HIDDEN TREASURES IN UNSPLICED EST DATA

Jan Engelhardt[1] *& Peter F. Stadler*[1−6]

[1]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center of Bioinformatics, Univ. Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany; [2]Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany; [3]RNomics Group, Fraunhofer IZI, Perlickstraße 1, D-04103 Leipzig, Germany; [4]Inst. f. Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Vienna, Austria; [5]Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg, Denmark; [6]The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, New Mexico, USA

jane@bioinf.uni-leipzig.de, studla@bioinf.uni-leipzig.de

Abstract. Several classes of exclusively – or at least predominantly – unspliced non-coding RNAs have been described in the last years, including totally and partially intronic transcripts and long intergenic RNAs. Functionally, they appear to be involved in regulating gene expression, at least in part by associating with the chromatin. Here we systematically analyze the distribution of unspliced ESTs in the human genome. Most appear in clusters overlapping or in the close vicinity of annotated RefSeq genes. Partially Intronic (PIN) unspliced ESTs show complex patterns of overlap with the intron/exon structure of the RefSeq gene. Distinctive patterns of CAGE tags indicate that a large class of unspliced EST clusters forms long extensions of 3'UTRs, at least several hundreds of which probably appear also as independent uaRNAs.

1 INTRODUCTION

Noncoding RNA (ncRNA) constitutes a significant portion of the mammalian transcriptome [1]. These transcripts form by no means a homogeneous class, however. A large sub-class of long ncRNAs (lncRNAs) that includes mRNA-like RNAs [2] differs from their protein-coding siblings only in coding capacity: these transcripts are capped, spliced, and polyadenylated. At least some of them, furthermore, are conserved over long evolutionary time-scales [3]. Nuclear retained ncRNAs, on the other hand, are often spliced transcripts but not polyadenylated. These "*dark matter* RNAs", which have remained largely un-annotated so far, can in fact be the dominating non-ribosomal RNA component in a mammalian cell [4]. In this contribution we focus on the unspliced transcripts and characterize their genomic distribution and their organization relative to the much better characterized spliced transcriptional output of the human genome. A survey of the literature suggests that there are three major groups of unspliced transcripts: intronic transcripts typically associated with protein-coding genes, transcripts associated with long 3'-UTRs, and independent unspliced RNAs found in intergenic regions.

Tens of thousands of totally and partially intronic transcripts (TINs and PINs) have been reported in the human and mouse transcriptomes [5, 6, 7], many of which are unspliced. A large fraction of these comprises unspliced long anti-sense intronic RNAs [8, 9]. Intronic transcripts have been implicated in gene regulation, presumably employing of a variety of different mechanisms [10]. Although a detailed analysis of nearly 40000 putative ncRNAs from RIKEN's FANTOM3 transcript data set showed that a large fraction of the intronic and intergenic transcripts are potentially internally primed from even longer transcripts [11], there are nevertheless thousands of transcripts for which there is no indication that they might be artifacts.

The 3' untranslated regions (3'UTRs) of eukaryotic genes not only regulate mRNA stability, localization and translation. In addition, a large number of 3'UTRs in animals can be decoupled from the protein-coding sequences to which they are normally linked. This independent expression of 3'UTR-derived RNAs (uaRNAs) is regulated and conserved. They appear to function as noncoding RNAs in trans [12]. In the form of independent uaRNAs, they are often detectable as unspliced ESTs.

The best-known examples of independently located unspliced ncRNAs are MALAT-1 and MENβ. These long ($\sim$ 8.7kb and $\sim$ 20kb, resp.) ncRNAs organize nuclear structures known as SC35 speckles and paraspeckles, respectively [13, 14, 15]. Both transcripts are spliced only infrequently [16] and are rather well-conserved [17]. In contrast, the even longer ($\sim$ 100kb) transcripts involved in the regulation of imprinted loci (e.g. Airn [18], its human analog [19], and KCNQ1OT1 [20]) are very poorly conserved. Similar macroRNAs have been observed in tumor cells [4].

Recent evidence demonstrates that non-coding RNAs can affect gene expression both in cis and in trans by modulating the chromatin structure [21]. In fact, reports that RNA is an integral component of chromatin have been published already in the 1970s [22].

Chromatin-associated RNAs (CARs) are predominantly non-polyadenylated [23]. Recently, deep sequencing was used to characterize 141 intronic regions and 74 intergenic regions harboring CARs [24].

Although a large body of unspliced EST data is available in public databases, they have received little attention as a source of information on non-coding RNAs apart from the seminal work Nakaya *et al.* [5]. Here we analyze these data in detail, focusing on their relationships with recently described types on unspliced and rarely spliced transcripts, such as nuclear retained species and independent UTR transcripts.

