Ultra short and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution

Lukas Windhager, Thomas Bonfert, Kaspar Burger, et al.

Genome Res. published online April 26, 2012
Access the most recent version at doi:10.1101/gr.131847.111

Ultra short and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution

Lukas Windhager[1*], Thomas Bonfert[1*], Kaspar Burger[2], Zsolt Ruzsics[3], Stefan Krebs[4], Stefanie Kaufmann[1], Georg Malterer[3], Anne L'Hernault[6], Markus Schilhabel[5], Stefan Schreiber[5], Philip Rosenstiel[5], Ralf Zimmer[1], Dirk Eick[2], Caroline C. Friedel[1#], Lars Dölken[6#]

Affiliations

1 Institute for Informatics, Ludwig-Maximilians-Universität München, Amalienstr. 17, 80333 Munich, Germany.

2 Department of Molecular Epigenetics, Helmholtz-Zentrum München, Center of Integrated Protein Science (CIPSM), Marchioninistrasse 25, 81377 Munich, Germany

3 Max von Pettenkofer-Institute, Ludwig-Maximilians-Universität München, Pettenkofer Str. 9a, 80336 Munich, Germany

4 Gene Center, Ludwig-Maximilians-Universität München, Feodor-Lynen-Str. 25, 81377 Munich, Germany

5 Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, Schittenhelmstraße 12, 24105 Kiel, Germany

6 Department of Medicine, University of Cambridge, Box 157, Level 5, Addenbrooke's Hospital, CB2 0QQ, Cambridge, UK

* contributed equally

corresponding authors

Corresponding authors

Caroline C. Friedel

Institute for Informatics

Ludwig-Maximilians-Universität München

Amalienstr.17

80333 Munich, Germany

e-mail: caroline.friedel@bio.ifi.lmu.de

Tel.: +49-89-2180-4056

Fax: +49-89-2180-99-4056

Lars Dölken

Department of Medicine

University of Cambridge

Box 157, Level 5

Addenbrooke's Hospital

Hills Road

CB2 0QQ Cambridge, UK

e-mail: ld408@medschl.cam.ac.uk

Tel.: +44 1223 761304

Running title: Progressive 4sU-tagging reveals RNA processing

Keywords: 4sU-tagging, newly transcribed RNA, RNA-seq, RNA processing, 4-thiouridine, non-coding RNA

Abstract

RNA synthesis and decay rates determine the steady-state levels of cellular RNAs. Metabolic tagging of newly transcribed RNA by 4-thiouridine (4sU) can reveal the relative contributions of RNA synthesis and decay rates. The kinetics of RNA processing, however, so far remained unresolved. Here, we show that ultra-short 4sU-tagging not only provides snap-shot pictures of eukaryotic gene expression but, when combined with progressive 4sU-tagging and RNA-seq, reveals global RNA processing kinetics at nucleotide resolution. Using this method, we identified classes of rapidly and slowly spliced/degraded introns. Interestingly, each class of splicing kinetics was characterized by a distinct association with intron length, gene length and splice site strength. For a large group of introns, we also observed long lasting retention in the primary transcript, but efficient secondary splicing or degradation at later time points. Finally, we show that processing of most, but not all small nucleolar (sno)RNA-containing introns is remarkably inefficient with the majority of introns being spliced and degraded rather than processed into mature snoRNAs. In summary, our study yields unparalleled insights into the kinetics of RNA processing and provides the tools to study molecular mechanisms of RNA processing and their contribution to the regulation of gene expression.

Introduction

RNA levels in a cell are determined by the rates of transcription, RNA processing and RNA decay. Regulation may occur at all three levels providing substantial flexibility for adaption to alterations in environmental conditions (Jing et al. 2005; Kim et al. 2009; Nilsen and Graveley 2010). Most studies focus on regulation at the transcriptional level but changes in RNA degradation rates may also significantly alter gene expression of coding and non-coding RNAs (Cazalla et al. 2010; Miller et al. 2011; Shalem et al. 2008). So far, little is known about the contribution of alterations in RNA processing to gene expression. Furthermore, despite the knowledge on the occurrence of multiple isoforms of transcripts, the dynamic mechanisms guiding tissue- and context-specific regulation of RNA processing (e.g. alternative splicing events) remain unknown. Research has been severely hampered by the lack of proper tools to study these processes with sufficient resolution. Next-generation sequencing of total cellular RNA (RNA-seq) allows studying the outcome of RNA processing at whole-transcriptome level at a given time point (Pan et al. 2008; Wang et al. 2008). This has recently resulted in the discovery of numerous new alternative isoforms of mammalian transcripts indicating that most multi-exon genes are alternatively spliced (Nilsen and Graveley 2010). The kinetics of RNA splicing and processing and thus the underlying regulatory mechanisms, however, can hardly be resolved with these techniques.

Metabolic labeling of newly transcribed RNA using 4-thiouridine (4sU-tagging), a naturally occurring uridine derivative, provides direct access to newly synthesized transcripts with minimal interference to cell growth and gene expression (Cleary et al. 2005; Dölken et al. 2008; Friedel et al. 2009; Kenzelmann et al. 2007; Melvin et al. 1978; Weintz et al. 2010). Following isolation of total cellular RNA and thiol-specific biotinylation, this can be quantitatively separated into tagged (newly transcribed) and untagged (pre-existing) RNA using streptavidin-coated magnetic beads. This allows bias-free analysis of RNA synthesis

and decay at high resolution. We and others have demonstrated that this approach provides access to the dynamics of RNA production and degradation in eukaryotic cells. Furthermore, it is directly compatible with microarray analysis (Dölken et al. 2008; Friedel and Dölken 2009; Friedel et al. 2009) and RNA-seq (Rabani et al. 2011; Schwanhäusser et al. 2011). However, only relatively long durations of 4sU-tagging were used together with RNA-seq so far. Here, we show that ultra-short 4sU-tagging with as little as 5 min labeling time can be combined with RNA sequencing to provide high-quality sequencing data. The combination of ultra-short and progressive 4sU-tagging from 5 to 60 min labeling time then allows unparalleled insights into the kinetics of RNA processing, in particular RNA splicing and processing of non-coding RNAs.

Results

Ultra-short 4sU-tagging is compatible with RNA-seq in human B-cells

Newly transcribed RNA obtained by 4sU-tagging contains substantially greater amounts of large, unprocessed transcripts than regularly found in total cellular RNA. This is readily visualized by electrophoretic analysis (Dölken et al. 2008). When shortening the duration of 4sU-tagging the average age of nascent transcripts in newly transcribed RNA decreases. We thus hypothesized that RNA-seq combined with progressive reduction of the duration of 4sU-tagging could be employed to study the kinetics of RNA processing. For this purpose, we performed a time course experiment of 4sU-tagging in DG75 human B-cells consisting of five samples with 60, 20, 15, 10 and 5 min of 4sU-tagging. At the end of 4sU exposure, cells were harvested using Trizol, total cellular RNA was prepared and newly transcribed RNA was purified. The relative abundance of newly transcribed, tagged RNA in total cellular RNA decreased from 3.5% of total RNA after 1 h 4sU-tagging to about 0.8% after 5 min (Figure 1 A). Newly transcribed RNA from all five labeling conditions was subjected to RNA-seq

analysis using sequencing by ligation (SOLiD II, Applied Biosystems). In the following, we will refer to these samples as '5 min 4sU-RNA' to '60 min 4sU-RNA'. In addition, total and untagged RNA following 60 min of 4sU-tagging were sequenced. As newly transcribed RNA may not yet be poly-adenylated, no poly-A selection was performed. For each sample, between 2.8 and 4.4 million reads could be uniquely mapped by aligning them first to the human transcriptome and the remaining unaligned reads to the human genome. All samples passed standard quality controls.

We first assessed the contribution of intronic and exonic sequences in the seven RNA samples. As predicted, the number of reads mapping to intronic sequences increased with reduced duration of 4sU-tagging from 18.9% in untagged RNA to 75.9% in 5 min 4sU-RNA (Figure 1 B). As excised introns are generally believed to be rapidly degraded (Clement et al. 1999; Lamond et al. 1988; Nam et al. 1997) this indicates the presence of large amounts of unspliced pre-mRNAs in the newly transcribed RNA samples. Our data thus provide solid evidence that as little as 5 min of 4sU-tagging yields high quality newly transcribed RNA fully compatible with next generation sequencing. If none of the transcripts in 5 min 4sU-RNA had undergone any splicing events, the intronic reads would have been predicted to contribute ~89% instead of 75.9% of all reads (Figure 1 B). Thus, a substantial fraction of cellular transcripts in 5 min 4sU-RNA has already undergone splicing events with >65% (conservative estimate) of all introns already decayed (see Supplementary Methods on how this estimate was obtained).

Similar to the changes in the contribution of intronic reads over time, we observed a strong correlation between the number of reads crossing exon-intron or exon-exon junctions and the duration of 4sU-tagging (Figure 1 C). Exon-intron junction reads result from unspliced or partially spliced transcripts. Accordingly, their contribution considerably decreased with the

duration of 4sU-tagging (from 1.1% in 5 min 4sU-RNA to 0.29% in untagged RNA). Conversely, the frequency of exon-exon junction reads increased from 1.9% in 5 min 4sU-RNA to 12% in untagged RNA. In conclusion, these results show that progressive 4sU-tagging visualizes the maturation of transcripts over time.

