

Identification of a dinucleotide signature that discriminates coding from non-coding long RNAs

Damien Ulveling, Marcel E Dinger, Claire Francastel, Florent Hubé

▶ **To cite this version:**

Damien Ulveling, Marcel E Dinger, Claire Francastel, Florent Hubé. Identification of a dinucleotide signature that discriminates coding from non-coding long RNAs. Frontiers in Genetics, 2014, 5, 10.3389/fgene.2014.00316 . hal-02127278

HAL Id: hal-02127278

https://u-paris.hal.science/hal-02127278

Submitted on 13 May 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

frontiers in GENETICS

HYPOTHESIS AND THEORY ARTICLE
published: 09 September 2014
doi: 10.3389/fgene.2014.00316

Identification of a dinucleotide signature that discriminates coding from non-coding long RNAs

Damien Ulveling[1], Marcel E. Dinger[2], Claire Francastel[1] and Florent Hubé[1]*

[1] CNRS UMR7216, Epigenetics and Cell Fate, Université Paris Diderot, Sorbonne Paris Cité, Paris, France
[2] The University of Queensland Diamantina Institute, The University of Queensland, Brisbane, QLD, Australia

Edited by:
Manja Marz, University of Marburg, Germany

Reviewed by:
David Langenberger, ecSeq Bioinformatics, Germany
Pedro Miramontes, Universidad Nacional Autónoma de México, Mexico

***Correspondence:**
Florent Hubé, UMR7216 - Epigénétique et Destin Cellulaire, Université Paris 7 Diderot, Bâtiment Lamarck - 4ème étage, Case Courrier 7042, 35, rue Hélène Brion, 75013 Paris, France
e-mail: florent.hube@univ-paris-diderot.fr

To date, the main criterion by which long ncRNAs (lncRNAs) are discriminated from mRNAs is based on the capacity of the transcripts to encode a protein. However, it becomes important to identify non-ORF-based sequence characteristics that can be used to parse between ncRNAs and mRNAs. In this study, we first established an extremely selective workflow to define a highly refined database of lncRNAs which was used for comparison with mRNAs. Then using this highly selective collection of lncRNAs, we found the CG dinucleotide frequencies were clearly distinct. In addition, we showed that the bias in CG dinucleotide frequency was conserved in human and mouse genomes. We propose that this sequence feature will serve as a useful classifier in transcript classification pipelines. We also suggest that our refined database of "bona fide" lncRNAs will be valuable for the discovery of other sequence characteristics distinct to lncRNAs.

Keywords: ncRNA, mRNA, CG dinucleotide, sequence biais, pseudogene, intron, exon, database

INTRODUCTION

In the early sixties, the discovery of ribosomal RNA (rRNA) and transfer RNA (tRNA) (Rosset and Monier, 1963; Holley et al., 1965) was the first step toward the identification of different classes of so-called non-protein-coding RNAs (ncRNA or npcRNA). In recent years, ncRNAs have become regarded as key regulatory molecules, and data assigning new functions to these RNAs continue to accumulate exponentially (Mercer et al., 2009; Clark and Mattick, 2011; Mattick, 2011). The primary class, typically referred to as housekeeping or infrastructural ncRNAs (Figure S1), includes tRNA, rRNA, and small nuclear or nucleolar RNA (snRNA and snoRNA) (reviewed in Yoshihisa, 2006; Kawaji and Hayashizaki, 2008). The class of small/short ncRNAs, such as microRNAs (miRNA), short interfering RNAs (siRNA) and piwi-interacting RNAs (piRNA), has also been extensively studied in the last decade, including their biogenesis, function and mechanisms of action, and are now known to be essential regulators of a number of biological processes (Yoshihisa, 2006; Ghildiyal and Zamore, 2009; Li et al., 2010; Farazi et al., 2011). As opposed to these well-documented classes of RNAs, a growing number of longer transcripts are classified into various categories, according to their function, subcellular localization, or genomic proximity with respect to protein-coding genes (e.g., overlapping, antisense, bidirectional). These ncRNAs are often referred to as the dark matter of the genome even though they have been shown to represent the majority of distinct transcripts that arise from mammalian genomes (Mattick, 2001, 2003; Kapranov et al., 2007; Kapranov and St Laurent, 2012). The advent of whole transcriptome sequencing, which has exposed the prevalence of ncRNA transcription, in combination with the profusion of molecular functions operated by these transcripts, has led to an increasing interest and awareness in ncRNAs over the last decade (Chen and Carmichael, 2009; Wilusz et al., 2009). Unifying and discriminating characteristics of ncRNAs remain an important challenge for the further understanding of these biomolecules.