2 MATERIALS & METHODS

2.1 Data

The annotation track 'Human EST' for genome assembly hg19 was downloaded from the UCSC genome browser (May 2010). Starting from 'ESTs including unspliced' we removed all ESTs contained in the 'Spliced EST' subset in order to obtain the unspliced subset. The remaining data, however, still contains sequences with sometimes long, intron-like, gaps since all transcript which mapped without canonical splice sites have not been included in the 'Spliced EST' subset. We therefore excluded all ESTs with more than 30 deleted nucleotides compared to the reference genome. The cutoff was chosen since even smaller introns are extremely rare [25, 26]. We furthermore discarded ESTs mapped with more than 5% mismatches. We obtained 3,275,035 unspliced ESTs at 408,769 loci, Tab. 1.

A possible source of contamination is nuclear encoded mitochondrial DNA (NUMT) [27]. Since the mitochondrial transcripts remain unspliced at least in mammals [28], it is impossible to reliably distinguish unspliced ESTs mapping to recent NUMTs from fragments of mitochondrial transcripts. An annotation track for human NUMTs was recently provided in [29].

The gene locations and structure were taken from the 'RefSeq Genes' track [30] downloaded from the UCSC genome browser. CAGE data for the hg17 assembly were downloaded from RIKEN [31]. Their coordinates were converted to hg19 using the UCSC liftover tools. Coordinates of chromatin-associated RNAs (CARs) were taken from the supplemental material of [24] and lifted over to hg19.

2.2 Analysis pipeline

While the reading direction of spliced ESTs can be determined with high accuracy from the asymmetric structure of the splice sites, unspliced ESTs data in practice have to be regarded as undirected. The only exception are ESTs arising from polyT-primed libraries, provided the polyA-tail is part of the sequenced fragment. However, there are also internally-primed ESTs containing A-rich regions [32], so that even such a sequence-based detection of the reading direction is not reliable in all cases. Unspliced ESTs are therefore treated as undirected annotation items.

Table 1: Summary of Human EST data.
Analysis of the EST annotation track of hg19 downloaded from the UCSC genome browser in May 2010.

Type	number
all ESTs	8,266,338
spliced ESTs	4,559,208
$> 30nt$ gaps	256,852
$> 5\%$ mismatches	168,024
overlap with NUMTs	7,219
unspliced ESTs	3,275,035

We also have to bear in mind that ESTs are, by definition, not full-length mRNAs. In particular they may cover an unspliced stretch of a spliced ESTs only and thus appear unspliced. We therefore are interested in particular in genomic loci where unspliced ESTs cluster together. To this end we define unspliced ESTs that overlap each other or that are separated by at most 30 nt as members of the cluster. In order to avoid artifacts of both the experimental and the computations procedures we only consider unspliced EST clusters comprising at least three individual ESTs, leaving 101,230 unspliced EST cluster for further analysis.

In order to computer overlaps with existing annotation we used the Table Browser integrated in the UCSC genome browser and the "Operate on Genomic Intervals"-tools of Galaxy Browser [33] as well as R scripts [34] for more complex operations involving multiple tracks.

3 RESULTS

3.1 ESTs "within range" of RefSeq genes

The majority of unspliced EST clusters overlaps, or at least lies in close proximity of, already annotated RefSeq genes. This begs the question whether unspliced ESTs are just a by-product of "normal" spliced transcripts. To this end we compare the abundance of spliced and unspliced ESTs with the range of annotated RefSeq genes (including a 5kb flanking region), Fig. 1. Although there is a weak correlation ($r^2 = 0.47$) within the range of spliced RefSeq genes, we observe a surprisingly variability in the numbers of unspliced ESTs, with relative abundances of spliced and unspliced sequences typically varying by a factor of ten or more. Conversely, there is a surprisingly large number of spliced ESTs overlapping with annotated unspliced RefSeq genes. In these loci, the correlation between spliced and unspliced output is even less pronounced.

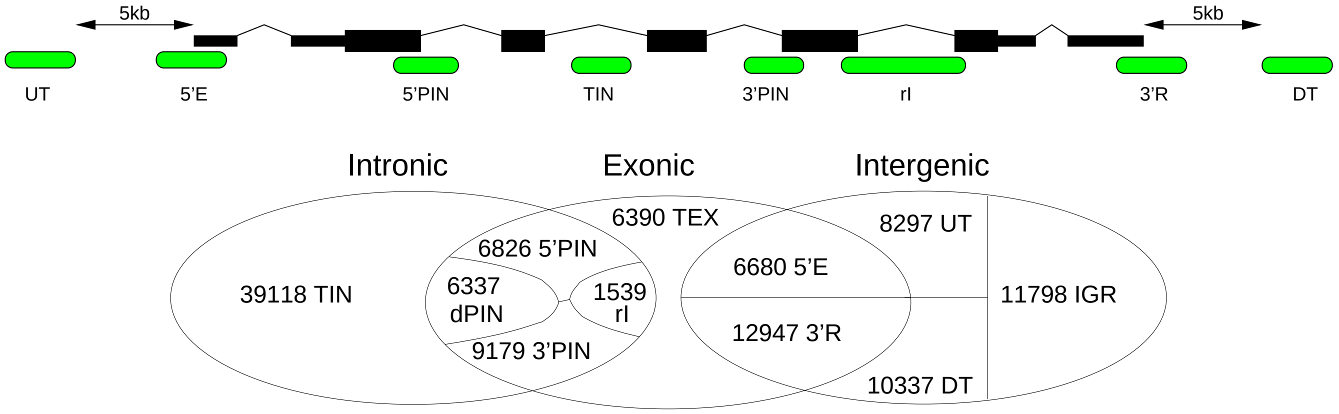
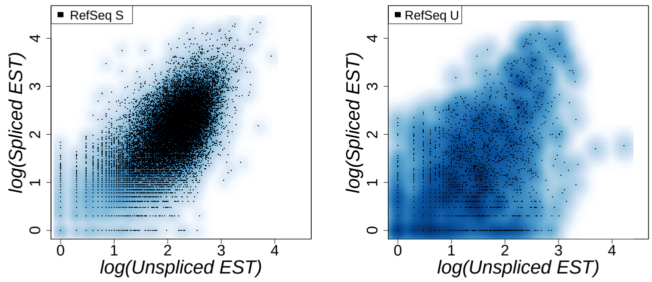


