P., Hayashizaki Y., Pesole G., and Mattick J.S., 2006, Discrimination of non-protein-coding transcripts from protein-coding mRNA, RNA Biol, 3: 40-48
http://dx.doi.org/10.4161/rna.3.1.2789
PMid:17114936

Furuno M., Kasukawa T., Saito R., Adachi J., Suzuki H., Baldarelli R., Hayashizaki Y., and Okazaki Y., 2003, CDS

annotation in full-length cDNA sequence, Genome Res, 13: 1478-1487
http://dx.doi.org/10.1101/gr.1060303
PMid:12819146 PMCid:PMC403693

Gish W., and States D.J., 1993, Identification of protein coding regions by database similarity search, Nat Genet, 3: 266-272
http://dx.doi.org/10.1038/ng0393-266
PMid:8485583

Gough J., Karplus K., Hughey R., and Chothia C., 2001, Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure, J Mol Biol, 313: 903-919
http://dx.doi.org/10.1006/jmbi.2001.5080
PMid:11697912

Hatzigeorgiou A.G., Fiziev P., and Reczko M., 2001, DIANA-EST: a statistical analysis, Bioinformatics, 17: 913-919
http://dx.doi.org/10.1093/bioinformatics/17.10.913
PMid:11673235

Imanishi T., Itoh T., Suzuki Y., O'donovan C., Fukuchi S., Koyanagi K.O., Barrero R.A., Tamura T., Yamaguchi-Kabata Y., Tanino M., Yura K., Miyazaki S., Ikeo K., Homma K., Kasprzyk A., Nishikawa T., Hirakawa M., Thierry-Mieg J., Thierry-Mieg D., Ashurst J., Jia L., Nakao M., Thomas M.A., Mulder N., Karavidopoulou Y., Jin L., Kim S., Yasuda T., Lenhard B., Eveno E., Yamasaki C., Takeda J., Gough C., Hilton P., Fujii Y., Sakai H., Tanaka S., Amid C., Bellgard M., Bonaldo Mde F., Bono H., Bromberg S.K., Brookes A.J., Bruford E., Carninci P., Chelala C., Couillault C., De Souza S.J., Debily M.A., Devignes M.D., Dubchak I., Endo T., Estreicher A., Eyras E., Fukami-Kobayashi K., Gopinath G.R., Graudens E., Hahn Y., Han M., Han Z.G., Hanada K., Hanaoka H., Harada E., Hashimoto K., Hinz U., Hirai M., Hishiki T., Hopkinson I., Imbeaud S., Inoko H., Kanapin A., Kaneko Y., Kasukawa T., Kelso J., Kersey P., Kikuno R., Kimura K., Korn B., Kuryshev V., Makalowska I., Makino T., Mano S., Mariage-Samson R., Mashima J., Matsuda H., Mewes H.W., Minoshima S., Nagai K., Nagasaki H., Nagata N., Nigam R., Ogasawara O., Ohara O., Ohtsubo M., Okada N., Okido T., Oota S., Ota M., Ota T., Otsuki T., et al., 2004, Integrative annotation of 21,037 human genes validated by full-length cDNA clones, PLoS Biol, 2: e162
http://dx.doi.org/10.1371/journal.pbio.0020162
PMid:15103394 PMCid:PMC393292

Johnson J.M., Edwards S., Shoemaker D., and Schadt E.E., 2005, Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments, Trends Genet, 21: 93-102
http://dx.doi.org/10.1016/j.tig.2004.12.009
PMid:15661355

Kapranov P., Cheng J., Dike S., Nix D.A., Duttagupta R., Willingham A.T., Stadler P.F., Hertel J., Hackermuller J., Hofacker I.L., Bell I., Cheung E., Drenkow J., Dumais E., Patel S., Helt G., Ganesh M., Ghosh S., Piccolboni A., Sementchenko V., Tammana H., and Gingeras T.R., 2007a, RNA maps reveal new RNA classes and a possible function for pervasive transcription, Science, 316: 1484-1488
http://dx.doi.org/10.1126/science.1138341
PMid:17510325

Kapranov P., Willingham A.T., and Gingeras T.R., 2007b, Genome-wide transcription and the implications for genomic organization, Nat Rev Genet, 8: 413-423
http://dx.doi.org/10.1038/nrg2083
PMid:17486121

Kiyosawa H., Mise N., Iwase S., Hayashizaki Y., and Abe K., 2005, Disclosing hidden transcripts: mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized, Genome Res, 15: 463-474
http://dx.doi.org/10.1101/gr.3155905
PMid:15781571 PMCid:PMC1074361