Any attempt to identify or predict new lncRNAs implies that they can be associated with specific features such as structural, thermodynamic, or even sequence and base composition. Whereas small ncRNAs seem to be conserved among species (Quach et al., 2009; Jan et al., 2011), lncRNAs appear to have evolved independently and do not exhibit strong conservation during evolution (Marques and Ponting, 2009). This may explain the few attempts to examine and identify specific features for this class of lncRNA. In addition, available databases for non-protein-coding RNAs typically suffer from certain redundancy and mixtures of various classes of ncRNA.

Here, we first describe the definition of a database of lncRNAs that can be used for examination of sequence-specific features. In addition to extensive literature mining, it is based on the exclusion of hypothetical or predicted sequences, of short RNAs and of sequences that may introduce biases because of redundancy (isoforms, repeats, pseudogenes). Second, we demonstrate the utility of this database by identify a conserved sequence signature of ncRNAs, the CG dinucleotide enrichment, that can be used to effectively discriminate lncRNA from mRNAs.

DEFINING A REFERENCE DATABASE OF lncRNAs

AVAILABLE DATABASES

There are a number of comprehensive ncRNA databases, which cover various classes of ncRNAs, include housekeeping RNAs,

such as tRNAs, snoRNAs and rRNAs (Sprinzl et al., 1998; Wuyts et al., 2004; Griffiths-Jones et al., 2005), small RNAs, such as miRNAs and piRNAs (Griffiths-Jones, 2004), and lncRNAs. However, each of these databases have limitations in their applicability to sequence analysis. The Rfam database contains thousands of mammalian RNA, the majority of which are infrastructural RNAs, predicted using co-variance models from multiple-sequence alignments of genomic datasets, with little direct experimental support for their transcription (Griffiths-Jones et al., 2003). The literature-curated subset of RNAdb comprises approximately two-thirds of miRNAs and snoRNAs (Pang et al., 2005). Likewise, the HGNC (HUGO Gene Nomenclature Classification) database, which contains only human entries (Seal et al., 2011), discriminates non-protein-coding gene loci from infrastructural RNA genes, pseudogenes or antisense sequences of coding genes, is contaminated by genes hosting snoRNA or clusters of miRNA. A certain redundancy is also caused by the presence of non-coding isoforms of mRNA (Hube et al., 2006, 2011; Dinger et al., 2008, 2011; Ulveling et al., 2011a,b) and non-coding transcripts overlapping or antisense of coding transcripts.

WORKFLOW TO RETAIN ONLY BONA FIDE lncRNA

Therefore, we sought to develop a highly filtered set of lncRNAs that was amenable to sequence analysis. We used the lncRNAdb (Amaral et al., 2011) and the HGNC database using the "gene with no protein product locus type" track (Bruford et al., 2008) as main sources of entries included in this collection. We decided to generate this specific collection through a pipeline to keep "bona fide" and accurate lncRNAs. The pipeline consisted of three steps, that (1) excluded "contaminant" RNA, (2) prequalified sequences, and (3) confirmed the candidate (Figure S2).