Intron decay is correlated to intron length, gene length and splice site strength

Interestingly, the decrease in the number of exon-intron junction reads appeared to be delayed compared to the decrease in the intronic reads themselves (Figure 1 C). As the relative contribution of long introns to the number of intronic reads is higher than to the number of exon-intron reads (Supplementary Figure 1 A), we hypothesized this to be due to a faster decay of long introns. In this case, intronic read numbers, which are disproportionately determined by long introns, would decrease faster with increasing labeling time than exon-intron junction read numbers. Indeed, no delay could be observed when read numbers were analyzed selectively for three approximately equally sized groups of introns with similar length (Supplementary Figure 1 B-D).

To quantify the rate of decay for each intron, ratios of intron reads in 60 min 4sU-RNA compared to 5 min 4sU-RNA (60/5 min ratios) were calculated and correlated with intron length. Thus, high 60/5 min ratios correspond to low decay rates and vice versa. On the individual intron level, a correlation with intron length was basically non-existent (spearman rank correlation r_s=-0.037, p-value for a correlation different from zero=0.0039). However, when introns were binned according to length and average 60/5 min ratios were calculated for each bin, a very clear trend was observed (Figure 1 D). Interestingly, the highest 60/5 min ratios, i.e. the lowest decay rates, were not observed for the shortest introns, but for introns in a range of ~300-400 nucleotides (nt). Decay rates increased both for introns longer as well as shorter than this range, with the highest decay rates observed for introns longer than ~1500 nt.

Accordingly, after normalizing intronic read counts to intron length, the lag between the loss of intronic reads and exon-intron junction reads over time disappeared (Figure 1 E).

To investigate whether these findings were influenced by any bias introduced by our experimental setup or computational analysis, we correlated 60/5 min ratios and intron length with other intron features. While there was no significant correlation between the relative position of an intron in a gene and its 60/5 min ratio or length, gene length was correlated to both characteristics (r_s= -0.24 for 60/5 min ratios and r_s=0.38 for intron length, p-value<10^{-15} in both cases). Here, the correlation between gene length and 60/5 min ratio was very clear when binning according to gene length (r_s=-0.94): as average gene length increased, average 60/5 min ratios decreased. An analysis of 60/5 min ratios binned first according to gene length into three groups and then for each group binned according to intron length, indicated that the contributions of gene and intron length were mostly independent of each other (Supplementary Figure 1 E). Irrespective of gene length, intron length and 60/5 min ratios showed the same characteristic correlation with a peak of 60/5 min ratios in the low-to-medium intron length range (300-400 nt) and decreasing ratios on either side.

Finally, we investigated whether splice site strength had any effect on splicing kinetics. Splice site strength was quantified in terms of splice site scores calculated with MaxEntScan (Yeo and Burge 2004). Similar to intron length, there was only an extremely weak correlation between splice site scores (5' only, 3' only and 5'+3' scores) and 60/5 min ratios (r_s = -0.055, -0.044 and -0.072, respectively) or intron length (r_s = 0.055, 0.024 and 0.046, respectively). However, when binning introns according to the splice site strength, there was a trend towards higher 60/5 min ratios (slower splicing kinetics) for very weak splice sites (Supplementary Figure 1 F). This effect was mostly independent of intron length (Supplementary Figure 1 G).

Distinct classes of introns are defined by their splicing kinetics

To further investigate differences in the kinetics of intron processing, we first focused on the most highly expressed genes as intron expression levels in newly transcribed RNA although substantially higher than in total RNA are much lower than expression levels of the surrounding exons. This is due to the large fraction of introns (>65%) already spliced and decayed in 5 min 4sU-RNA. Expression levels of genes were quantified in terms of reads per kilobase of gene per million mapped reads (RPKM) after normalizing for mappability (see methods) and the analysis was focused on genes with an RPKM $\geq$ 11 in all RNA samples (525 genes). Since ribosomal proteins are generally highly expressed, they, as well as other genes involved in translation and ribosome biogenesis, were enriched in this set. Other enriched functions included RNA splicing, ATPase activity, cell cycle and regulation of apoptosis (see Supplementary Table 1 for a full list). For these genes, we distinguished between introns absent (RPKM < 0.5: 1,014 introns, Supplementary Table 2) or present (5,838 introns, Supplementary Table 3) in 5 min 4sU-RNA. Even after excluding 50 absent introns (~5%) that were shorter than the read length (35 nt) and, thus, could not contain any intronic reads, absent introns were significantly shorter than present ones (Wilcoxon test, p-value<10^{-15}). Furthermore, they were located closer to the 3' end of the gene than present introns (Wilcoxon test, p-value=0.0042) with 12% of the absent introns being the last intron of the gene compared to 7% for present introns (Fisher's exact test, p-value<10^{-6}). This suggests that at least some of these introns were part of longer transcript versions that were not transcribed in this form in the DG75 cells. For other introns, possible explanations for their absence in 5 min 4sU-RNA might be 1) very fast co-transcriptional splicing, 2) problems in sequencing or 3) problems in mapping these parts of the pre-mRNA, e.g. due to repetitive sequences. Interestingly, in many absent introns, both neighboring exons were well expressed and precisely delimited. This indicates rapid co-transcriptional splicing and intron degradation rather than sequencing bias. In addition, there was no significant increase for the absent

introns in the frequency of repetitive sequences identified by RepeatMasker (Smit et al. 1996-2010) or the frequency of non-unique read mappings. Notably, the fraction of absent intron positions contained within repetitive sequences was actually significantly smaller than for present ones (Wilcoxon test, p-value<10^{-9}). These analyses confirm that numerous transcripts in 5 min 4sU-RNA (>65%) had already been spliced and their introns had been degraded.

Remarkably, exon-intron junction reads for these absent introns were regularly observed, although only at around 60% of the level of introns present in 5 min 4sU-RNA (Wilcoxon test, p-value<10^{-15}). The frequency of exon-exon junction reads was reduced to a similar degree. This indicates that the proportion of completed splicing events including the fusion of neighboring exons was not significantly higher than for present introns. A possible explanation for this observation is that major parts of the absent introns are already degraded while the exon-intron junctions are still connected to the splicing machinery in a state prior to exon fusion.

An example for such an absent intron is shown for the *RPL23A* gene (Figure 2 A). In this case, the absent intron is surrounded by a retained intron contained in an alternative transcript of *RPL23A.* According to the Ensembl annotation, this alternative transcript is not translated. It is important to note that the absence of this intron is not due to repetitive sequences in this region, which might result in non-uniquely mapped reads and their subsequent exclusion during mapping. Although two repeats were identified by RepeatMasker in this absent intron, sequence divergence to the corresponding repeat consensus was rather high (6.2% and 11.3%, respectively). Accordingly, only one additional non-uniquely mapping read was found in all samples. This confirms that the low expression of this intron is not an artifact created by the removal of non-unique mappings. Interestingly, the intron retention in *RPL23A* is still observed at considerable levels in 60 min 4sU-RNA, but hardly detectable in total and

untagged RNA. This suggests that this alternative splicing variant is erroneously produced at considerable levels but quality control mechanisms of the cell later on result in its delayed but nevertheless highly efficient removal. This may either result from decay of the whole transcript or secondary splicing events removing only the retained intron.

To systematically mine for such and other distinctive kinetics of intron processing, we used cluster analysis on the introns still present in 5 min 4sU-RNA. Introns were clustered according to their relative abundance compared to the expression level of the gene (intron/gene ratio). The intron/gene ratio approximately represents the fraction of pre-mRNAs or transcripts still containing the intron. A stable clustering was obtained by repeated k-means clustering (see methods). The resulting 20 largest clusters represented 85% of introns (Supplementary Figure 2 A). Although distinct differences between introns of the same gene were observed, they were significantly enriched in the same clusters (Fisher's exact test, p-value$<10^{-15}$). This indicates that introns of the same genes tend to have similar splicing kinetics.

A visual inspection of the clusters identified distinctive subgroups of clusters with different absolute 5 min intron/gene ratios but similar trends across the time-course, i.e. similar relative ratio changes over time. For random data sets analyzed as controls, no distinctive subgroups were observed (Supplementary Figure 2 B). To systematically identify such subgroups of splicing patterns, the 20 largest clusters were clustered again after normalizing the intron/gene ratios to 5 min 4sU-RNA. Thus, exon/intron ratios in 5 min 4sU-RNA were set to 100%. This resulted in 4 distinct classes of intron splicing (Figure 2 B and Supplementary Table 3). The largest class, Class 1, representing 3,908 introns, was characterized by an almost linear slope in intron/gene ratios up to 20 min 4sU-RNA. Class 2 (580 introns) showed a smaller slope than Class 1, suggesting a lower splicing rate. In contrast, Class 3 (338 introns) was

characterized by a more rapid decrease in intron/gene ratios between 5 and 10 min 4sU-RNA, indicating a faster splicing rate. Finally, Class 4 (109 introns) showed a very interesting pattern. Here, intron/gene ratios remained remarkably stable from 5 to 60 min 4sU-RNA, but then dropped significantly some time later to reach similar low levels in total and untagged RNA as observed in the other classes. Exemplary introns for Class 1 and 4 are shown in Figure 2 C-D. Please note that the division into different classes was not always clear-cut. For instance, although cluster 3 (238 introns) is assigned to Class 2, it shows a similar trend as Class 4 introns with stable intron levels up to 20-60 min and subsequent decay. In general, Class 2 constitutes a weaker version of the delayed decay in Class 4, while Class 3 contains extreme cases of rapid intron removal and decay.