Figure 2: Classification of unspliced EST clusters w.r.t. their location relative to RefSeq genes. With the exception of totally intronic RNAs (TINs) and clusters in the upstream (UT) and downstream (DT) region within 5kb, all other classes partially overlap RefSeq exons: 5' and 3' partially intronic RNAs (5'PIN, 3'PIN), EST clusters overlapping 3' UTR and downstream region (3'R) or 5'UTR and upstream region (5'E), resp., and clusters covering complete introns (rI) are distinguished in the statistical analysis. Furthermore, we record totally exonic clusters (TEX). Below, the data are summarized as a Venn diagram. Some clusters cannot be classified unambiguously, mostly because two or more RefSeq genes may overlap the same locus. For instance, 6,337 "dPINs" are classified as both 5' and 3' PIN. Another 6,164 EST clusters show more complex patterns of overlap with the annotation; these are not included in the diagram.



spliced ($r^2 = 0.47$) unspliced ($r^2 = 0.22$)

Figure 1: Scatter plots of the relative abundance of spliced and unspliced ESTs within the range of both spliced and unspliced RefSeq genes show only a relatively poor correlation. The correlation coefficients are computed from the logarithms of EST counts.

We therefore analyzed in detail the relative location of unspliced ESTs clusters and components of RefSeq genes, Fig. 2. This also connects our analysis to previous work on unspliced intronic transcripts [9, 5, 7]. Compared to [5], we can work with a much larger data set, comprising 39,117 (vs. 5678) *Totally INtronic* (TIN) and 27,151 (vs. 9132) *Partially INtronic* (PIN) unspliced EST clusters. Interestingly, the number of detected TINs has increased by a factor of 6.9, while PINs increased by only a factor of 3, suggesting that the coverage of TINs is much farther from saturation than that of the PINs. The protein-coding part of RefSeq genes overlaps 38,675 unspliced EST clusters. Only 4% of these clusters, however, are located completely in the coding region.

In order to obtain a more fine-grained view of distribution of TINs and PINs, we distinguish PINs depending on whether they overlap the exon/intron boundary at the donor or the acceptor side, Fig. 2. Furthermore we treat coding region and UTRs of coding RefSeq genes separately.

In addition to the unspliced ESTs overlapping the body of the RefSeq genes we also consider EST clusters in the immediate proximity, here defined as within 5k of the annotated ends of the RefSeq entry. There are 8,297 unspliced EST clusters in upstream region of RefSeq genes and 10,377 downstream, see Fig. 2.

3.2 Independent UTR-derived RNAs

Unspliced EST clusters show a bias towards the 3' end of the RefSeq gene. There are more than 23,000 clusters of unspliced ESTs overlapping, or located within 5000 nt downstream of, the 3' UTR of RefSeq genes. An impressive example is shown in Fig. 3. We frequently find that CAGE tags, i.e., markers for transcription start sites, are located within long unspliced 3'UTRs. This supports the observation in [12] that these 3'UTRs are also transcribed in an independent mode.

586,242 of the approximately 1.6 million CAGE tags overlap with a total of 41,136 unspliced EST clusters. Figure 4 shows that CAGE tags are predominantly located at the ends of the unspliced EST clusters. The symmetry of the diagram is expected because we treat unspliced ESTs as undirected. In fact, we can use the CAGE tags to determine the reading direction of those

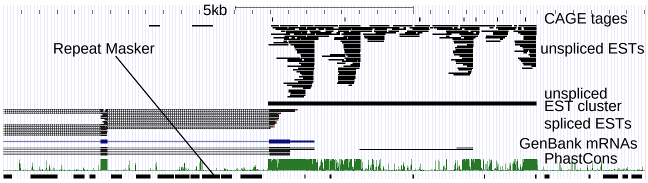


Figure 3: The 3' UTR of the RSF gene is extended by a single cluster of unspliced ESTs covering about 7 kb. Compared to the large number of unspliced ESTs, only a few spliced ESTs cover the inner exons of RSF. The presence of several CAGE tags in the extended UTR region suggests that independent uaRNAs are produced from this locus [hg19 Chr.11: 77386000-77358000]. This is further supported by two unspliced GenBank mRNAs that map to the extended UTR. Is it interesting to note that a large fraction of the extended UTR is very well conserved and nearly devoid of repetitive sequence.