(1) The first criterion to define "bona fide" regulatory lncRNA candidates was to eliminate known infrastructural RNA and small RNAs, as well as pseudogenes and RNA antisense to annotated protein-coding genes. Indeed, we reasoned that these latter types of transcripts, inherently redundant in sequence and base composition with their cognate protein-coding RNA, would introduce biases in the search for specific features to distinguish lncRNA from other types of RNA.
(2) The second step pre-qualified selected RNA by collecting only human entries, with a validated RefSeq status ("inferred," "model," "predicted," "provisional" and "wgs" are not curated and were therefore excluded, whereas the "validated" status indicated that the RefSeq record was reviewed and subsequently included) and a clearly identified NR_ access number (corresponding to a mix of non-coding transcripts including structural RNAs, transcribed pseudogenes, and others).
(3) The pre-collection was then re-checked by genome mapping using UCSC and GenBank database (NCBI) (Benson et al., 2011) and manually curated based upon extensive literature analysis to validate the uniqueness of retained sequences, the absence of associated protein, and any associated function as functional ncRNA. Additional annotation information (available in the Table S1) was derived from the GenBank database and from the literature (Name/Alias, RefSeq number, chromosome, exon count, length of transcripts, ORF max and Link to disease).

With the concern not to introduce biases of representativeness, we decided to qualify only the longest isoform, if any. Ultimately, our collection (Table S1) contains 52 unique confidently characterized human entries. This dataset contains RNAs that present a median and mean length of 1.5 and 3.1 kb, respectively.

SPECIFIC FEATURES ASSOCIATED WITH "BONA FIDE" ncRNA

To identify specific features that may increase the prediction accuracy of yet unknown lncRNAs, we performed an analysis of sequence composition (dinucleotides index) using the above-defined reference lncRNA collection compared to randomly sampled collections of 50 mRNAs or 50 pseudogene RNA sequences.

DATASETS COMPOSITION AND DATA CROSS-VALIDATION

Human mRNA and pseudogene transcripts were selected randomly from the HGNC database and the corresponding nucleotide sequences were extracted from GenBank. We grouped sequences in datasets of 50 sequences that were cross-validated between each other. Briefly, dinucleotide relative abundance (DRA) was calculated as defined below for each dataset and compared using the Student's *t*-test. Five mRNA and five pseudogene RNA datasets showing no statistical difference were kept for further analyses. Results obtained with mRNA datasets were then compared to those obtained by Bulmer (1987) to further validate our method. Under-representation of dinucleotides CG and TA were observed in both studies (Bulmer, 1987 and the present work, **Figure 1A**), validating the power of the method used here.

To compare data obtained from mRNA, pseudogene and non-coding RNA datasets and to limit bias in the internal composition that may subsist, we have chosen to standardize and correct the DRA for each dinucleotide to that obtained from the entire genome. Human genome chromosomes were obtained from the Genome Reference Consortium Human genome build 37 (GRCh37; ftp://ftp.ncbi.nih.gov/genomes/). The same DRA calculation was performed and validated using data available on the Guide to Human Genome Website (http://www.cshlp.org).

DRA AND TRANSCRIPT SIGNATURE CALCULATION

The transcript signature (TS) is defined as the ratio of its DRA to the DRA of the whole human genome. The DRA is defined as $\rho_{XY} = f^*_{XY}/f^*_X f^*_Y$ where f_X is the frequency of the nucleotide X and f^*_{XY} is the frequency of the dinucleotide XY. The symbol ρ^* measures the abundance of dinucleotides relative to what would be expected from the component base frequencies. Hence, ρ^* (actually $\rho^* - 1$) can also be referred to as the dinucleotide bias.

The relative counts of each nucleotide / dinucleotide are computed within each transcript sequence using the "count" function from the "seqinr" package in the R environment (Charif et al., 2005) (http://cran.r-project.org/web/packages/seqinr/index.html).