Monitoring the fate of alternative splicing products

Interestingly, the Class 4 introns and to a lesser degree Class 2 introns showed a pattern very similar to the known intron retention example shown in Figure 2 A. This suggests that these introns likely constitute novel alternative splicing events resulting in retained introns. To provide evidence for this hypothesis, degradation patterns for known alternative splicing products were investigated. For this purpose, exon boundaries were recalculated based on Ensembl transcripts and the re-defined exons were classified either as core exons, retained introns or alternative 3' or 5' exon ends (see Supplementary Material). Exon expression levels were divided by the overall gene expression level (exon/gene ratios) and normalized to 5 min 4sU-RNA (Figure 2 E). As expected, core exon/gene ratios were stable across all time points. In contrast, for retained introns the same pattern was observed as for Class 4 and some Class 2 introns. The same pattern was also seen for alternative 3' and 5' ends, however, much less pronounced. This implies that alternative transcripts containing retained introns are present at substantial numbers (>30%) until 60 min 4sU-RNA, but are subsequently degraded quite rapidly, e.g. by nonsense-mediated decay or by secondary splicing events.

Properties of splicing classes

Intron length, gene length and splice site strength were analyzed to reveal their contribution in defining the kinetics of splicing in the four classes of introns. Interestingly, significant differences in the distribution of these features were observed (Figure 3 A). Class 1 introns were on average longer (by 36%, Wilcoxon test, FDR corrected p-value $<10^{-12}$), contained in longer genes (by 76%, p-value $<10^{-15}$) and showed increased splice site strength (by 4.5%, p-value $<10^{-6}$) compared to introns of the other classes. As all of these factors were positively correlated with intron decay rates, it is not surprising that these introns are lost relatively fast compared to the majority of other introns (Class 2 and 4). Strikingly, Class 3, which was characterized by the most rapid intron decay, contained shorter introns and shorter genes than Class 1, while splice site strength was neither significantly increased nor decreased. In contrast, Classes 2 and 4, which also contained significantly shorter introns, were characterized by both significantly shorter gene length and weaker splice sites compared to Classes 1 and 3 (Wilcoxon test, p-value$<10^{-5}$). These results suggest that for short introns, gene length and splice site strength make a marked difference resulting in either very fast or very slow splicing and intron decay. In particular, reduced splice site strength may result in the retention of introns (resulting in Class 4 or Class 2 introns) due to poor recognition of splice sites by the splicing machinery. For the latter two classes, the strongest difference was in intron length with Class 2 having significantly longer introns than Class 4. Thus, the four classes are characterized by distinct combinations of intron length, gene length and splice site strength.

Characteristics of snoRNA processing

Interestingly, both Class 2 and Class 4 were significantly enriched for small nucleolar RNA (snoRNA) precursor introns. SnoRNAs are small non-coding RNAs mainly involved in the chemical modification of other non-coding RNAs, in particular rRNAs. They are mostly

encoded within introns of protein-coding or non-coding genes and are excised from these larger RNAs during splicing (Dieci et al. 2009; Hirose et al. 2003). In our case, 22 out of 580 Class 2 introns (3.8%, Fisher's exact test, FDR corrected p-value=0.0027) and 7 out of 109 Class 4 introns (6.4%, Fisher's exact test, FDR corrected p-value=0.011) contained snoRNAs. In contrast, snoRNA containing introns were under-represented in Class 1 (52 out of 3,908 introns, 1.3%, p-value=0.00024). To analyze this observation in more detail, the RNA-seq data were used to investigate processing of snoRNA transcription units. Almost all human snoRNAs (>90%) are processed from introns with only a single snoRNA generated from each intron (Dieci et al. 2009). For these snoRNAs, splicing is generally required to make the 5'- and 3'-ends of the snoRNA-containing intron accessible to trimming by nuclear RNAses. As the majority of snoRNA-containing genes were found to be among the most highly transcribed genes, this group of rather small non-coding RNAs was ideally suited for further analysis. Again, we focused on the 5,838 well expressed introns of which 121 contained snoRNAs. Of these snoRNA precursor introns, 88 (72%) were assigned to one of the four classes. The remaining ones were not assigned to the 20 largest clusters. Please note that in this and all previous analyses reads mapping to snoRNAs were assigned only to the snoRNAs and not used for calculating the intron levels.

As Classes 2 and 4 showed reduced intron decay rates, we wondered whether this might be a general feature of snoRNA containing introns. Indeed, intron/gene ratios of snoRNA precursor introns were significantly increased in all RNA fractions compared to all other introns (Figure 3 B). This indicates that introns containing snoRNAs are characterized by significantly reduced splicing and degradation rates. Nevertheless, the shape of the curve is the same as for Class 1 with a linear decrease within the first 20 minutes and no "bulge" as observed for Classes 2 and 4. Accordingly, snoRNA intron processing generally appears to be a continuous process and not a case of temporary intron retention and delayed

processing/decay. Remarkably, splice site scores of snoRNA containing introns were not significantly reduced compared to other introns. This suggests that the reduced decay rates are not due to poor recognition of the splice site by splicing factors. In contrast, both intron length and length of the corresponding genes were reduced on average by about 40% (Wilcoxon test, p-value = 0.0005) and 60% (p-value < 10^{-15}), respectively. As both short-to-medium intron length and short gene length have been found to be associated with reduced intron decay rates, even if snoRNA encoding introns were excluded from the analysis, this may provide a possible explanation for the smaller splicing rates of snoRNA introns.

One representative example of a snoRNA precursor intron is shown in Figure 3 C. Here, the *SNORD4B* snoRNA is excised from an intron of the *RPL23A* gene. In this case, large parts of the precursor intron were clearly present until 60 min 4sU-RNA, whereas only the mature snoRNA was detected in total and untagged RNA. Interestingly, the mature snoRNAs can be identified by reads starting predominantly at the 5'-end of the snoRNA. These sequences most likely represent cloning products of mature snoRNAs not affected by the fragmentation procedure applied during library preparation. Any fragmentation of these small non-coding RNAs probably resulted in the loss of the resulting snoRNA fragments as the standard SOLiD protocol includes a size selection during library preparation to remove cDNAs without inserts.

Due to this apparent cloning/sequencing bias, determination of snoRNA stability using the RNA-seq data is not possible. Thus, we analyzed snoRNA half-lives obtained using Affymetrix Gene ST 1.0 microarrays in DG75-eGFP and two other human B-cell lines (DG75-10/12 and BCBL-1) (Dölken et al. 2010) based on newly transcribed / total RNA ratios. Surprisingly, snoRNAs were highly enriched among the most short-lived transcripts in all three B-cell lines in the microarray data (Figure 4 A). Here, snoRNAs represented >20 of the 40 most short-lived transcripts in all three human B-cell lines. This was surprising as these

small non-coding RNAs are involved in metabolic processes, which are usually associated with long-lived transcripts (Friedel et al. 2009). In addition, there is so far no evidence for snoRNAs being used up and metabolized while exerting their function in rRNA maturation. Furthermore, snoRNAs are small RNA molecules (60-300 nt (Kiss 2002)). Therefore, they are prone to reduced capture rates during 4sU-tagging due to their very low uridine (and thus low 4sU) content (Friedel et al. 2009; Miller et al. 2009). Capture rates can be assessed by comparing the $\log_2$ (newly transcribed/total RNA) ratios to the uridine content of transcripts (Figure 4 B). As expected, average log-ratios decreased for very short transcripts with very low uridine content. For virtually all snoRNAs, however, these ratios were not only not reduced, but rather surprisingly high for transcripts of such short length. If we had corrected for size-dependent reduced capture efficiency using computational methods (Miller et al. 2011; Miller et al. 2009), these ratios would have been even higher resulting in even shorter RNA half-lives.

Remarkably, the apparent short snoRNA half-lives from microarrays could also not be verified by northern blot analysis. Here, snoRNA levels for 8 short-lived, 7 medium-lived and 3 long-lived snoRNAs (according to the microarray data) were determined following 6 h transcriptional arrest with either Flavipiridol (FL) or Actinomycin D (Act-D). The very short-lived c-myc transcript and the long-lived glyceraldehyde-3-phosphate dehydrogenase (GAPDH) served as controls. While a reduction in snoRNA levels following transcriptional arrest was observed for three snoRNAs (*SNORD22*, *SNORA38B* and *SNORA73A*), expression of the majority of snoRNAs remained unaltered (Figure 4 C). Although a rapid stabilization of these snoRNAs due to transcriptional arrest cannot be ruled out completely, the alterations were much smaller than for other short-lived RNAs (e.g. *MYC* or *FOS*) known to be rapidly stabilized after transcriptional arrest. We conclude that most mature snoRNAs molecules are not highly unstable, in contrast to what was implied by the microarray data. However, as the

majority of snoRNAs are well expressed in human B-cells and were highly consistent in three different cell lines, these measurements are very unlikely to be due to technical problems or artefacts of the microarray measurements.

