Identification of potentially new bifunctional RNA based on genome-wide data-mining of alternative splicing events

Damien Ulveling, Claire Francastel, Florent Hubé

Identification of potentially new bifunctional RNA based on genome-wide data-mining of alternative splicing events

Damien Ulveling, Claire Francastel and Florent Hubé*

Université Paris Diderot, CNRS UMR7216, Epigenetics and Cell Fate, Paris, France

* Corresponding author:
Dr. Florent Hubé
UMR7216 - Epigénétique et Destin Cellulaire
Université Paris 7 Diderot
Bâtiment Lamarck - 4ème étage
Case Courrier 7042
35, rue Hélène Brion
75013 PARIS
Tél: +33 1 57 27 89 32
Fax: +33 1 57 27 89 10
E-mail: florent.hube@univ-paris-diderot.fr

Key words:
Intron retention; alternative splicing; non-coding RNA; bifunctional RNA; Steroid Receptor RNA Activator SRA

Highlights:
Bifunctional RNA are RNA that carry both RNA-translatable and RNA-intrinsic functions; Alternative splicing might regulate coding and non-coding capabilities.

Abstract:

It is now evident that the transcriptional output of the genome is much more complex than estimates based on the number of protein-coding genes, and that non-coding RNA widely increase the source of regulatory molecules, a role previously ascribed to proteins. Furthermore, the recent characterization of bifunctional RNA, *i.e.* RNA for which both coding capacity and activity as functional RNA have been reported, adds an additional degree of complexity. Based on the SRA (*Steroid Receptor RNA Activator*) model, where bifunctionality is regulated by alternative splicing, we hypothesized that similar cases, not yet formally tested experimentally, might exist. Using freely available data from high-throughput sequencing projects, we propose here a bioinformatical identification of mRNA whose ORF are disrupted by alternative splicing events, especially by intron retention, and potentially representing a cognate non-coding RNA. Our data mining approach revealed that the human genome contains around 300 possibilities of potentially new bifunctional RNA.

Completion of the Human Genome Project and, later on, advances in whole genome tiling arrays and high-throughput cDNA/EST sequencing have led to at least two unexpected outcome which have advanced our somewhat dichotomous vision of genome expression which placed genes and their messenger RNA products in the limelight for decades, whereas the remainder was relegated as junk DNA. The predicted 50-100,000 protein-coding gene units, that would reflect the great complexity of the human organism, were first revised downwards to about 20,000 coding genes [1]. In parallel, it strengthened the view that genomes are widely transcribed while mRNA represents only 1-2% of this transcriptional output. The vast majority of these RNA was classified as non-coding RNA (ncRNA) [2; 3] primarily owing to the fact that, by definition, no open reading frame (ORF) long enough to be considered, nor protein, have been associated with them. Generally speaking, this classification incorporates transcripts generated by pervasive transcription outside of annotated gene loci for which no function has been identified yet, although many studies are continuing to add to our knowledge of the fundamental role of these RNA in most biological processes (for a review see [4]). Most non-coding RNA, and especially long non-coding RNA, are transcribed by RNA polymerase type II, capped and polyadenylated, undergo splicing, and can not be discriminated from mRNA [5-8], except by their protein-coding potential. This view has been challenged recently, with the concept of bifunctional RNA (for a review, see [9] and Figure 1). In essence, bifunctional RNA hold the dual capacity of serving as both intermediate molecules translated into protein and functional RNA. Several examples have been described recently in the animal kingdom, with p53 being a prime example (nicely reviewed by M. Candeias in this issue). Both p53 mRNA and protein can interact with the Mdm2 protein with opposing effects on p53 synthesis and proteasome-mediated