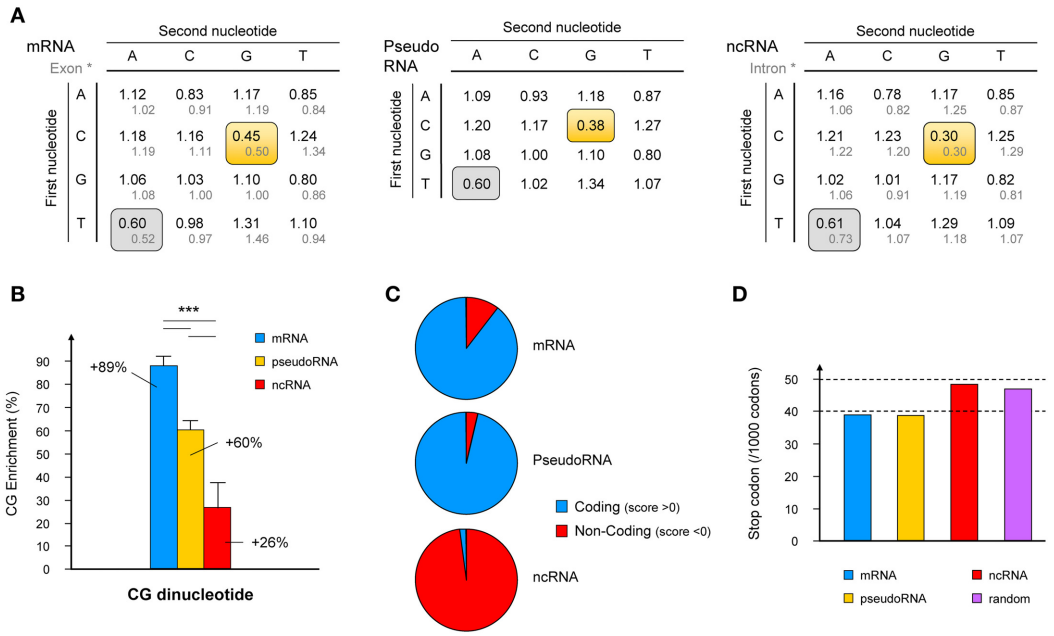


FIGURE 1 | Characterization of specific features of the "bona fide" lncRNA database. (A) Frequencies of occurrence of dinucleotides amongst the "bona fide" lncRNAs compared to that in mRNAs and pseudogenic RNAs (pseudoRNA) and compared to published dinucleotide frequencies in intronic and exonic sequences (Bulmer, 1987) (gray text). Frequencies of underrepresented dinucleotides are framed in gray where no difference is observed, or yellow where differences between mRNA, pseudoRNA and lncRNA are observed. **(B)** The CG dinucleotide signature for mRNAs, pseudoRNAs and lncRNAs is expressed as a% enrichment over the frequency of CG dinucleotide in the whole human genome. Histograms represent mean values ± s.e.m. ***p-value < 0.005 (student's t-test, two-sided). **(C)** Raw data obtained from CPC (Coding Potential Calculator; http://cpc.cbi.pku.edu.cn) using the three databases (mRNA, pseudoRNA and lncRNA) were plotted according to the number of sequences presenting negative (non-coding prediction) or positive (coding capacity) scores. **(D)** Using data extracted from EMBOSS CUSP tool (http://emboss.sourceforge.net), which creates a codon usage table from a nucleotide sequence, the number of stop codons per 1000 bases is represented for the three databases and a set of random sequences generated using the Random DNA Sequence Generator software (http://users-birc.au.dk/biopv/php/fabox).

COMPARISON OF TRANSCRIPT SIGNATURE (TS)

The TS characterizes the enrichment of each dinucleotide normalized to that of the whole human genome. We observed an absence of variation for all dinucleotides except for TA and CG. The dinucleotide TA is broadly under-represented in most prokaryotic sequences and in all eukaryotic genomes (Karlin, 1998). In addition, it is well established that the human genome exhibits extreme CG under-representation owing to the methylation-deamination conversion of CG to TG/CA (Gentles and Karlin, 2001).