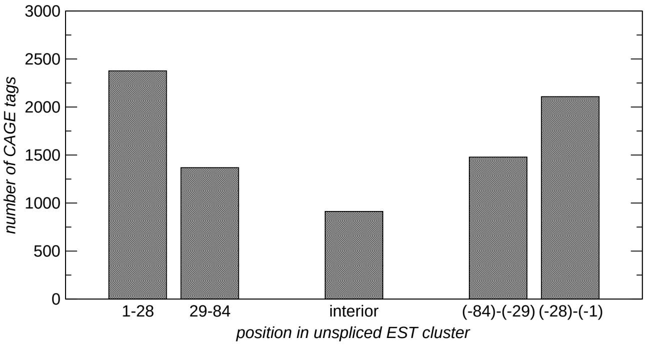


Figure 4: Distribution of CAGE tags on unspliced EST clusters (with length ≥ 170). The number of tags is divided through the length of the region.

clusters in which the tag distribution unambiguously concentrates on one end, which is the 5' end. Here, we require that the density of CAGE tags in the first 84 nt is at least threefold higher than in the remainder of the cluster. We furthermore required at least 10 CAGE tags at the 5' end. We found at total of 3350 such clusters with a CAGE-supported transcription start and an unambiguous reading direction. A subset of 483 overlap at least 100 CAGE tags each.

Not surprisingly, many clusters of unspliced ESTs correspond to the 5' ends or 5' extensions of annotated RefSeq genes, Tab. 2. As expected, most of these clusters share the reading direction with the RefSeq gene. This is also the case of PINs. For the relatively small number of TINs with a strong CAGE signal we find a 2:1 ratio of sense and antisense orientation, consistent with previous observations [5].

Table 2: Unspliced EST clusters whose reading direction is determined by CAGE tags and their orientation relative to the surrounding RefSeq gene. Only clusters overlapping with at least 10 individual CAGE tags are considered. RefSeq TSS are all EST clusters that overlap the 5'UTR or that are located completely within the upstream region. TINs and PINs are defined in Fig. 2. All unspliced EST clusters that have an overlap with the 3' UTR but not with the 5' UTR are interpreted as uaRNAs. "Sense" and "antisense" are relative to the reading direction of the RefSeq gene.

Type	sense	antisense	all
RefSeq TSS	2368	199	14,864
TINs	63	34	36,145
PINs	1668	99	17,546
uaRNAs	25	11	17,985

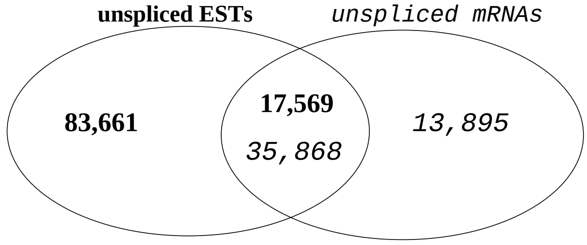


Figure 5: Overlap of unspliced EST clusters with annotated unspliced mRNAs. At least 5% overlap is required.

A particular class of most unspliced transcripts are the 215 chromatin associated RNAs (CARs) [24]. Of these, 145 (67%) overlap unspliced EST clusters, and 179 are located within the range of RefSeq genes ±5 kb of flanking region. We therefore evaluated in more detail how the CARs are distributed relative to the organization of their RefSeq gene: 43% are located in the 3' area (7 entirely in the 3'UTR, 16 partially overlapping the 3'UTR, and 40 in the 3' flanking region), which only 16% are associated with the 5'UTR or 5' flanking region. 19% of the CARs overlap the coding region of RefSeq gene.

3.3 Unspliced mRNAs

35,374 of 49,764 annotated unspliced mRNAs are located within the extended RefSeq regions, of which 29,015 overlap RefSeq exons (including annotated unspliced RefSeq genes). This unspliced mRNA set contains many (partially) overlapping entries, however, so that the number of mRNAs outside the RefSeq regions exceeds the number of unspliced EST clusters. We find that 72% of the unspliced mRNAs overlap with about 17% of the unspliced ESTs, Fig. 5. Among these unspliced mRNAs are several famous transcripts, Tab. 3.

MALAT-1, for instance, appears among the loci with the highest coverage of unspliced ESTs. Together with MENβ, which is located in the same genomic region, it belongs to a class of long nuclear retained transcripts

Table 3: Coverage of some well-known long ncRNAs by unspliced ESTs. The first group is annotated as predominantly unspliced. The second group has annotated spliced isoforms.