degradation [10; 11]. Historically, the founding member of this new class of ncRNA exhibiting the ability to encode for proteins was SRA (*Steroid Receptor RNA Activator*), first identified as a structural ncRNA molecule in hormone receptor complexes, characterized by discrete stem-loop structures required for the co-activator function of SRA [12] extensively reviewed by Cooper et al. in this issue. A few years later, new SRA isoforms were identified, exhibiting an additional exon upstream of the core exons, containing two initiating methionines and a predicted open reading frame (ORF) of 236/224 amino acids [13; 14], for which two associated SRAP proteins were detected shortly afterwards [13; 15; 16]. Interestingly, the existence of both coding and non-coding SRA transcripts seems to be regulated, at least in part, by the differential splicing of the first intron of SRA. Therefore, the regulation of alternative splicing on this type of molecule might regulate the balance between RNA and protein-coding functions and influence the overall effect of SRA expression on the regulation of gene expression and cell differentiation [6; 13-19]. In this respect, we recently established that while SRA ncRNA was an enhancer of MyoD transcriptional activity and myogenic differentiation, SRAP prevented this SRA RNA-dependant co-activation through interaction of SRAP with its RNA counterpart.

Such an example of a genetic locus producing both coding and non-coding RNA, depending primarily on an event of alternative splicing, is to our knowledge an isolated case. But it can be predicted that bifunctional RNA could be more widespread than expected and could apply to hundreds of mRNA that could function as functional RNA or, conversely as non-coding RNA that could hide a capacity to encode peptides, simply because these aspects have rarely been formally addressed [20; 21]. We propose here to take advantage of the massive amount and diversity of data generated by sequencing projects to perform a bioinformatical identification of

mRNA with ORF disrupted by alternative splicing, that could potentially represent new bifunctional RNA for which experimental approaches may uncover a functional implication (Figure 2).

The number of introns in a genome seems to increase with organismal complexity since it greatly increased during evolution [20], from 4 introns in the whole genome of *Giardia intestinalis* to about 7.8 per gene in humans. With the number of coding genes estimated between 19,000 and 25,000, the human genome contains about 150,000 to 200,000 introns [1; 20]. A particular case in the general process of alternative splicing of introns leads to maintenance of certain introns. In vertebrates, this retention of introns affects predominantly short introns, *i.e.* shorter than 200 nucleotides in length [22]. Although this only accounts for about 3.1% of all reported alternative splicing events, hundreds of genes may be affected [20]. Whether this retention of introns produces functional and stable molecules, as reported in the case of SRA, remains to be explored.

As shown in Figure 3, the vast majority of the observed retention of introns (1962 out of 2241 reported events) fall within coding sequences (CDS), which only represent about half of the full mRNA length [23], suggesting a non-random distribution of retained introns. Those retained at the extremities of untranslated regions (UTR), 140 events in 5'UTR and 139 and 3'UTR, might contribute to the stabilization (or destabilization) of the associated transcript or might offer target sites for miRNA and therefore participate in the regulation of mRNA degradation. The appearance of stop codon(s) in such regions would not interfere with the protein coding capacity of the transcripts. Indeed, in 620 cases, retention of an intron did not change the reading frame, but could potentially lead to the formation of additional

protein domains of up to 66 amino acids since we selected introns shorter than 200 nucleotides. This type of alternative splicing has been known for decades to participate in the enrichment of the diversity of proteins produced from a single gene and potentially diversification of protein function. However, when introns are retained within the CDS, the appearance of a stop codon would contribute to the production of a truncated peptide (if inserted close to the 3'-end) or the absence of a protein product (if inserted downstream of the start codon). We found here that the translatability of about 318 genes might indeed be interrupted by the retention of an intron in the first third of the CDS. This type of event was typically disregarded owing to the absence of protein products associated with them, but in the light of recent data, they might also contribute to the diversification of the information carried by genes, by producing functional RNA.

During our data-mining for alternative splicing events described above and in the legend of Figure 2, we identified 318 potentially new bifunctional RNA. SRA, the first reported case of retention of an intron that influences the production of two distinct molecular entities with opposing functions in the same pathways, which served as the base for this study, was indeed found in this cluster of 318 transcripts, validating the potential and accuracy of the approach employed. One could argue that these alternatively spliced and untranslated mRNA are usually considered as defective RNA and thought to be directed to nonsense mediated decay (NMD) pathways to be degraded. However, the database we used contained entries exhibiting various types of alternative splicing events and the denoted 'retained intron' was based on an analysis of splicing graphs, conserved between human and mouse, and further confirmed by their presence in EST or cDNA database, indicating that they may well exist under physiological conditions [24]. Of course, these

potentially new bifunctional RNA await experimental validation to test their functionality at the RNA level.