Consistent with these data, the dinucleotide TA exhibited an under-representation at the same level for all types of transcripts (about 20%). In contrast, the CG dinucleotide was enriched in all three collections of RNA compared to the entire human genome. More importantly, the CG dinucleotide occurs at almost twice the frequency in mRNA than in the whole genome and this signature clearly distinguishes coding RNA from "bona fide" ncRNA (**Figure 1B** and Figure S3). As mentioned above, transcripts originating from pseudogenes, which retain sequence similarities with the gene from which they derive although lacking coding capacity, exhibit an intermediate CG dinucleotide signature.

CPC DETECTION

We used a recently described computational tool, CPC (Coding Potential Calculator; http://cpc.cbi.pku.edu.cn), which is a Support Vector Machine-based classifier that takes into consideration multiple protein features (peptide length, amino acid composition, protein homologs, secondary structure, and protein alignment) to distinguish mRNAs from ncRNAs (Kong et al., 2007), to compare the coding capacity of the three databases (mRNA, pseudoRNA and "bona fide" ncRNA). For each category of RNA, the number of sequences presenting a negative (non-coding prediction) or positive (coding capacity) score was plotted (**Figure 1C**). We found that over 90% of mRNAs indeed exhibited protein-coding capacity, as well as pseudogenic RNAs. As mentioned above, transcripts originating from pseudogenes exhibit a high sequence homology with the coding genes from which they evolved. It is therefore not surprising that, except for the presence of stop-codons, pseudogenic RNAs exhibit a comparable capacity to encode proteins. In contrast, only one sequence from our "bona fide" lncRNA database was considered as a potentially coding transcript. The WBSCR26 lncRNA contains a putative 240 nt long ORF (80 aa) out of the 471 nt of the transcript (NR_026690, frame 3). As expected, the remaining 98% of ncRNAs were indeed detected as non-coding transcripts, once again validating the strength of the method used to identify "bona fide" ncRNAs.

It should be noted that the identification of "bona fide" lncRNAs as non-coding transcripts could not rely solely on the absence of a sufficiently long ORF to be considered. Indeed, as

shown in Figure S4, ∼50% of the transcripts are predicted to contain an ORF longer than if occurring by chance (Dinger et al., 2008; Ulveling et al., 2011b). Although there was a slight increase (4% vs. 5% for mRNA and ncRNA, respectively) in the number of stop codons in lncRNAs, which indeed is comparable to that found in randomly chosen sequences (**Figure 1D**).

In summary, "bona fide" lncRNA cannot be distinguished from randomly chosen mRNA and pseudogene transcripts in terms coding capacity. However, we uncovered a CG dinucleotide signature that clearly discriminates these "bona fide" lncRNAs from mRNAs.

USE OF CG TRANSCRIPT SIGNATURE TO DISCRIMINATE BETWEEN lncRNAs AND mRNAs

To assess whether the CG transcript signature that we identified was a conserved and consistent feature and could be used to discriminate between non-coding and coding RNAs, we decided to test its power against human and murine reference gene transcript sequence files. The two databases (human.rna.fna and mouse.rna.fna) were downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/) and cleaned of "contaminant" sequences (all RNAs containing "partial," "predicted," "transcript variant" that were > 1, "NR_" RefSeq prefix and "RIKEN" in their title were discarded) to retain only sequences with a "NM_" RefSeq status and used to build a mRNA dataset. The lncRNA dataset was selected on the basis of the "NR_" Refseq prefix. The human collected mRNA and lncRNA datasets contained 18,999 and 6056 non-redundant transcripts, respectively. In parallel, murine datasets were composed of 19,101 and 1116 transcripts corresponding to mRNAs and lncRNAs, respectively.

The distributions of the CG transcript signature for each large dataset are noticeably different, both in humans and mouse (**Figure 2**). Consistent with data obtained with the reduced collection of 52 "bona fide" lncRNAs (**Figure 1B**), the distribution corresponding to lncRNAs is clearly shifted to the left, toward a lower representation of CG dinucleotides (mean 1.96; median 1.81). In clear contrast, the profile corresponding to mRNA is shifted toward a higher representation of CG dinucleotides (mean 2.29; median 2.18).