As microarray probes cannot distinguish between mature snoRNAs and their precursors, we hypothesized that microarrays measured the short RNA half-lives of instable snoRNA precursors, rather than of the mature snoRNAs. Accordingly, the very short half-lives observed for many "snoRNA" probe sets in the microarray data likely represent the half-lives of their parental introns. Thus, they correspond to rapid splicing and decay of the parental introns, which are degraded rather than processed to a (stable) mature snoRNA. In conclusion, the short snoRNA half-lives observed by microarray analysis actually reflect inefficient processing of many snoRNAs, i.e. only a small fraction of introns is actually processed to mature snoRNAs while the majority of introns are simply spliced and degraded. Interestingly, very short microarray-derived half-lives were observed only for snoRNAs encoded within introns of non-coding and protein-coding genes, but not for snoRNAs with independent promotors, which provides further support to this conclusion (Figure 4 D).

In contrast to most other snoRNAs, snoRNAs of the *SNORD116* cluster were found to possess rather long RNA half-lives (>5 h) in the microarray data (Figure 4 D). This would be consistent with substantially more efficient snoRNA processing. It is important to note that their long RNA half-lives were not due to low signal intensities of the corresponding probe sets. These snoRNAs all derive from introns of the *SNURF–SNRPN* transcription unit (Runte et al. 2001), which is expressed from the imprinted *SNURF-SNRPN* domain. It encodes for at least two long, paternally expressed non-coding RNAs, which both encode for multiple intronic snoRNAs (Royo and Cavaille 2008; Vitali et al. 2010). Little is known about their biogenesis and function. The second cluster of snoRNAs (*SNORD115*) encoded in this

domain is brain-specific (Cavaille et al. 2000) and was hardly expressed at all in the B-cells under study. Although found only at relatively low levels in our RNA-seq data, the mature snoRNAs of the *SNORD116* cluster were expressed at greater levels than the surrounding intronic sequences (Figure 5 A). Furthermore, reads derived from the *SNORD116* cluster generally started at the 5'-end of the snoRNA even in 5 min 4sU-RNA (Figure 5 B) which is in stark contrast to the other intronic snoRNAs (Figure 5 C). This strongly indicates that these reads derive from mature snoRNAs. Accordingly, most *SNORD116* snoRNAs were already fully processed in 5 min 4sU-RNA, which indicates highly efficient snoRNA processing within the *SNORD116* cluster.

Discussion

Metabolic labeling of newly transcribed RNA with 4-thiouridine (4sU-tagging) is superior to simply analyzing total cellular RNA and allows distinguishing alterations in RNA synthesis and decay (Dölken et al. 2008; Friedel et al. 2009; Rabani et al. 2011). In this study, we show that characteristics of RNA processing can be examined by performing ultra-short and progressive 4sU-tagging combined with RNA-seq. In human B-cells, as little as 5 min of 4sU-tagging provided sufficient amounts of newly transcribed RNA to perform RNA-seq. Previously, 4sU-tagging was shortened to 10-15 min but only combined with microarray (Dölken et al. 2008) and nCounter technology (Rabani et al. 2011) as not enough RNA material was thought to be recoverable for sequencing (personal communication). RNA-seq was only performed following much longer labeling (e.g. 2 h (Schwanhäusser et al. 2011)). In order to shorten the duration of 4sU labeling down to 5 min and still obtain sufficient newly transcribed RNA for next-generation sequencing, we increased the amount of total RNA we started with from 70 to 150 microgram. In addition, we benefitted from a slightly increased efficiency of 4sU uptake by cells growing in suspension (B-cells) compared with adherent

cells (murine fibroblasts). As 4sU-incorporation is strongly dependent on the employed 4sU concentration, this approach is readily adaptable to other cell types by simply increasing the 4sU concentration for ultra-short labeling.

The high purity of the newly transcribed RNA samples was confirmed by the much higher percentage of intronic reads and exon-intron junction reads inversely correlated with the duration of labeling. Interestingly, our data showed that >65% of intronic sequences had already been spliced and decayed in 5 min 4sU-RNA, which highlights the need to perform ultra-short 4sU-tagging. To further characterize processing of introns, we analyzed the relative loss of intronic reads over time (5 min 4sU-RNA to 60 min 4sU-RNA) to define splicing kinetics for >5,800 introns and correlated this with various features. Remarkably, intron decay rates were found to be influenced by both intron and gene length. The correlation to intron length was mostly independent of gene length as it was also observed for genes of similar length. As these observations were based on the comparison of read counts in 60 min 4sU-RNA to 5 min 4sU-RNA, it is unlikely that they are due to sequencing artifacts or insufficient length normalization. These problems would affect intron read counts in both samples to a similar degree and, thus, they would not be relevant when calculating ratios of read counts for each intron. In conclusion, although we cannot completely exclude a minor bias introduced during sample and library preparation negatively affecting sequence reads derived from very large genes/pre-mRNAs, our data indicate that any bias introduced did not substantially affect the major findings of this study.

Interestingly, numerous introns were already absent in 5 min 4sU-RNA. Although some of these probably represented non-transcribed variants at the 3’ end of a gene and some may simply reflect sequencing bias, a considerable number were surrounded by well-expressed exons suggestive of very fast splicing. Remarkably, we still observed exon-intron junction

reads for these absent introns albeit at a lower frequency than for introns present in 5 min 4sU-RNA. Furthermore, the number of exon-exon junction reads was not increased compared to the number of exon-intron junction reads. This hints at degradation of these introns concurrently with splicing, which would be consistent with previous results showing that tethering of exons to the RNA polymerase II results in correct splicing of these exons even though the intron connecting them may no longer be continuous, e.g. due to ribozyme-mediated cleavage and partial degradation (Dye et al. 2006).

When looking more closely at the kinetics of splicing, we identified clusters of introns spliced with distinct kinetics accompanied by an enrichment for similar splicing kinetics within a given gene. It is important to note that the reported splicing kinetics were also observed for less abundant genes with more diverse functions. When classes were extended by so far unclassified present introns (RPKM $\geq$0.5) of genes with a lower minimum expression value (RPKM $\geq$5) (2,468 genes; 29,947 introns), the same trends were observed for each class as before (Supplementary Figure 2 C). The most interesting kinetics were observed in introns of Class 4 and less pronounced in Class 2. These showed relative high intron levels even in 60 min 4sU-RNA - indicative of retained introns. This was associated with reduced splice site strength most likely resulting in missed splice site recognition contributing to intron retention. Nevertheless, introns levels dropped substantially in total and untagged RNA. This is consistent with a delayed but nevertheless efficient removal of these intronic sequences. Many known and at least 100 novel alternatively spliced transcripts containing retained introns showed this pattern. This temporally delayed but nevertheless eventually efficient removal of retained introns may either be due to RNA degradation by nonsense-mediated RNA decay or may represent novel secondary, i.e. post-transcriptional splicing events. In principle, the two could be distinguished as secondary splicing events would only remove the initially retained introns while nonsense-mediate decay would also affect levels of the surrounding exons and

thus overall transcript RNA half-lives. Unfortunately, the fraction of transcripts of a gene still containing retained introns in 5 min 4sU-RNA was not much larger than 30% on average. Thus, the effect on half-lives was not large enough to differentiate the two alternatives based on the current data. Further studies will thus need to employ knock-out/down approaches targeting nonsense-mediated RNA decay to answer this interesting question. Nevertheless, both alternatives suggest highly effective cellular quality control measures to ensure correct splicing, as the abundance of transcripts containing retained introns in total and untagged RNA was very low.

Another interesting finding of this study was the poor processing efficiency of many but not all human snoRNAs. In microarray data of newly transcribed/total RNA ratios derived from 1 h 4sU-tagging (Dölken et al. 2010) snoRNAs seemed to be the most short-lived cellular transcripts in three human B-cell lines. Northern blot analysis of mature snoRNAs after 6 h of transcriptional arrest clearly excluded such short RNA half-lives for most of the mature snoRNAs, thereby at first contradicting the microarray-based measurements. However, as microarray probe sets cannot differentiate between a small mature snoRNA and its much larger precursor, these “seemingly” short RNA half-lives only reflect inefficient processing of snoRNAs from their much larger precursors, i.e. degradation of the parental introns without processing to the mature snoRNAs. Poor processing efficiency of snoRNAs derived from intronic sequences indicates competition between splicing factors (resulting in subsequent intron degradation) and the snoRNA processing machinery. Evidence for this hypothesis is provided by the observation that introns containing snoRNAs are spliced and degraded much slower than other introns. It is tempting to speculate that the low basal processing efficiency of many snoRNAs may offer the opportunity for significant regulation of snoRNA expression levels by modifying their processing efficiency. Indeed, up-regulation of snoRNA levels has recently been reported for cells expressing a mutant form of the carboxy-terminal domain

(CTD) of the large subunit of RNA polymerase II (Sims et al. 2011). Expression of a CTD mutant deficient in arginine methylation resulted in a significant increase of steady state levels of a variety of snoRNAs and snRNAs (small nuclear RNAs), while the levels of all other categories of RNAs, e.g. mRNAs, remained unaffected. It will be interesting to test whether this mutant shows alterations in RNA splicing, which might account for the alterations in snoRNA levels. An alternative explanation for the large number of snoRNA-containing introns being spliced and degraded rather than processed into mature snoRNAs would be that the expression levels of snoRNA-binding proteins may not be sufficiently high to bind all newly synthesized snoRNA precursors and thereby prevent their degradation. In this case, snoRNA levels would not be defined transcriptionally but by the abundance of their respective snoRNA-binding proteins.