Acknowledgements

F.H. was supported by “Association Française contre les Myopathies” and the “Association le Cancer du Sein, Parlons-en!”. Work in C.F. lab is supported by research founding from grants from LNCC (Ligue Nationale Contre le Cancer).

Figure 1. Influence of alternative splicing events on the transcriptional output of a gene. In addition to the classical enrichment of protein diversity, alternative splicing might also contribute to the diversification of the information carried by genes, by producing functional RNA instead of a protein product. For example, SRA RNA exists as coding and non-coding isoforms, through alternative splicing of the first intron, leading to the production of a protein-coding mRNA whereas its retention leads to the production of a non-protein-coding functional RNA (ncRNA), forming the basis for the concept of “bifunctional RNA”. (ORF, Open Reading Frame).

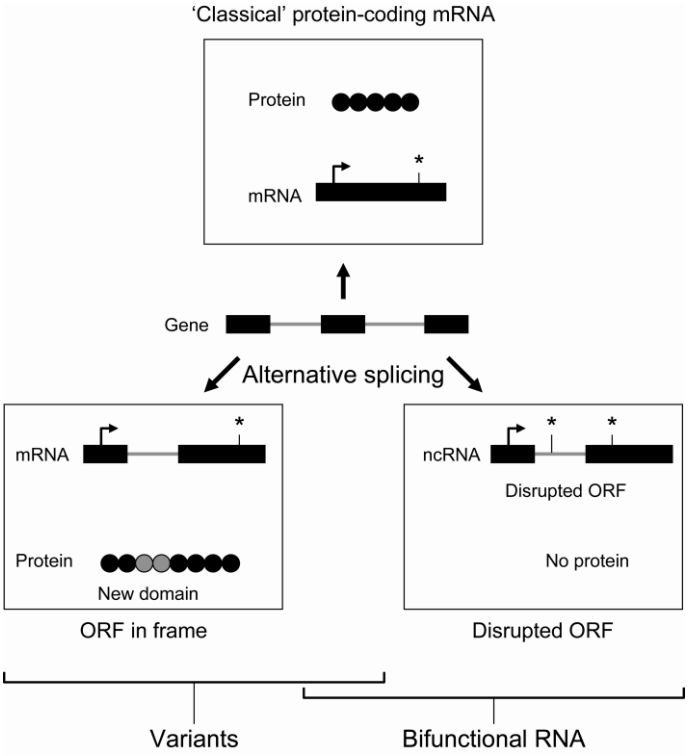

Figure 2. Workflow for data mining of alternative splicing events. Datasets of ‘Human Introns’, ‘Alternative Events’ and ‘Genes’ were retrieved from the UCSC table browser (http://genome.ucsc.edu), using "Genes and gene prediction", “Alt events” or "RefSeq Genes" tracks, respectively, from the hg19 assembly covering the whole human genome. Using Galaxy (http://main.g2.bx.psu.edu) and ‘Operate on Genomic Intervals’ tool, we have selected only introns shorter than 200 nucleotides and 100% covered in 'Alternative Events’ datasets. Location of intron retention relative to the start and stop codons was obtained by genomic coordinate comparison (from ‘Genes’ dataset). ORF disruption was then assessed using a custom Perl script.

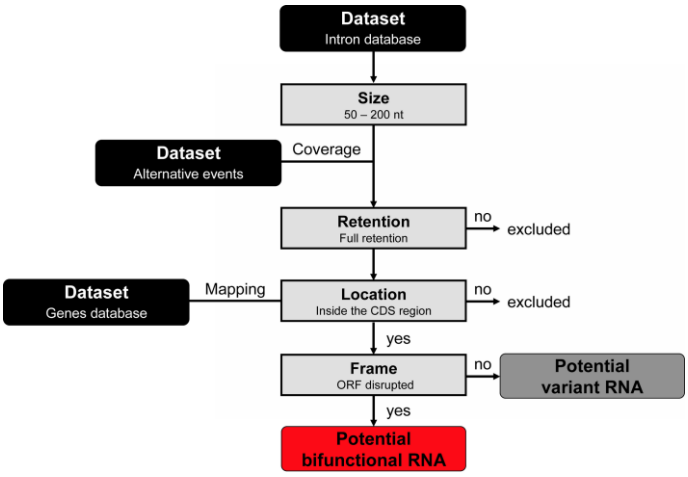

Figure 3. Number of entries identified throughout the workflow described in Figure 2, analyzing retention of introns in mRNA. UTR, CDS and ORF stands for untranslated region, coding sequence region and open reading frame, respectively.