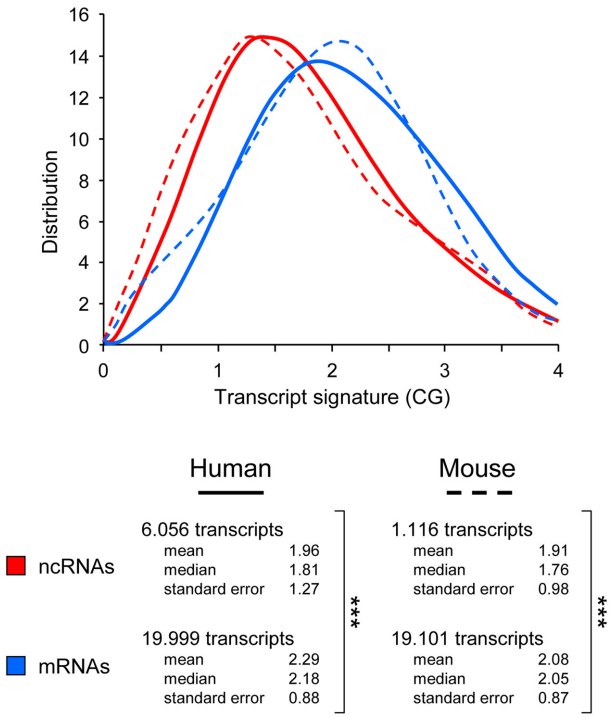


FIGURE 2 | Use of the CG dinucleotide frequency to categorize whole genome transcripts. Distribution of transcript signature scores (CG) obtained from ncRNA, mRNA and all grouped transcripts in human and murine sequences. Human and mouse transcripts in were downloaded from NCBI (human.rna.fna and mouse.rna.fna, respectively) and filtered to specifically select lncRNA and mRNA sequences. Briefly, lncRNAs were selected using NR_ as RefSeq accession number filter, and mRNAs were depleted using "partial," "predicted," "RIKEN," "transcript variant" (with a number >1 to only keep the first one) and "NR_" as keywords. Number of RNA sequences used for the distribution plots, including the mean, median, and standard error for each dataset. ***p-value < 0.005 (student's t-test, two-sided).

CONCLUSION

We demonstrate that the collection of "bona fide" lncRNAs presented here serves as a powerful resource to detect novel unifying features for lncRNAs and distinguish them from other classes of transcripts.

It is interesting to note that lncRNAs exhibit sequence characteristics, at the levels of nucleotides and dinucleotides, similar to that previously described for inherently non-coding sequences like intronic and intergenic regions (Bulmer, 1987; Gentles and Karlin, 2001). Remarkably, the CG dinucleotide composition that we identified discriminates lncRNAs from mRNAs, and, to a lesser extent, lncRNAs from pseudogenic transcripts. Indeed, pseudogenic RNAs, although non-coding, share sequence features with the coding gene from which they originate. Similarly, although still anecdotal, bifunctional RNAs, which can operate both as a functional RNA and an mRNA for the production of a protein, also exhibit intermediate CG dinucleotide signature (not shown). Although this observation is preliminary, it has been revealed after performing a thorough curation of existing datasets to reduce biases introduced by redundancy (e.g., homologs, antisense, isoforms, and pseudogenes) or a mixture of sequences (in terms of length, class, and species).

Even if the number of these "bona fide" lncRNAs is limited, this set will increase as new experimental evidence supports functional roles for unclassified lncRNAs. Meanwhile, we believe that this collection will help uncover additional structural, thermodynamic or sequence features specific for strict non-coding RNAs, and will provide an interesting index classification index for lncRNAs.

To date and to our knowledge the CG dinucleotide represents the first sequence feature that allows discrimination between lncRNA and mRNA that does not depend on coding potential.