Kong L., Zhang Y., Ye Z.Q., Liu X.Q., Zhao S.Q., Wei L., and Gao G., 2007, CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine, Nucleic Acids Res, 35: W345-349
http://dx.doi.org/10.1093/nar/gkm391
PMid:17631615 PMCid:PMC1933232

Lin M.F., Carlson J.W., Crosby M.A., Matthews B.B., Yu C., Park S., Wan K.H., Schroeder A.J., Gramates L.S., St Pierre S.E., Roark M., Wiley K.L., Jr., Kulathinal R.J., Zhang P., Myrick K.V., Antone J.V., Celniker S.E., Gelbart W.M., and Kellis M., 2007, Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes, Genome Res, 17: 1823-1836
http://dx.doi.org/10.1101/gr.6679507
PMid:17989253 PMCid:PMC2099591

Lin M.F., Jungreis I., and Kellis M., 2011, PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions, Bioinformatics, 27: i275-282
http://dx.doi.org/10.1093/bioinformatics/btr209
PMid:21685081 PMCid:PMC3117341

Liu J., Gough J., and Rost B., 2006, Distinguishing protein-coding from non-coding RNAs through support vector machines, PLoS Genet, 2: e29

http://dx.doi.org/10.1371/journal.pgen.0020029 PMid:16683024 PMCid:PMC1449884

Lottaz C., Iseli C., Jongeneel C.V., and Bucher P., 2003, Modeling sequencing errors by combining Hidden Markov models, Bioinformatics, 19 Suppl 2: ii103-112 http://dx.doi.org/10.1093/bioinformatics/btg1067 PMid:14534179

Maeda N., Kasukawa T., Oyama R., Gough J., Frith M., Engstrom P.G., Lenhard B., Aturaliya R.N., Batalov S., Beisel K.W., Bult C.J., Fletcher C.F., Forrest A.R., Furuno M., Hill D., Itoh M., Kanamori-Katayama M., Katayama S., Katoh M., Kawashima T., Quackenbush J., Ravasi T., Ring B.Z., Shibata K., Sugiura K., Takenaka Y., Teasdale R.D., Wells C.A., Zhu Y., Kai C., Kawai J., Hume D.A., Carninci P., and Hayashizaki Y., 2006, Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs, PLoS Genet, 2: e62 http://dx.doi.org/10.1371/journal.pgen.0020062 PMid:16683036 PMCid:PMC1449903

Maher B., 2012, ENCODE: The human encyclopaedia, Nature, 489: 46-48 http://dx.doi.org/10.1038/489046a PMid:22962707

Prasanth K.V., and Spector D.L., 2007, Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum, Genes Dev, 21: 11-42 http://dx.doi.org/10.1101/gad.1484207 PMid:17210785

Punta M., Coggill P.C., Eberhardt R.Y., Mistry J., Tate J., Boursnell C., Pang N., Forslund K., Ceric G., Clements J., Heger A., Holm L., Sonnhammer E.L., Eddy S.R., Bateman A., and Finn R.D., 2012, The Pfam protein families database, Nucleic Acids Res, 40: D290-301 http://dx.doi.org/10.1093/nar/gkr1065 PMid:22127870 PMCid:PMC3245129

Solda G., Makunin I.V., Sezerman O.U., Corradin A., Corti G., and Guffanti A., 2009, An Ariadne's thread to the identification and annotation of noncoding RNAs in eukaryotes, Brief Bioinform, 10: 475-489 http://dx.doi.org/10.1093/bib/bbp022 PMid:19383843

Sun K., Chen X., Jiang P., Song X., Wang H., and Sun H., 2013a, iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data, BMC Genomics, 14 Suppl 2: S7 PMid:23445546 PMCid:PMC3582448

Sun L., Luo H., Bu D., Zhao G., Yu K., Zhang C., Liu Y., Chen R., and Zhao Y., 2013b, Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts, Nucleic Acids Res, 41: e166 http://dx.doi.org/10.1093/nar/gkt646 PMid:23892401 PMCid:PMC3783192

Wang L., Park H.J., Dasari S., Wang S., Kocher J.P., and Li W., 2013, CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model, Nucleic Acids Res, 41: e74 http://dx.doi.org/10.1093/nar/gkt006 PMid:23335781 PMCid:PMC3616698

Washietl S., Findeiss S., Muller S.A., Kalkhof S., Von Bergen M., Hofacker I.L., Stadler P.F., and Goldman N., 2011, RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data, RNA, 17: 578-594 http://dx.doi.org/10.1261/rna.2536111 PMid:21357752 PMCid:PMC3062170