Introns from hg19 build : 326 536
Introns shorter then 200 nt : 38 800
Introns crossing with alternative events : 2241
Included in 5'-UTR : 140
Included in CDS : 1962
Included in 3'-UTR : 139
ORF disrupted by intron retention: 1342
ORF undamaged by intron retention : 620
ORF disrupted in the first third of the protein : 318

References

[1] G. Ast, The alternative genome, Sci.Am. 292 (2005) 40-47.

[2] M. Pheasant, J.S. Mattick, Raising the estimate of functional human sequences, Genome Res. 17 (2007) 1245-1253.

[3] J.S. Mattick, Genome-sequencing anniversary. The genomic foundation is shifting, Science 331 (2011) 874.

[4] T.R. Mercer, M.E. Dinger, J.S. Mattick, Long non-coding RNAs: insights into functions, Nat.Rev.Genet. 10 (2009) 155-159.

[5] P. Carninci, Y. Hayashizaki, Noncoding RNA transcription beyond annotated genes, Curr.Opin.Genet.Dev. 17 (2007) 139-144.

[6] M.E. Dinger, K.C. Pang, T.R. Mercer, J.S. Mattick, Differentiating protein-coding and noncoding RNA: challenges and ambiguities, PLoS.Comput.Biol. 4 (2008) e1000176.

[7] S. Griffiths-Jones, Annotating noncoding RNA genes, Annu.Rev.Genomics Hum.Genet. 8 (2007) 279-298.

[8] D. Ulveling, C. Francastel, F. Hube, When one is better than two: RNA with dual functions, Biochimie (2010).

[9] D. Ulveling, C. Francastel, F. Hube, When one is better than two: RNA with dual functions, Biochimie 93 (2011) 633-644.

[10] M.M. Candeias, L. Malbert-Colas, D.J. Powell, C. Daskalogianni, M.M. Maslon, N. Naski, K. Bourougaa, F. Calvo, R. Fahraeus, P53 mRNA controls p53 activity by managing Mdm2 functions, Nat.Cell Biol. 10 (2008) 1098-1105.

[11] D. Ulveling, C. Francastel, F. Hube, When one is better than two: RNA with dual functions, Biochimie (2010).

[12] R.B. Lanz, B. Razani, A.D. Goldberg, B.W. O'Malley, Distinct RNA motifs are important for coactivation of steroid hormone receptors by steroid receptor RNA activator (SRA), Proc.Natl.Acad.Sci.U.S.A 99 (2002) 16081-16086.

[13] S. Chooniedass-Kothari, E. Emberley, M.K. Hamedani, S. Troup, X. Wang, A. Czosnek, F. Hube, M. Mutawe, P.H. Watson, E. Leygue, The steroid receptor RNA activator is the first functional RNA encoding a protein, FEBS Lett. 566 (2004) 43-47.

[14] E. Emberley, G.J. Huang, M.K. Hamedani, A. Czosnek, D. Ali, A. Grolla, B. Lu, P.H. Watson, L.C. Murphy, E. Leygue, Identification of new human

coding steroid receptor RNA activator isoforms, Biochem.Biophys.Res.Commun. 301 (2003) 509-515.

[15] F. Hube, J. Guo, S. Chooniedass-Kothari, C. Cooper, M.K. Hamedani, A.A. Dibrov, A.A. Blanchard, X. Wang, G. Deng, Y. Myal, E. Leygue, 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 (2006) 418-428.