ACKNOWLEDGMENTS

This work was supported by the ≪ Association le Cancer du Sein, Parlons-en ! ≫, and by AFM (Association Française contre les Myopathies).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2014.00316/abstract

REFERENCES

Amaral, P. P., Clark, M. B., Gascoigne, D. K., Dinger, M. E., and Mattick, J. S. (2011). lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res.* 39, D146–D151. doi: 10.1093/nar/gkq1138

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2011). GenBank. *Nucleic Acids Res.* 39, D32–D37. doi: 10.1093/nar/gkq1079

Bruford, E. A., Lush, M. J., Wright, M. W., Sneddon, T. P., Povey, S., and Birney, E. (2008). The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res.* 36, D445–D448. doi: 10.1093/nar/gkm881

Bulmer, M. (1987). A statistical analysis of nucleotide sequences of introns and exons in human genes. *Mol. Biol. Evol.* 4, 395–405.

Charif, D., Thioulouse, J., Lobry, J. R., and Perriere, G. (2005). Online synonymous codon usage analyses with the ade4 and seqinR packages. *Bioinformatics* 21, 545–547. doi: 10.1093/bioinformatics/bti037

Chen, L. L., and Carmichael, G. G. (2009). Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA. *Mol. Cell* 35, 467–478. doi: 10.1016/j.molcel.2009.06.027

Clark, M. B., and Mattick, J. S. (2011). Long noncoding RNAs in cell biology. *Semin. Cell Dev. Biol.* 22, 366–376. doi: 10.1016/j.semcdb.2011.01.001

Dinger, M. E., Gascoigne, D. K., and Mattick, J. S. (2011). The evolution of RNAs with multiple functions. *Biochimie* 93, 2013–2018. doi: 10.1016/j.biochi.2011.07.018

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. doi: 10.1371/journal.pcbi.1000176

Farazi, T. A., Spitzer, J. I., Morozov, P., and Tuschl, T. (2011). miRNAs in human cancer. *J. Pathol.* 223, 102–115. doi: 10.1002/path.2806

Gentles, A. J., and Karlin, S. (2001). Genome-scale compositional comparisons in eukaryotes. *Genome Res.* 11, 540–546. doi: 10.1101/gr.163101

Ghildiyal, M., and Zamore, P. D. (2009). Small silencing RNAs: an expanding universe. *Nat. Rev. Genet.* 10, 94–108. doi: 10.1038/nrg2504

Griffiths-Jones, S. (2004). The microRNA Registry. *Nucleic Acids Res.* 32, D109–D111. doi: 10.1093/nar/gkh023

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. R. (2003). Rfam: an RNA family database. *Nucleic Acids Res.* 31, 439–441. doi: 10.1093/nar/gkg006

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., and Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* 33, D121–D124. doi: 10.1093/nar/gki081

Holley, R. W., Pgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., et al. (1965). Structure of a ribonucleic acid. *Science* 147, 1462–1465. doi: 10.1126/science.147.3664.1462

Hube, F., Guo, J., Chooniedass-Kothari, S., Cooper, C., Hamedani, M. K., Dibrov, A. A., et al. (2006). Alternative splicing of the first intron of the steroid receptor RNA activator (SRA) participates in the generation of coding and noncoding RNA isoforms in breast cancer cell lines. *DNA Cell Biol.* 25, 418–428. doi: 10.1089/dna.2006.25.418

Hube, F., Velasco, G., Rollin, J., Furling, D., and Francastel, C. (2011). Steroid receptor RNA activator protein binds to and counteracts SRA RNA-mediated activation of MyoD and muscle differentiation. *Nucleic Acids Res.* 39, 513–525. doi: 10.1093/nar/gkq833

Jan, C. H., Friedman, R. C., Ruby, J. G., and Bartel, D. P. (2011). Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. *Nature* 469, 97–101. doi: 10.1038/nature09616

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., et al. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316, 1484–1488. doi: 10.1126/science.1138341

Kapranov, P., and St Laurent, G. (2012). Dark Matter RNA: existence, function, and controversy. *Front. Genet.* 3:60. doi: 10.3389/fgene.2012.00060

Karlin, S. (1998). Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr. Opin. Microbiol.* 1, 598–610. doi: 10.1016/S1369-5274(98)80095-7

