研究报告
Research Report

长链非编码 RNA 基因组序列信息的预测

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

1 哈尔滨工业大学生命科学技术学院, 150001, 哈尔滨;
2 哈尔滨医科大学生物信息科学与技术学院, 150081, 哈尔滨
通讯作者: yanyou1225@gmail.com; 作者

计算分子生物学, 2013 年, 第 2 卷, 第 12 篇 doi: 10.5376/cmb.cn.2013.02.00012

推荐引用：

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)

摘 要 编码和非编码基因的二进制分类在近 50 年被简化。全基因组转录组研究表明，存在成千上万的长非编码 RNAs(lncRNAs)，与此同时，其功能也被慢慢发现。lncRNA 的准确鉴定是 lncRNA 系统表征的最初步骤。lncRNA 转录模式的多样性质疑着可用的非编码 RNA 预测算法。到目前为止，lncRNA 的预测大多依赖于基因组序列和跨物种对齐信息。在这里，我们介绍可以从编码蛋白的转录中区分 lncRNA 的主要策略。特别地，最近可用的机器学习算法表明对基于转录组装配的转录物 lncRNA 能够有效的快速，精确鉴定大量假定 lncRNA，这将提供对 lncRNA 生物学基础的理解。

关键词 下一代测序, 预测, 计算方法, 机器学习, RNA 序列

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, yanyou1225@gmail.com; Authors

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 introduce 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

许多研究已经表明哺乳动物基因组的转录组比先前预期更有说服力和复杂(Kapranov et al., 2007b; Djebali et al., 2012)。一种使得大多数哺乳动物基因组被转录的物质已经被人们所熟知，它曾经被称为“暗物质”(Johnson et al., 2005)。令人惊讶的是，直到最近几年(Maher, 2012)，非编码转录组受到越来越多的关注。在过去几年中，大量小 RNA(长度<200 nt)的发现和功能分析已经主导了非编码 RNA 领域。基于分子生物学特征，例如基因

收稿日期：2013 年 11 月 24 日
接受日期：2013 年 12 月 10 日
发表日期：2013 年 12 月 27 日
基金项目：本研究由中国国家自然科学基金会[31171383,31271558,31371478,31371334]资助。

表 1 可用于 lncRNA 识别的软件的子集

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/

组，结构和翻译特征，这些小 RNA 可进一步分组成不同类别(例如, miRNA, piRNA 和内源 siRNA) (Dinger et al., 2008)。相比之下，lncRNA 可能是小 RNA 的前体，一般认为许多 lncRNA 被转录为聚腺苷酸化或非聚腺苷酸化的独立转录物(Kiyosawa et al., 2005)，然而非聚腺苷酸化的 RNA 至今尚未得到很好的研究。非聚腺苷酸化的转录物可以携带大量的 lncRNA，其在早期 EST 和 cDNA 数据中可能没有被充分代表(Solda et al., 2009)。目前，由于生物学是完全不同的，原则上没有工具允许可靠地鉴定长(> 200 nt)和短(<200 nt)转录物。到目前为止，许多工具和算法可用于 lncRNA 预测。

虽然 lncRNA 具有不同种类的调节功能，但是 lncRNA 的生物学重要性仍然可能被低估(Prasanth and Spector, 2007)。这部分是由于 lncRNA 与基因组序列中的蛋白质编码 mRNA 相似，并且缺乏区别于其他类别的非编码 RNA 的明显特征。几个 lncRNAs 的功能特点和 lncRNAs 的监管重要性仍然存在争论。lncRNAs 的大规模鉴定遇到了两难问题，即不同项目之间的重叠率通常很差，虽然使用类似的 cDNA 文库构建(Carninci et al., 2005; Imanishi et al., 2004)，但是突出了区分 lncRNA 和蛋白质编码 RNA 的困难。计算方法和指标是从基因组序列中有效识别 lncRNAs 和区别于蛋白质编码转录本的候选方法。我们列出了表 1 中易于使用的软件工具。

1 从蛋白质编码 RNA 中区分 lncRNA 的基本策略

用于区分lncRNA和蛋白质编码RNA序列主要有两种方法：基于开放阅读框(ORF)的方法和基于比较基因组分析的方法。

1.1开放阅读框(ORF)长度

ORF长度是最常用的区分lncRNA与蛋白质编码RNA的方法，并且仍在最近的算法中广泛使用。偶然地，在非编码RNA中的推定的ORF预期明显短于蛋白质编码RNA(Dinger et al., 2008; Solda et al., 2009)。300 nt的阈值(推定的100个密码子)通常用于筛选蛋白质编码RNA。根据这一点，Swiss-Prot数据库中> 95％的蛋白质序列长度> 100 aa。对于几个良好表征的lncRNA，H19，Xist，Gtl2和Kcnq1ot1阈值似乎有些任意，因为所有具有假定的ORFs> 100密码子，都违反了基于此阈值的规则(Dinger et al., 2008; Solda et al., 2009)。因此，在给定的截断值下，ORF长度度量出现问题。此外，一些相对短的蛋白质(<100 aa)偶然被错误分类为lncRNA。