Gene	Chr.	approx.loc.	# u-ESTs
MALAT1	11	65.265	16,829
MENβ	11	65.190	2,816
KCNQ1OT1	11	2.661	95
PTCSC	8	134.067	6
XIST	X	73.040	1,338
TUG1	22	31.365	914
HOTAIR	12	54.356	12

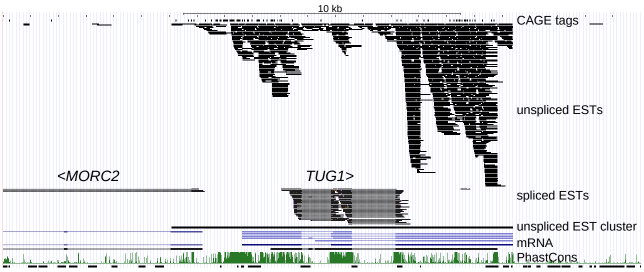


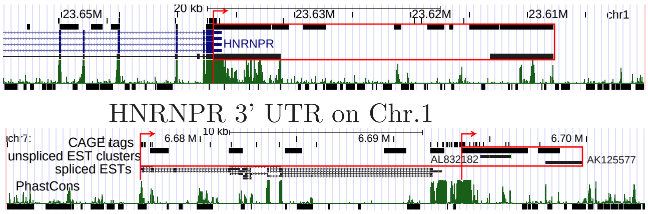
Figure 6: TUG1 (hg19 Chr.22: 31365634-31375380) has been reported as a spliced transcript. The available EST data (927 unspliced *versus* 60 spliced ESTs), however suggest that there are also unspliced isoforms from this locus. A cluster of CAGE tags marks the (divergent) start sites of TUG1 and MORC2.

involved in the organization of nuclear speckles [16, 35], see also [17] and the references therein. Other examples, such as KCNQ1OT1 [20] and PTCSC [36] are clearly visible in our data, albeit with moderate coverage.

Unspliced ESTs are also reported as parts of spliced transcripts, in particular those with very long exons such as XIST [37]. In other cases, such as TUG1 [38] we observe predominantly unspliced ESTs that cover nearly the entire primary transcript, even though the genomic location is annotated by the spliced forms.

9302 EST clusters are located outside spliced RefSeq genes and their 5kB area and do not intersect any annotated human mRNA. 446 of these are within a 5kB range of known unspliced RefSeq genes, leaving 8856 candidates for novel, predominantly unspliced genes. Another 1418 of these "intergenic" clusters overlap already annotated unspliced mRNAs.

A manual inspection showed, however, that many of them are conspicuously well-conserved and can be identified as recent retropseudogenes [39] of known spliced human genes. This concerns in particular the EST clusters with the largest numbers of ESTs. We therefore compared the sequences of clusters against all human proteins (Ensembl release 61) with `blastx` to determine the fraction of such clusters that correspond to retropseudogenes. 2728 loci (31%) derive from protein-coding genes. These unspliced EST clusters might be mapping artifacts and cannot be reliably distinguished from revived retrogenes [40].



HNRNPR 3' UTR on Chr.1

ncRNA on Chr.7

Figure 7: Examples of "intergenic" unspliced EST clusters that probably are part of very long extensions of the 3'UTR. Clusters of CAGE tags indicate transcription start sites (red arrows). Potential uaRNAs are marked by red boxes.

The largest class among the remaining loci are probably long extensions of 3'UTRs of both coding and non-coding spliced transcripts, We suspect that in many of these case we might in fact (also) see uaRNAs. In the examples shown in Fig. 7 this hypothesis is supported by large CAGE tag clusters at the 5' end of the 3' UTRs.

In order to further investigate the 5853 intergenic unspliced EST clusters that were not recognized as likely retrogenes, we used `RNAz 2.0` [41, 42] as provided by the Vienna RNAz server [43] to detect signatures of stabilizing selection on RNA secondary structure and `RNAcode` [44] to find evidence for conserved protein-coding coding regions. This data set covers 3,887,395 nt and yields 1160 RNAz hits with a classification probability $P_{RNAz} > 0.5$ and 443 with $P_{RNAz} > 0.9$. Compared to the predicted 6880 low confidence and 2259 high confidence hits in the 30 Mb of ENCODE regions [42], this amounts to a very moderate enrichment by a factor of 1.3 to 1.5. Although it remains unclear whether this enrichment is confounded by the restriction of unspliced EST clusters to genomic regions with relatively high expression. It could hint a function of long unspliced regions in specific binding with proteins. An example for a conserved secondary structure element is shown in Fig. 8. In contrast to `RNAz`, we found only a small number of short hits with `RNAcode`, which at least mostly appear to be artifacts caused by short segments of coding region in retrogenes.