Kawaji, H., and Hayashizaki, Y. (2008). Exploration of small RNAs. *PLoS Genet.* 4:e22. doi: 10.1371/journal.pgen.0040022

Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L., et al. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* 35, W345–W349. doi: 10.1093/nar/gkm391

Li, J. Y., Yong, T. Y., Michael, M. Z., and Gleadle, J. M. (2010). Review: the role of microRNAs in kidney disease. *Nephrology.(Carlton.)* 15, 599–608. doi: 10.1111/j.1440-1797.2010.01363.x

Marques, A. C., and Ponting, C. P. (2009). Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol.* 10:R124. doi: 10.1186/gb-2009-10-11-r124

Mattick, J. S. (2001). Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.* 2, 986–991. doi: 10.1093/embo-reports/kve230

Mattick, J. S. (2003). Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* 25, 930–939. doi: 10.1002/bies.10332

Mattick, J. S. (2011). The central role of RNA in human development and cognition. *FEBS Lett.* 585, 1600–1616. doi: 10.1016/j.febslet.2011.05.001

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* 10, 155–159. doi: 10.1038/nrg2521

Pang, K. C., Stephen, S., Engstrom, P. G., Tajul-Arifin, K., Chen, W., Wahlestedt, C., et al. (2005). RNAdb–a comprehensive mammalian noncoding RNA database. *Nucleic Acids Res.* 33, D125–D130. doi: 10.1093/nar/gki089

Quach, H., Barreiro, L. B., Laval, G., Zidane, N., Patin, E., Kidd, K. K., et al. (2009). Signatures of purifying and local positive selection in human miRNAs. *Am. J. Hum. Genet.* 84, 316–327. doi: 10.1016/j.ajhg.2009.01.022

Rosset, R., and Monier, R. (1963). [Apropos of the presence of weak molecular weight RNA in the ribosomes of *Escherichia Coli*]. *Biochim. Biophys. Acta* 68, 653–656. doi: 10.1016/0926-6550(63)90495-X

Seal, R. L., Gordon, S. M., Lush, M. J., Wright, M. W., and Bruford, E. A. (2011). genenames.org: the HGNC resources in 2011. *Nucleic Acids Res.* 39, D514–D519. doi: 10.1093/nar/gkq892

Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., and Steinberg, S. (1998). Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* 26, 148–153. doi: 10.1093/nar/26.1.148

Ulveling, D., Francastel, C., and Hube, F. (2011a). Identification of potentially new bifunctional RNA based on genome-wide data-mining of alternative splicing events. *Biochimie* 93, 2024–2027. doi: 10.1016/j.biochi.2011.06.019

Ulveling, D., Francastel, C., and Hube, F. (2011b). When one is better than two: RNA with dual functions. *Biochimie* 93, 633–644. doi: 10.1016/j.biochi.2010.11.004

Wilusz, J. E., Sunwoo, H., and Spector, D. L. (2009). Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.* 23, 1494–1504. doi: 10.1101/gad.1800909

Wuyts, J., Perriere, G., and Van, D. E., Peer, Y. (2004). The European ribosomal RNA database. *Nucleic Acids Res.* 32, D101–D103. doi: 10.1093/nar/gkh065

Yoshihisa, T. (2006). tRNA, new aspects in intracellular dynamics. *Cell. Mol. Life Sci.* 63, 1813–1818. doi: 10.1007/s00018-006-6092-9

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 26 June 2014; paper pending published: 07 August 2014; accepted: 22 August 2014; published online: 09 September 2014.

*Citation: Ulveling D, Dinger ME, Francastel C and Hubé F (2014) Identification of a dinucleotide signature that discriminates coding from non-coding long RNAs. Front. Genet. **5**:316. doi: 10.3389/fgene.2014.00316*

This article was submitted to Non-Coding RNA, a section of the journal Frontiers in Genetics.

Copyright © 2014 Ulveling, Dinger, Francastel and Hubé. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.