Interestingly, RNA processing of intronic snoRNAs was not always found to be inefficient as exemplified by the majority of snoRNAs of the *SNORD116* cluster. Although we did observe weak expression of the intronic regions surrounding the snoRNAs, accumulation of reads derived from the 5' end of the mature snoRNAs in all samples provided evidence that mature snoRNAs were already generated within the first few minutes of synthesis and thus became apparent even in 5 min 4sU-RNA. This also indicates that snoRNA processing can indeed be both fast and efficient. For other snoRNAs with less efficient processing, this clear bias of read starts characteristic for mature snoRNAs was only observed later on indicating that the majority of their parental introns are not rapidly spliced and processed to snoRNAs. The underlying molecular mechanism of this substantial difference in processing efficiency of the paternally imprinted *SNORD116* snoRNAs remains to be elucidated.

RNA processing involves numerous complex and highly regulated molecular processes. This field of research has recently received a dramatic boost by the development of high-

throughput technological for studying RNA-protein interactions at nucleotide resolution. Both HITS-CLIP and PAR-CLIP can now be employed to reveal thousands of protein-RNA interaction sites in a single experiment and thus serve to unravel the underlying functional networks and molecular mechanisms (Hafner et al. 2010; Ule et al. 2003; Ule et al. 2006). Our approach now provides the means to directly study the functional consequences of these RNA-protein interactions on RNA processing. In this study, we restricted our approach to RNA splicing and the processing of the highly abundant snoRNAs. With more in depth sequencing and sequencing of shorter RNAs, RNA processing of other non-coding RNAs, such as miRNAs or lincRNAs, could also be studied in detail. One of the most important conclusions of this work is that RNA processing efficiency needs to be considered when studying the regulation of RNA expression levels. RNA processing is likely to provide a notable contribution to the regulation of both large and small non-coding RNAs. Application of RNA-seq combined with ultra-short and progressive 4sU-tagging will thus dramatically enhance our understanding of the underlying molecular mechanisms and regulatory networks.

Methods

Cell culture and 4sU-tagging

DG75-eGFP human B-cell lines were cultured in RPMI 1640 medium containing 10% fetal calf serum, 100 IU/ml Penicillin, 100 µg/ml Streptomycin and 2 mM L-Glutamin. Newly transcribed RNA was labeled for 5, 10, 15, 20 or 60 min using 500 µM 4sU. Total cellular RNA was prepared from cells using Trizol reagent (Invitrogen) following the modified protocol by Chomczynski *et al.* (Chomczynski and Mackey 1995). Newly transcribed RNA was purified from 150 µg (5-15 min 4sU), 100 µg (20 min 4sU) and 70 µg total RNA (60 min 4sU). Separation of total RNA into newly transcribed and untagged pre-existing RNA was performed as described (Dölken et al. 2008) using 100 µl streptavidin beads (Miltenyi Biotec).

Next-generation sequencing

RNA sequencing was performed using the sequencing-by-ligation SOLiD II platform (Applied Biosystems). A modified whole-transcriptome analysis (WTAK) protocol was employed. 300 ng of RNA were used in the library construction process. Volume of the samples was adjusted by vacuum centrifugation, as concentrations of the early newly transcribed RNA samples tended to be low after depletion. Integrity, size distribution and yield were monitored after RNaseIII fragmentation (10 minutes, 37°C) on an Agilent RNA 6000 PicoChip. If newly transcribed RNA samples resulted in larger than expected fragments, samples were re-digested using RNAseIII for additional 2 minutes at 37°C, re-purified and re-checked as described above. Libraries of individual samples were prepared using barcoded adaptors and sequenced on the SOLiD instrument using standard protocols. No deviation in quality or length of the newly transcribed RNA sequences was observed when comparing to other standard RNAs.

Read mapping

Reads were mapped in a 3-step process using the Bowtie alignment program (Langmead et al. 2009). First, reads were aligned to pre-rRNA sequences (18S, 5.8S, 28S and spacer regions). The remaining unmapped reads were aligned to all Ensembl transcripts (Ensembl version 60) excluding pseudo-genes and haplotypes to identify exonic and exon-exon junction reads (aligned reads overlapping an exon-exon junction by ≥1 nt). Reads that remained unmapped after step two were aligned to the human reference genome (GRCh37/hg19) to identify intron and exon-intron junction reads (overlapping an exon-intron junction by ≥1 nt). For each read the best alignment location was used. Reads mapping equally well to two different locations were discarded. The following Bowtie settings were used: seed region = first 20 nt, 3 mismatches allowed in the seed, 5 in the whole alignment.

Quantification of gene, exon and intron expression levels

Expression levels of genes, exons and introns were estimated using the standard RPKM measure (= number of reads per kilobase of gene, exon or intron per million mapped reads) (Mortazavi et al. 2008). Number of reads mapping to a gene were determined as the total number of exon and exon-junction reads for this gene. To calculate RPKM values for exons and introns respectively, only reads mapping completely within this region were used. To account for problems in mapping reads to repetitive sequence regions, the effective length of exons and introns was used instead of the actual length. The effective length was calculated in the following way. First, *in-silico* reads were simulated by sliding a window across gene regions with the size of the read length in the experiment (35 nt). Thus, the simulated read set contains exactly one read from each position in each gene. The simulated reads were then mapped using the same 3-step procedure described above. The effective length was then defined as the number of positions within the respective region (exon, intron or gene) which had exactly one correctly and uniquely mapped read starting at this position.

Intron clustering

Introns were clustered based on the ratios of intron RPKM to gene RPKM (intron/gene ratio) across all samples (feature vector for clustering = intron/gene ratios at 5, 10, 15, 20, 60 min 4sU-RNA, total and untagged RNA). Introns were first clustered 100 times using standard k-means clustering (k=10) and Euclidean distance metrics. Final clusters were then defined as introns always clustered together in the 100 k-means runs. To identify classes among clusters, representative introns were determined first for each cluster as those with the smallest Euclidian distance to the cluster median. Exon/intron ratios of the representatives were then normalized to the exon/intron ratios in 5 min 4sU-RNA. Cluster representatives were again clustered 100 times with k-means for values of k between 3 and 10. For each k, the most frequent clustering was determined. For the final clustering, the k was chosen that maximized this frequency (in this case, k=4 with a maximal frequency of ~90%).

Statistical tests

Significance of differences between two distributions was assessed using the Wilcoxon rank sum test. This test is based on the ranks of observations and, thus, does not require any assumptions on the type of distribution (non-parametric test). Significance of enrichment was evaluated using Fisher's exact test which tests for a significant association between two different types of classification. Correction for multiple testing was performed using the method by Benjamini and Hochberg (Benjamini and Hochberg 1995) for control of the false discovery rate (FDR).

Data Access

Sequencing data including raw sequencing reads have been submitted to the Gene Expression Omnibus (GEO) under series ID GSE31653.

Acknowledgement

We thank Bernd Rädle and Silvia Weide for excellent technical assistance. This work was funded by NGFN-plus #01GS0801 grants to L.D. and R.Z., LMUexcellent to L.D, and DFG grant FR2938/1-1 to C.C.F. D.E. was supported by DFG grants SFB684 and SFB/Transregio5, K.B. by a stipend of the José Carreras Leukämie Stiftung e.V. The sequencing facility in Kiel is supported by the Clusters of Excellence Future Ocean and Inflammation at Interfaces and NGFN-Plus Grant 'Systematic genomics of chronic inflammation'. SOLiD II sequencing was performed in a research and development cooperation with Applied Biosystems.

Figure Legends

Figure 1: Ultra-short and progressive 4sU-Seq reveals the kinetics of RNA splicing.

(A) Contribution of newly transcribed RNA to total RNA levels following different durations of 4sU-exposure. The relative contributions of purified newly transcribed RNA (4sU-tagged RNA) to input RNA (total RNA) measured spectrophotometrically (OD260) are shown (combined data of three biological replicates). Standard error is indicated by error bars. (B) Distribution of the number of reads mapped to exons, exon-exon junctions, exon-intron junctions and intron regions for 5-60 min 4sU-RNA, total and untagged RNA. Samples are ordered according to increasing age of RNA in these samples from 5 to 60 min 4sU-tagging to total RNA and finally untagged RNA. RNA in the untagged RNA samples is at least 60 min old. This visualizes the maturation of transcripts over time. The expected distribution of reads for completely unspliced RNA is shown in the left-most column (see Supplementary Methods). (C) Normalized read frequencies were calculated by first dividing read numbers by the total number of reads on protein-coding genes in the corresponding sample. Subsequently, frequencies for a specific read type were divided by the maximum frequency observed for the corresponding read type in any sample. Despite the overall smaller number of exon-exon

junction and exon-intron junction reads, their normalized read frequencies are similar to normalized frequencies of exon and intron reads, respectively. (D) Introns were binned into 10 equally-sized groups according to intron length. For each group, average ratios of intron read counts in 60 min compared to 5 min 4sU-RNA were calculated. Results are shown both for all introns of the 525 most highly expressed genes (excluding only those with 0 reads in 5 min 4sU-RNA) and introns present in 5 min 4sU-RNA (RPKM ≥ 0.5). Ratios were highest for introns in a range of 300-400 nt and decreased both for longer and shorter introns. This was observed independently of gene length (see Supplementary Figure 1 E). (E) After normalizing the number of intronic reads to intron length and exonic reads to exon length, the delay between the decrease in intronic reads and exon-intron junction reads, on the one hand, and the increase in exonic reads and exon-exon junction reads, on the other hand, disappeared. For this analysis only the 525 most highly expressed genes were considered. After normalization to intron and exon length, respectively, the same normalization was applied as in (C).