1.2 ORF保守

克服ORF长度问题的另一种方法是评估与已知蛋白质或蛋白质结构域可能的lncRNA转录物的推定ORF的相似性，因为给定转录物中ORF保守的发生可能指示真正的lncRNA，其将不同于那些没有跨物种直向同源物可以随机进化。许多研究将缺少ORF保守的转录物作为lncRNA。BLASTX(Gish and States, 1993)，rsCDS(Furuno et al., 2003)，Pfam和SUPERFAMILY(Gough et al., 2001)是基于ORF保守信息的程序。更有用的是，CGminer有能力筛选来自转录组的lncRNAs(Castrignano et al., 2004)。基于ORF保守的预测是有问题的，因为该方法受到当前蛋白质注释的限制，并且一些lncRNA如假基因从

蛋白质编码RNA(Duret et al., 2006)。

1.3比较序列分析

比较序列分析基于多个基因组比对中氨基酸序列的保守性来鉴定lncRNA。一种适用的度量是密码子取代频率(CSF)，其已经广泛用于大规模的lncRNA鉴定(Dinger et al., 2008; Solda et al., 2009)。该方法基于在候选序列和可能的同源序列之间的密码子中核苷酸取代的预期概率是有效的。然而，可以利用多个基因组序列比对其中固有的更多信息(Lin et al., 2007)。phyloCSF，一个新开发的算法，利用统计框架来比较基于蛋白质编码基因的模型和另一个模型与非编码基因(Lin et al., 2011)。不幸的是，没有自动化软件可用于实现该算法，使得新手难以遵循。类似地，RNA代码方法将基于核苷酸取代频率的信息整合到框架中，而没有机器学习组件来预测非编码RNA(Washietl et al., 2011)。尽管比较的方法可用于鉴定保守的lncRNA，但仍需要可以实现lncRNA转录物快速鉴定的其它方法。

2区分蛋白质编码RNA的lncRNAs的综合算法

尽管不同的方法在原理上不同，但是这些方法在性能上表现出广泛的一致性(Frith et al., 2006)。然而，如先前的研究所证明的，可以组合不同的方法以实现更好的效果。例如，使用统计学模型和比较方法的CRITICA算法(Badger and Olsen, 1999)显示在FANTOM cDNA集合的所选十种生物信息学方法中表现最佳(Frith et al., 2006)。其他算法使用统计学方法来整合不同类别的标记，例如聚腺苷酸化位点，剪接位点和序列同源性(Hatzigeorgiou et al., 2001)。例如，DIANA-EST使用人工神经网络方法和统计模型来区分编码区，ESTScan使用隐马尔可夫模型(Lottaz et al., 2003)。

最近的工具CONC(Liu et al., 2006)，CPC(Kong et al., 2007)，iSeeRNA(Sun et al., 2013a)，CPAT(Wang et al., 2013)和CNCI (Sun et al., 2013b)使用机器学习算法来区分蛋白质编码mRNA和lncRNAs。这些算法基于多个基因组衍生的和其他特征，例如推定的肽长度，推定的氨基酸组成，蛋白质同源物，RNA二级结构和多物种蛋白质比对，区分lncRNA和蛋白质编码RNA。

首先，CONC是一种算法和软件，可以基于机器学习算法将输入转录物分类为蛋白质编码RNA或非编码RNA(Liu et al., 2006)。CONC算法使用蛋白质相关特征，包括RNA二级结构，RNA溶剂可及表面积，除了序列组成熵，肽长度，蛋白质同源性和氨基酸频率。尽管CONC在高质量的全长cDNA上运行良好，但是对于大型数据集和缺少Web界面(Maeda et al., 2006)，运行速度很慢。此外，CONC只报告“编码”或“非编码”，但不提供具有详细解释和其他相关信息的结果。CPC使用三个ORF相关功能和三个BLASTX派生的功能，并将其纳入支持向量机（SVM）算法(Kong et al., 2007)。作者使用相同的数据集(5610蛋白编码和2670非编码RNA)作为CONC以获得训练的SVM模型(Kong et al., 2007)。尽管CPC使用的签名比CONC (6对180)少，但可以比较，CPC可以观察到更好的性能。易于使用的网络工具和独立版本的CPC都可以使用(表1)。