4 CONCLUSIONS

Unspliced EST data, although typically disregarded in transcriptome analysis, can provide interesting, and at least in part unexpected, insights into the structure of human transcriptome. They outline a major component of transcriptional output that totally or at least partially

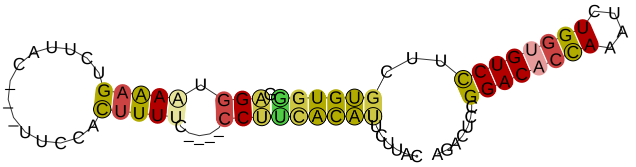


Figure 8: Example of conserved secondary structure element detected by `RNAz`. The motif located at hg19.chr14 107217995(-) is conserved among diverse mammals. The color code, from red to ochre and green indicates that 1, 2, or 3 different types of basepairs are observed, unsaturated colors indicate basepairs that cannot be formed by 1 or 2 of the 6 sequences in the alignment. Substitutions in stem regions are indicated by circles.

escapes splicing. Nevertheless, this valuable resource has not been mined comprehensively in the past. So far, only the partially and totally intronic transcripts (PINs and TINs), which constitute more than half of the clusters of unspliced ESTs, have received attention in systematic studies [5, 6, 7]. Our analysis confirmed the conclusions on these abundant class of transcripts. In addition, we find direct evidence of the independent transcription of PINs and TINs based on CAGE tag data.

The second-largest class of unspliced EST clusters forms extensions of the UTRs of well-annotated genes. On the 5'-side they provide additional information of the transcriptional start sites (TSS) of the annotated RefSeq genes themselves. Not surprisingly, the majority of clusters with clear support for a TSS from CAGE data falls into this class. More interestingly, however, the relative majority of UTR extensions is located at the 3'-end of RefSeq genes and forms typically extensive, several kb-long extensions. In line with the finding of ref. [12], a lot of these 3'UTR extensions contain a TSS for an independently transcribed uaRNA. We identify at least 25 likely uaRNAs with strong support from the CAGE data. Although TINs and PINs have been reported to be transcribed mostly independently of their surrounding RefSeq gene, only a small fraction of them has a sufficient concentration of CAGE tags for a recognizable TSS. This suggest that hundreds or even thousands of the 3'UTR extensions also give rise to independent uaRNAs.

The remaining set of unspliced EST clusters outside a 5kb range around RefSeq genes comprises 12% of the data. About a third of these clusters overlaps retrogenes and retropseudogenes. For these cases, it is often impossible to distinguish between between truely expressed loci and mapping artifacts of ESTs arising from the spliced original of the gene. The manual inspection of a random sample of the remaining cases shows that a large fraction of these ESTs clusters constitute even larger extensions of 3'UTRs. In many cases they also appear to be long 3'UTRs/uaRNAs belonging to previously unannotated, mostly non-coding, spliced transcripts.

In summary, the analysis of unspliced ESTs uncovers a largely unexplored realm of long transcripts. The frequently postulated connection between lack of splicing and nuclear retention, see e.g. [45], and the surprising overlap of chromatin associated transcripts suggests that this class of transcripts might be involved in chromatin organization and possibly other mechanisms of epigenetic control.

Acknowledgments

We thank Christian Otto for assistance with R scripting and Sonja J. Prohaska for helpful discussions. This work was funded in part by the DFG under the auspices of SPP-1174 "Deep Metazoan Phylogeny" and SPP-1258 "Sensory and regulatory RNAs in prokaryotes".

References

[1] The ENCODE Project Consortium. "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." *Nature*, vol. 447, pp. 799–816, 2007.

[2] M. Guttman, M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, J. L. Rinn, E. S. Lander, and A. Regev. "Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs." *Nat Biotechnol.*, vol. 28, pp. 503–510, 2010.

[3] M. Hiller, S. Findeiß, S. Lein, M. Marz, C. Nickel, D. Rose, C. Schulz, R. Backofen, S. J. Prohaska, G. Reuter, and P. F. Stadler. "Conserved introns reveal novel transcripts in *Drosophila melanogaster*." *Genome Res.*, vol. 19, pp. 1289–1300, 2009.

[4] P. Kapranov, G. St Laurent, T. Raz, F. Ozsolak, C. P. Reynolds, P. H. Sorensen, G. Reaman, P. Milos, R. J. Arceci, J. F. Thompson, and T. J. Triche. "The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA." *BMC Biol.*, vol. 8, p. 149, 2010.

[5] H. I. Nakaya, P. P. Amaral, R. Louro, A. Lopes, A. A. Fachel, Y. B. Moreira, T. A. El-Jundi, A. M. da Silva, E. M. Reis, and S. Verjovski-Almeida. "Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription." *Genome Biol.*, vol. 8, p. R43, 2007.

[6] R. Louro, H. I. Nakaya, P. P. Amaral, F. Festa, M. C. Sogayar, A. M. da Silva, S. Verjovski-Almeida, and E. M. Reis. "Androgen responsive intronic non-coding RNAs." *BMC Biol.*, vol. 5, p. 4, 2007.

[7] R. Louro, T. El-Jundi, H. I. Nakaya, E. M. Reis, and S. Verjovski-Almeida. "Conserved tissue expression signatures of intronic noncoding RNAs transcribed from human and mouse loci." *Genomics*, vol. 92, pp. 18–25, 2008.

[8] J. L. Rinn, G. Euskirchen, P. Bertone, R. Martone, N. M. Luscombe, S. Hartman, P. M. Harrison, F. K. Nelson, P. Miller, M. Gerstein, S. Weissman, and M. Snyder. "The transcriptional activity of human chromosome 22." *Genes Dev.*, vol. 17, pp. 529–540, 2003.