Figure 2: Characterization of distinct intron splicing kinetics.

(A) Decay of an alternative splicing isoform of the *RPL23A* gene. Read densities for 5-60 min 4sU-RNA as well as total and untagged RNA are shown in shades of grey. This figure and all others showing read densities were created using the UCSC genome browser (Kent et al. 2002). The corresponding UCSC genome browser session containing read density values for all 525 genes analyzed can be accessed via http://www.bio.ifi.lmu.de/en/4sU-Seq. Exons are indicated by boxes, introns by lines. Regions that are part of non-coding transcripts are indicated by boxes of smaller height. The known retained intron not translated is additionally indicated in the last line in black. The intron marked in grey is absent in all 4sU-RNA samples (RPKM<0.5), which is indicative of very fast splicing. In total and untagged RNA, the retained intron is also absent, indicating either nonsense-mediated decay of the alternative splicing product or a secondary splicing event. (B) Higher level clustering of the intron

clusters identified in the first clustering step (see also Supplementary Figure 2). Clustering was performed based on intron/gene ratios of cluster representatives normalized to 5 min 4sU-RNA levels. Four classes of intron processing were identified: Class 1 constitutes "normal" decay with a linear slope in the first 20 min; Class 2 shows reduced decay rates; Class 3 shows increased decay rates; Class 4 represents retained introns that are stable during the first 60 min but are eventually spliced and degraded. (C) Read density in all samples for an exemplary Class 1 intron. (D) Read density for an exemplary Class 4 intron. (E) Exon/gene ratios in all samples for core exons, exons constituting retained introns, as well as alternative 3' and 5' ends. Again, exon/gene ratios were normalized to ratios in 5 min 4sU-RNA.

Figure 3: Properties of intron classes and snoRNA processing.

(A) Log_2 fold-changes of the median in each class compared to the median of all other classes are shown for intron length (white), gene length (grey) and splice site scores (black) calculated using MaxEntScan (Yeo and Burge 2004). An asterisk indicates a statistical significant difference compared to all other classes (Wilcoxon test, FDR corrected p-value ≤ 0.05). With the exception of splice site scores in Class 3, distributions showed significant differences for all classes and all features analyzed. (B) Intron/gene ratios are shown separately for snoRNA precursor introns (white) and all other introns (grey). Intron/gene ratios for precursor introns were significantly increased compared to all other introns at all times. (C) Read density plot illustrating the processing of an exemplary precursor intron to the mature snoRNA. One can observe both the substantially slower decay of the intronic sequences surrounding the mature snoRNAs, which was typical for intronic snoRNAs, and the characteristic 5' read start pattern of mature snoRNAs in total and untagged RNA.

Figure 4: Poor processing efficiency of many, but not all human snoRNAs.

(A) Distribution of RNA half-lives based on the DG75 microarray data for snoRNAs (green) compared to all other genes (dashed). (B) Comparison of the transcript's uracil count against

the $\log_2$ (newly transcribed RNA / total RNA ratios) allows the identification of reduced capture rates for short, i.e. uracil-poor, transcripts. The colors indicate the local density of points (yellow=high density, blue=low density). For genes with several transcripts, uracil number was calculated as the average of all transcripts. If no bias in capture rate were observed, log-ratios should average around zero (grey line), indicative of no association of RNA half-life with uracil content. Local regression (black line) suggests a slight bias in this example reflecting reduced capture efficiency of small transcripts as seen in previous studies. For snoRNAs (green), log ratios are considerable higher than expected for their short length. (C) Northern blot data for 15 snoRNAs after transcriptional inhibition by treatment with either 5µM Actinomycin D (Act-D) or 800nM Flavopiridol (FL) for 6 h. 28S EtBr served as loading control. For almost all snoRNAs, expression levels after transcriptional inhibition remained unaltered indicative of their high stability. (D) Distribution of RNA half-lives based on the DG75 microarray data for snoRNAs with different expression strategies: intronic (nc) = snoRNA expressed as part of an intron of a non-coding gene; intronic (pc) = snoRNA expressed as a part of an intron of a protein-coding gene; independent = snoRNA with independent promoter; *SNORD116* = *SNORD116* cluster snoRNAs.

Figure 5: Processing of the SNORD116 cluster.

(A) Read density for the *SNORD116* cluster region (overview). (B) Read density plot for a typical snoRNA of the *SNORD116* cluster only shows reads at the 5'-end of the mature snoRNA, but no intronic sequences, thus representing sequencing of the mature, unfragmented snoRNA. (C) Read density plot for a typical intronic snoRNA. Here, intronic sequences were observed until 60 min 4sU-RNA.

References

Benjamini Y and Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**: 289-300.

Cavaille J, Buiting K, Kiefmann M, Lalande M, Brannan CI, Horsthemke B, Bachellerie JP, Brosius J and Huttenhofer A. 2000. Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proc Natl Acad Sci U S A* **97**: 14311-14316.

Cazalla D, Yario T and Steitz JA. 2010. Down-regulation of a host microRNA by a Herpesvirus saimiri noncoding RNA. *Science* **328**: 1563-1566.

Chomczynski P and Mackey K. 1995. Short technical reports. Modification of the TRI reagent procedure for isolation of RNA from polysaccharide- and proteoglycan-rich sources. *Biotechniques* **19**: 942-945.

Cleary MD, Meiering CD, Jan E, Guymon R and Boothroyd JC. 2005. Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. *Nat Biotechnol* **23**: 232-237.

Clement JQ, Qian L, Kaplinsky N and Wilkinson MF. 1999. The stability and fate of a spliced intron from vertebrate cells. *RNA* **5**: 206-220.

Dieci G, Preti M and Montanini B. 2009. Eukaryotic snoRNAs: a paradigm for gene expression flexibility. *Genomics* **94**: 83-88.

Dölken L et al. 2010. Systematic analysis of viral and cellular microRNA targets in cells latently infected with human gamma-herpesviruses by RISC immunoprecipitation assay. *Cell Host Microbe* **7**: 324-334.

Dölken L et al. 2008. High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA* **14**: 1959-1972.

Dye MJ, Gromak N and Proudfoot NJ. 2006. Exon tethering in transcription by RNA polymerase II. *Mol Cell* **21**: 849-859.

Friedel CC and Dölken L. 2009. Metabolic tagging and purification of nascent RNA: implications for transcriptomics. *Mol Biosyst* **5**: 1271-1278.

Friedel CC, Dölken L, Ruzsics Z, Koszinowski UH and Zimmer R. 2009. Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. *Nucleic Acids Res* **37**: e115.

Hafner M et al. 2010. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**: 129-141.

Hirose T, Shu MD and Steitz JA. 2003. Splicing-dependent and -independent modes of assembly for intron-encoded box C/D snoRNPs in mammalian cells. *Mol Cell* **12**: 113-123.

Jing Q, Huang S, Guth S, Zarubin T, Motoyama A, Chen J, Di Padova F, Lin SC, Gram H and Han J. 2005. Involvement of microRNA in AU-rich element-mediated mRNA instability. *Cell* **120**: 623-634.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM and Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12**: 996-1006.