[16] F. Hube, G. Velasco, J. Rollin, D. Furling, C. Francastel, Steroid receptor RNA activator protein binds to and counteracts SRA RNA-mediated activation of MyoD and muscle differentiation, Nucleic Acids Res. 39 (2011) 513-525.

[17] C. Cooper, J. Guo, Y. Yan, S. Chooniedass-Kothari, F. Hube, M.K. Hamedani, L.C. Murphy, Y. Myal, E. Leygue, Increasing the relative expression of endogenous non-coding Steroid Receptor RNA Activator (SRA) in human breast cancer cells using modified oligonucleotides, Nucleic Acids Res. 37 (2009) 4518-4531.

[18] A. Stark, M.F. Lin, P. Kheradpour, J.S. Pedersen, L. Parts, J.W. Carlson, M.A. Crosby, M.D. Rasmussen, S. Roy, A.N. Deoras, J.G. Ruby, J. Brennecke, E. Hodges, A.S. Hinrichs, A. Caspi, B. Paten, S.W. Park, M.V. Han, M.L. Maeder, B.J. Polansky, B.E. Robson, S. Aerts, H.J. van, B. Hassan, D.G. Gilbert, D.A. Eastman, M. Rice, M. Weir, M.W. Hahn, Y. Park, C.N. Dewey, L. Pachter, W.J. Kent, D. Haussler, E.C. Lai, D.P. Bartel, G.J. Hannon, T.C. Kaufman, M.B. Eisen, A.G. Clark, D. Smith, S.E. Celniker, W.M. Gelbart, M. Kellis, Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures, Nature 450 (2007) 219-232.

[19] C.D. Warden, S.H. Kim, S.V. Yi, Predicted functional RNAs within coding regions constrain evolutionary rates of yeast proteins, PLoS.One. 3 (2008) e1559.

[20] T.E. Koralewski, K.V. Krutovsky, Evolution of exon-intron structure and alternative splicing, PLoS.One. 6 (2011) e18055.

[21] T. Kondo, S. Plaza, J. Zanet, E. Benrabah, P. Valenti, Y. Hashimoto, S. Kobayashi, F. Payre, Y. Kageyama, Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis, Science 329 (2010) 336-339.

[22] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C.

Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R.A. Gibbs, D.M. Muzny, S.E. Scherer, J.B. Bouck, E.J. Sodergren, K.C. Worley, C.M. Rives, J.H. Gorrell, M.L. Metzker, S.L. Naylor, R.S. Kucherlapati, D.L. Nelson, G.M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D.R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H.M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R.W. Davis, N.A. Federspiel, A.P. Abola, M.J. Proctor, R.M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D.R. Cox, M.V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G.A. Evans, M. Athanasiou, R. Schultz, B.A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W.R. McCombie, B.M. de la, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J.A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D.G. Brown, C.B. Burge, L. Cerutti, H.C. Chen, D. Church, M. Clamp, R.R. Copley, T. Doerks, S.R. Eddy, E.E. Eichler, T.S. Furey, J. Galagan, J.G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L.S. Johnson, T.A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W.J. Kent, P. Kitts, E.V. Koonin, I. Korf, D. Kulp, D. Lancet, T.M. Lowe, A. McLysaght, T. Mikkelsen, J.V. Moran, N. Mulder, V.J. Pollara, C.P. Ponting, G. Schuler, J. Schultz, G. Slater, A.F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y.I. Wolf, K.H. Wolfe, S.P. Yang, R.F. Yeh, F. Collins, M.S. Guyer, J. Peterson, A. Felsenfeld, K.A. Wetterstrand, A. Patrinos, M.J. Morgan, J.P. de, J.J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y.J. Chen, Initial sequencing and analysis of the human genome, Nature 409 (2001) 860-921.

[23] P.A. Galante, N.J. Sakabe, N. Kirschbaum-Slager, S.J. de Souza, Detection and evaluation of intron retention events in the human transcriptome, RNA. 10 (2004) 757-765.

[24] C.W. Sugnet, W.J. Kent, M. Ares, Jr., D. Haussler, Transcriptome and genome conservation of alternative splicing events in humans and mice, Pac.Symp.Biocomput. (2004) 66-77.