[9] E. M. Reis, H. I. Nakaya, R. Louro, F. C. Canavez, A. V. Flatschart, G. T. Almeida, C. M. Egidio, A. C. Paquola, A. A. Machado, F. Festa, D. Yamamoto, R. Alvarenga, C. C. da Silva, G. C. Brito, S. D. Simon, C. A. Moreira-Filho, K. R. Leite, L. H. Camara-Lopes, F. S. Campos, E. Gimba, G. M. Vignal, H. El-Dorry, M. C. Sogayar, M. A. Barcinski, A. M. da Silva, and S. Verjovski-Almeida. "Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer." *Oncogene*, vol. 23, pp. 6684–6692, 2004.

[10] R. Louro, A. S. Smirnova, and S. Verjovski-Almeida. "Long intronic noncoding RNA transcription: expression noise or expression choice?" *Genomics*, vol. 93, pp. 291–298, 2009.

[11] K. J. Nordström, M. A. Mirza, M. S. Almén, D. E. Gloriam, R. Fredriksson, and H. B. Schiöth. "Critical evaluation of the FANTOM3 non-coding RNA transcripts." *Genomics*, vol. 94, pp. 169–176, 2009.

[12] T. R. Mercer, D. Wilhelm, M. E. Dinger, G. Soldà, D. J. Korbie, E. A. Glazov, V. Truong, M. Schwenke, C. Simons, K. I. Matthaei, R. Saint, P. Koopman, and J. S. Mattick. "Expression of distinct RNAs from 3' untranslated regions." *Nucl. Acids Res.*, 2010. Doi: 10.1093/nar/gkq1158.

[13] Y. T. F. Sasaki, T. Ideue, M. Sano, T. Mituyama, and T. Hirose. "MENϵ/β noncoding RNAs are essential for structural integrity of nuclear paraspeckles." *Proc Natl Acad Sci USA*, vol. 106, pp. 2525–2530, 2009.

[14] H. Sunwoo, M. E. Dinger, J. E. Wilusz, P. P. Amaral, J. S. Mattick, and D. L. Spector. "MEN ϵ/β nuclear-retained non-coding RNAs are upregulated upon muscle differentiation and are essential components of paraspeckles." *Genome Res.*, vol. 19, pp. 347–359, 2009.

[15] Y. S. Mao, H. Sunwoo, B. Zhang, and D. L. Spector. "Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs." *Nat. Cell Biol.*, vol. 13, pp. 95–101, 2011.

[16] J. Hutchinson, A. W. Ensminger, C. M. Clemson, C. R. Lynch, J. B. Lawrence, and A. Chess. "A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains." *BMC Genomics*, vol. 8, p. 39, 2007.

[17] P. F. Stadler. "Evolution of the long non-coding RNAs MALAT1 and MENβ/ϵ." In C. E. Ferreira, S. Miyano, and P. F. Stadler, eds., "Advances in Bioinformatics and Computational Biology, 5th Brazilian Symposium on Bioinformatics," vol. 6268 of *Lecture Notes in Computer Science*, pp. 1–12. Heidelberg: Springer Verlag, 2010.

[18] C. I. M. Seidl, S. H. Stricker, and D. P. Barlow. "The imprinted Air ncRNA is an atypical RNAPII transcript that evades splicing and escapes nuclear export." *EMBO J*, vol. 25, pp. 1–11, 2006.

[19] I. Y. Yotova, I. M. Vlatkovic, F. M. Pauler, K. E. Warczok, P. F. Ambros, M. Oshimura, H. C. Theussl, M. Gessler, E. F. Wagner, and D. P. Barlow. "Identification of the human homolog of the imprinted mouse Air non-coding RNA." *Genomics*, vol. 92, pp. 464–473, 2008.

[20] L. Redrup, M. R. Branco, E. R. Perdeaux, C. Krueger, A. Lewis, F. Santos, T. Nagano, B. S. Cobb, P. Fraser, and W. Reik. "The long noncoding RNA Kcnq1ot1 organises a lineage-specific nuclear domain for epigenetic gene silencing." *Development*, vol. 136, pp. 525–530, 2009.

[21] J. Whitehead, G. K. Pandey, and C. Kanduri. "Regulation of the mammalian epigenome by long noncoding RNAs." *Biochim Biophys Acta*, vol. 1790, pp. 936–947, 2009.

[22] I. J. Paul and J. D. Duerksen. "Chromatin-associated RNA content of heterochromatin and euchromatin." *Mol. Cell. Biochem.*, vol. 9, pp. 9–16, 1975.

[23] A. Rodríguez-Campos and F. Azorín. "RNA is an integral component of chromatin that contributes to its structural organization." *PLoS ONE*, vol. 2, p. e1182, 2007.

[24] T. Mondal, M. Rasmussen, G. K. Pandey, A. Isaksson, and C. Kanduri. "Characterization of the RNA content of chromatin." *Genome Res.*, vol. 20, pp. 899–907, 2010.