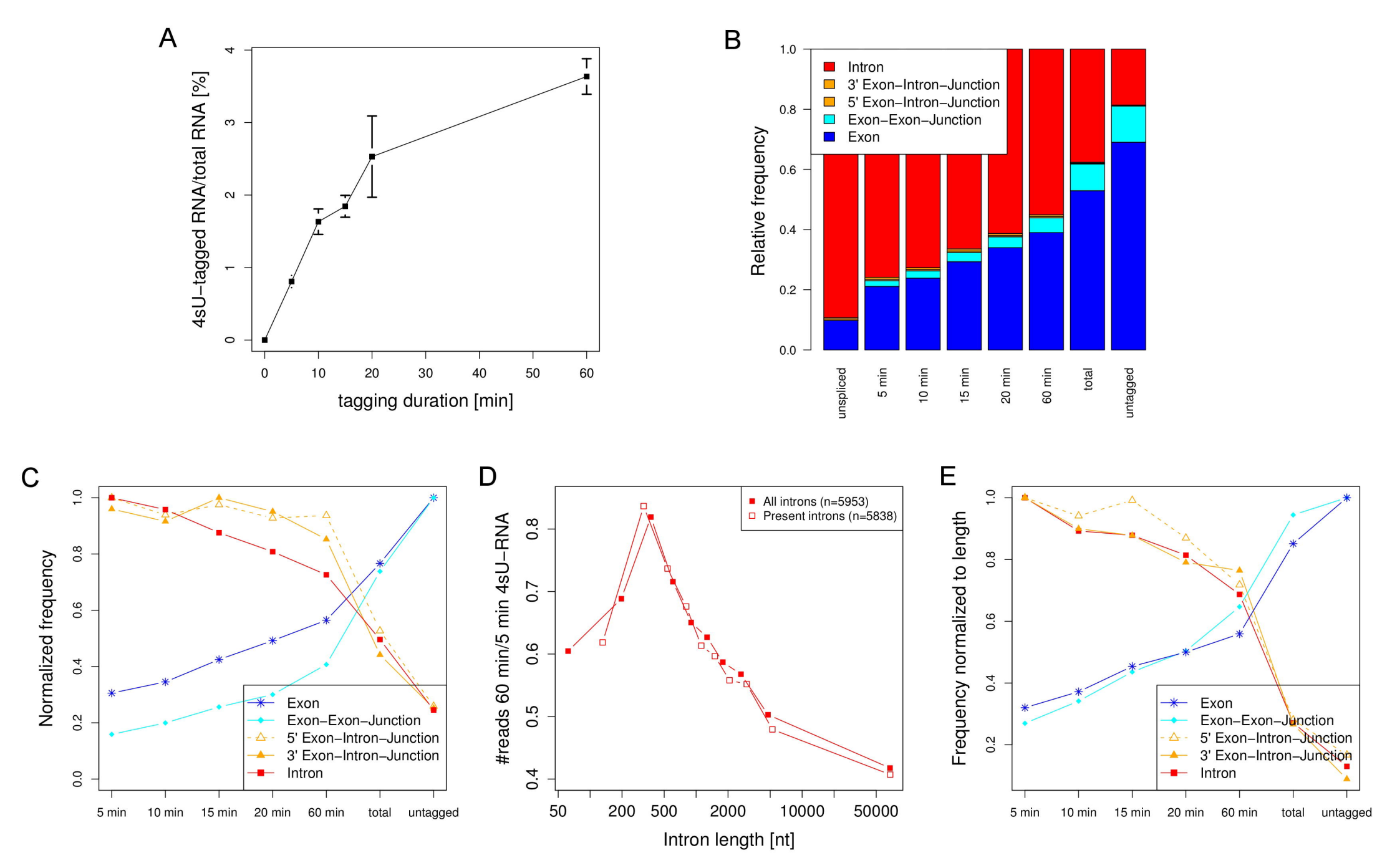
Kenzelmann M et al. 2007. Microarray analysis of newly synthesized RNA in cells and animals. *Proc Natl Acad Sci U S A* **104**: 6164-6169.

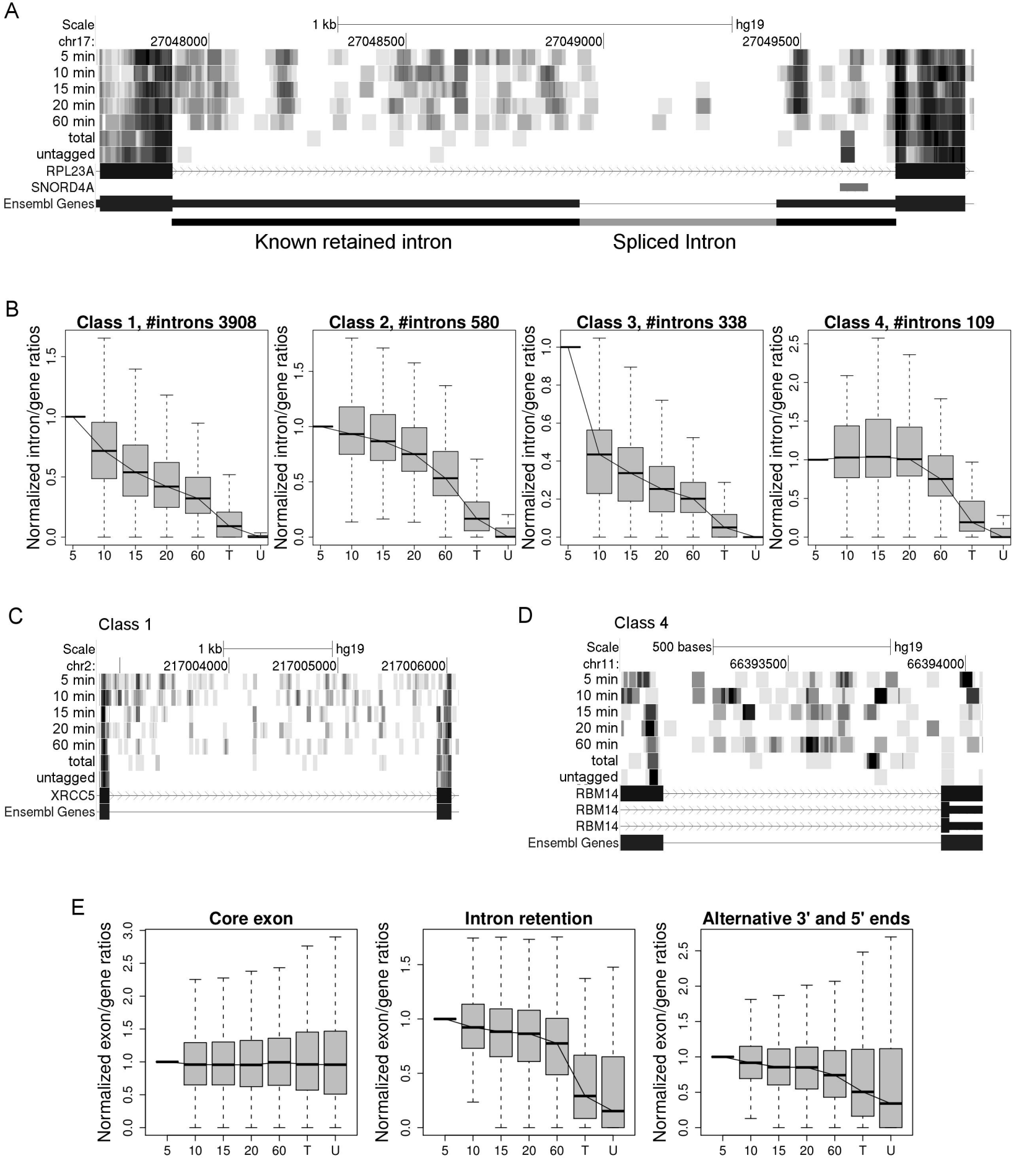
Kim HD, Shay T, O'Shea EK and Regev A. 2009. Transcriptional regulatory circuits: predicting numbers from alphabets. *Science* **325**: 429-432.

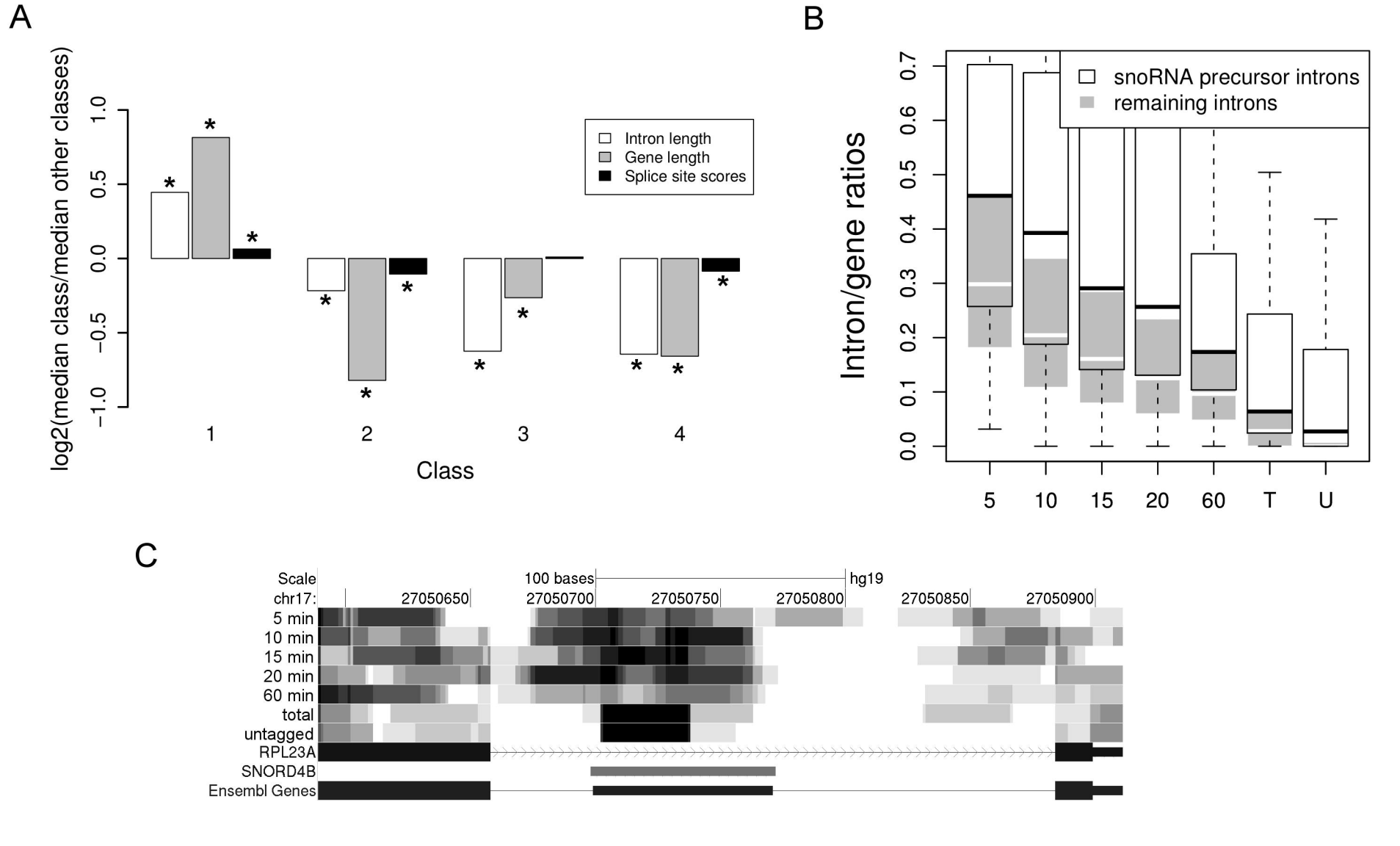
Kiss T. 2002. Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell* **109**: 145-148.