虽然蛋白质同源物对于提高预测准确性是非常有用的，但是使用这种信息的这些程序可能不适合于从被忽视的物种如真菌等中预测。PORTRAIT是一种基于SVM的软件(Arrial et al., 2009)，旨在克服这一障碍，考虑EST排序错误，移码和截断信息。iSeeRNA是最近发布的基于SVM的独立工具。它被证明对lncRNAs具有高准确性，平衡的特异性和灵敏度。iSeeRNA运行速度快，这是从转录组装配数据过滤候选lincRNA的替代工具。编码非编码指数(CNCI)是最近的另一种lncRNA鉴定工具，通过使用邻近核苷酸三联体(ANT)的基因组序列衍生信息。CNCI可有效区分lncRNA与蛋白质编码转录物，其对于具有不完全末端和顺反义对的lncRNA特别有用。CNCI适合于来自较少研究的物种的转录组装配数据，因为CNCI可以仅基于转录物序列的核苷酸频率有效地预测非编码转录物。Wang et al.(2013)发现ORF相关特征和六聚体使用偏差特征是区分蛋白质编码和lncRNA转录物预测的有效特征，并将它们整合到逻辑回归模型中(Wang et al., 2013)。基于训练的模型，他们开发了CPAT，其具有高精度和速度。通常，CPAT在运行时间方面比CPC和CSF算法快四个数量级的速度，适合于转录组装数据用于不断增长的RNA- seq群落。总之，SVM框架通过组合多个歧视特征似乎优于以前的非整合方法，目前代表的非编码RNA预测的开创性工具。

然而，应该注意一个重要的问题，由不完全逆

转录，基因组污染和大规模测序中pre-mRNA的内部启动引起的全长转录本序列的不完整性可以强烈影响这些工具的准确性。鉴于大多数lncRNA的低表达水平，推定的lncRNA可能不能通过转录组装配软件有效地组装。

3讨论

许多年前，当新基因组测序时，基因组注释是一项具有挑战性的任务。近年来，由于深层转录组测序，任务甚至更加迫切。鉴定非编码RNA序列现在是基因组元件注释中最重要的步骤之一。考虑到大多数新型lncRNAs比蛋白质编码RNAs保守和物种特异性，通过基于比对的算法检测lncRNA似乎不切实际。

在本研究中，我们已经证明基因组序列和序列衍生特征是区分lncRNAs从蛋白质编码RNA的算法的基础，并且可以有效地反映蛋白质编码和lncRNA转录本的固有属性。尽管不同的研究使用不同的基因组序列特征，但是仅有少数基于区分性基因组序列的特征有效地提高预测能力，并且还可以显著降低计算成本。因为本文中提到的工具仅仅基于序列内在组成，它们可能适用于仅具有很少注释的信息的物种。

还应该注意的是，将给定的转录物分类为蛋白质编码或lncRNA类别的方法是在RNA作为蛋白质编码或非编码的假设下进行的。然而，在真实的RNA世界中，大量的RNA可以是双功能的，也就是说，它们可以充当蛋白质或作为调节性长非编码RNA(Dinger et al., 2008)。尽管所提及的工具在区分蛋白质编码和lncRNA中是强有力的，但大量假定的lncRNA可基于长ORF(假定的H19的100个密码子)错误地分类为蛋白质编码RNA，其广泛应用于最近的算法。

由于增加的下一代数据是通过大规模RNA-seq技术生成的，所以人们对lncRNA的预测越来越感兴趣。用于预测lncRNA的许多工具是可得到的，但是还需要具有更高可靠性和更快速度的软件工具，这将有助于从RNA-seq数据的组装转录物中过滤出生物学相关的lncRNA候选物。

作者贡献

JL负责稿件的撰写，HBL和HL负责材料的收集。QW和YZ负责实验方案的构想、设计与操作。所有作者阅读并同意最终的文本。

致谢

感谢中国国家自然科学基金会[31171383,31271558,31371478,31371334]对本研究的资助。

参考文献

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

Badger J.H., and Olsen G.J., 1999, CRITICA: coding region identification tool invoking comparative analysis, Mol. Biol. Evol., 16: 512-524

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

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

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

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

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

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

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

Gish W., and States D.J., 1993, Identification of protein coding regions by database similarity search, Nat Genet, 3: 266-272

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

Hatzigeorgiou A.G., Fiziev P., and Reczko M., 2001, DIANA-EST: a statistical analysis, Bioinformatics, 17: 913-919

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

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

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:

Genome-wide transcription and the implications for genomic organization, Nat Rev Genet, 8: 413-423

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

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

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

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

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

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

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

1484-1488

Kapranov P., Willingham A.T., and Gingeras T.R., 2007b,

physical cDNAs, PLoS Genet, 2: e62

Maher B., 2012, ENCODE: The human encyclopaedia, Nature, 489: 46-48

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

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

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

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

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

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

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