[25] J. D. Hawkins. "A survey on intron and exon lengths." *Nucleic Acids Res.*, vol. 16, pp. 9893–9908, 1988.

[26] X. Hong, D. G. Scofield, and M. Lynch. "Intron size, abundance, and distribution within untranslated regions of genes." *Mol Biol Evol*, vol. 23, pp. 2392–2404, 2006.

[27] E. Hazkani-Covo, R. Zeller, and W. Martin. "Molecular poltergeists: Mitochondrial DNA copies (numts) in sequenced nuclear genomes." *PLoS Genet.*, vol. 6, p. e1000834, 2010.

[28] J. Asin-Cayuela and C. M. Gustafsson. "Mitochondrial transcription and its regulation in mammalian cells." *Trends Biochem Sci.*, vol. 32, pp. 111–117, 2007.

[29] J. Tsuji. *Immigrants to the Nucleus; Analysis of Mitochondrially Derived Nuclear Genomic Regions (NUMT).* Master's thesis, University of Tokyo, Department of Computational Biology, Graduate School of Frontier Science, 2010. K-02364.

[30] K. D. Pruitt, T. Tatusova, and D. R. Maglott. "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." *Nucleic Acids Res.*, vol. 35, pp. D61–D65, 2007.

[31] H. Kawaji, T. Kasukawa, S. Fukuda, S. Katayama, C. Kai, J. Kawai, P. Carninci, and Y. Hayashizaki. "CAGE basic/analysis databases: the CAGE resource for comprehensive promoter analysis." *Nucleic Acids Res*, vol. 34, pp. D632–D636, 2006.

[32] M. Furuno, K. C. Pang, N. Ninomiya, S. Fukuda, M. C. Frith, C. Bult, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, J. S. Mattick, and H. Suzuki. "Clusters of internally primed transcripts reveal novel long noncoding RNAs." *PLoS Genetics*, vol. 2, p. e37, 2006.

[33] J. Goecks, A. Nekrutenko, J. Taylor, and Galaxy Team. "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." *Genome Biol.*, vol. 11, p. R86, 2010.

[34] J. Gagneur and R. Bourgon. *genomeIntervals: Operations on genomic intervals*, 2009. R package version 1.4.0.

[35] R. Lin, M. Roychowdhury-Saha, C. Black, A. T. Watt, E. G. Marcusson, S. M. Freier, and T. S. Edgington. "Control of RNA processing by a large noncoding RNA over-expressed in carcinomas." *FEBS Lett.*, vol. 585, pp. 671–671, 2011.

[36] H. He, R. Nagy, S. Liyanarachchi, H. Jiao, W. Li, S. Suster, J. Kere, and A. de la Chapelle. "A susceptibility locus for papillary thyroid carcinoma on chromosome 8q24." *Cancer Res.*, vol. 69, pp. 625–631, 2009.

[37] E. A. Elisaphenko, N. N. Kolesnikov, A. I. Shevchenko, I. B. Rogozin, T. B. Nesterova, N. Brockdorff, and S. M. Zakian. "A dual origin of the *Xist* gene from a protein-coding gene and a set of transposable elements." *PLoS One*, vol. 3, p. e2521, 2008.

[38] T. L. Young, T. Matsuda, and C. L. Cepko. "The noncoding RNA taurine upregulated gene 1 is required for differentiation of the murine retina." *Current Biology*, vol. 15, pp. 501–512, 2005.

[39] E. J. Devor and K. A. Moffat-Wilson. "Molecular and temporal characteristics of human retropseudogenes." *Hum. Biol.*, vol. 75, pp. 661–672, 2003.

[40] Z. Yu, M. Morais, D Ivanga, and P. M. Harrison. "Analysis of the role of retrotransposition in gene evolution in vertebrates." *BMC Bioinformatics*, vol. 8, p. 308, 2007.

[41] S. Washietl, I. L. Hofacker, and P. F. Stadler. "Fast and reliable prediction of noncoding RNAs." *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 2454–2459, 2005.

[42] A. R. Gruber, S. Findeiß, S. Washietl, I. L. Hofacker, and P. F. Stadler. "`RNAz 2.0`: improved noncoding RNA detection." *Pac. Symp. Biocomput.*, vol. 15, pp. 69–79, 2010.

[43] A. R. Gruber, R. Neuböck, I. L. Hofacker, and S. Washietl. "The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures." *Nucleic Acids Res.*, vol. 35, pp. W335–W338, 2007.

[44] S. Washietl, S. Findeiß, S. Müller, S. Kalkhof, M. von Bergen, I. L. Hofacker, P. F. Stadler, and N. Goldman. "RNAcode: robust prediction of protein coding regions in comparative genomics data." *RNA*, vol. 17, pp. 578–594, 2011.

[45] A. B. Eberle, V. Hessle, R. Helbig, W. Dantoft, N. Gimber, and N. Visa. "Splice-site mutations cause rrp6-mediated nuclear retention of the unspliced RNAs and transcriptional down-regulation of the splicing-defective genes." *PLoS ONE*, vol. 5, p. e11540, 2010.