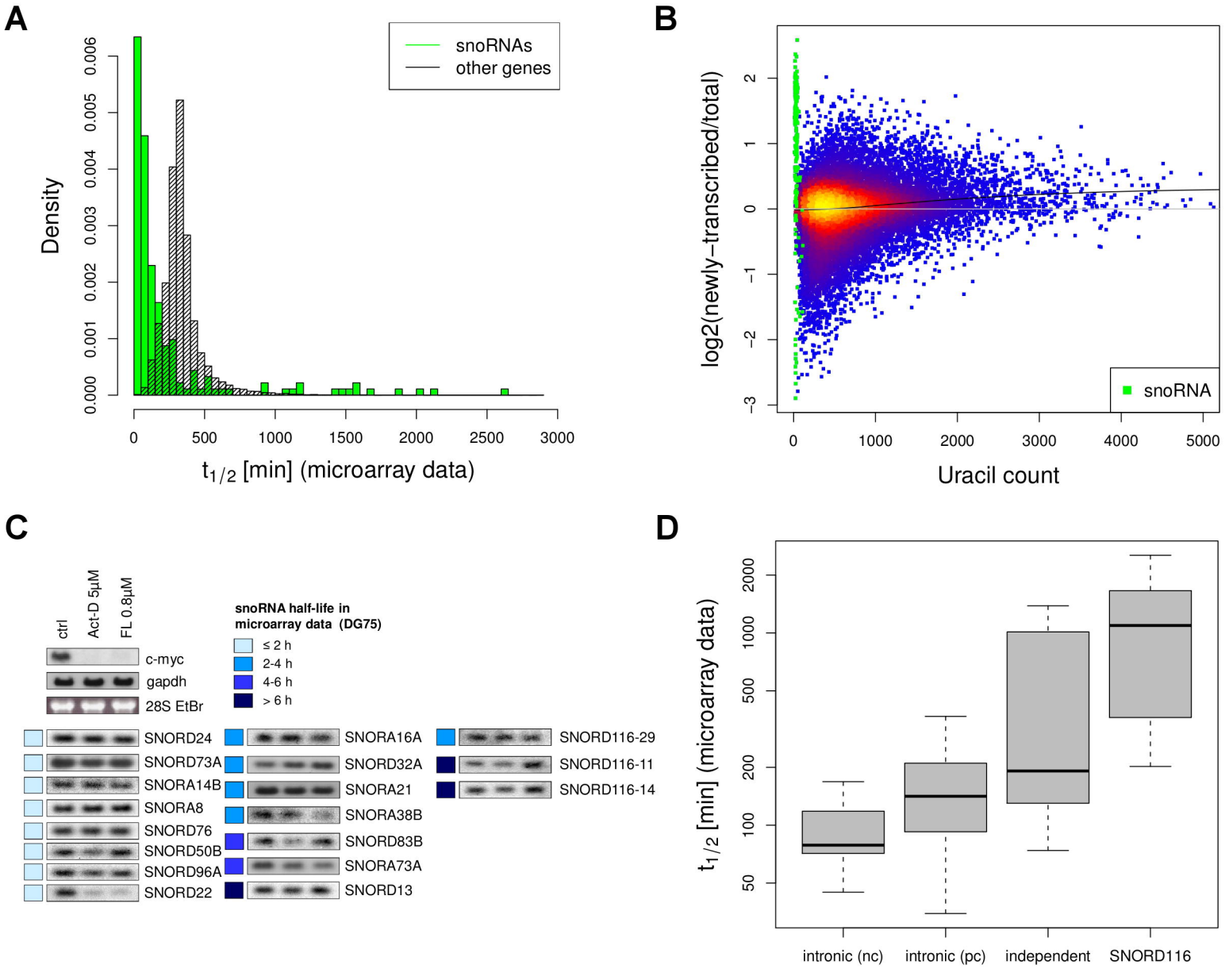
Lamond AI, Konarska MM, Grabowski PJ and Sharp PA. 1988. Spliceosome assembly involves the binding and release of U4 small nuclear ribonucleoprotein. *Proc Natl Acad Sci U S A* **85**: 411-415.

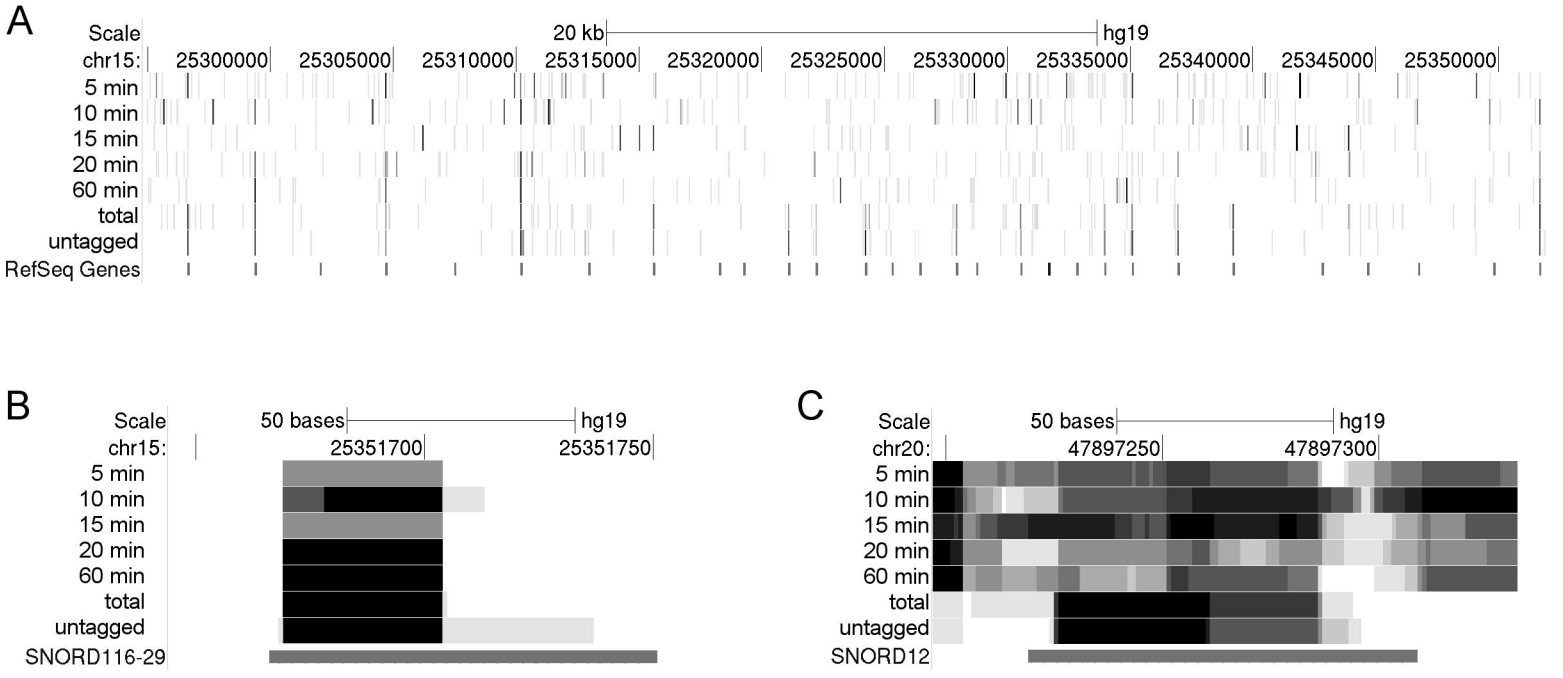
Langmead B, Trapnell C, Pop M and Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.
Melvin WT, Milne HB, Slater AA, Allen HJ and Keir HM. 1978. Incorporation of 6-thioguanosine and 4-thiouridine into RNA. Application to isolation of newly synthesised RNA by affinity chromatography. *Eur J Biochem* **92**: 373-379.
Miller C et al. 2011. Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol Syst Biol* **7**: 458.
Miller MR, Robinson KJ, Cleary MD and Doe CQ. 2009. TU-tagging: cell type-specific RNA isolation from intact complex tissues. *Nat Methods* **6**: 439-441.
Mortazavi A, Williams BA, McCue K, Schaeffer L and Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**: 621-628.
Nam K, Lee G, Trambley J, Devine SE and Boeke JD. 1997. Severe growth defect in a Schizosaccharomyces pombe mutant defective in intron lariat degradation. *Mol Cell Biol* **17**: 809-818.
Nilsen TW and Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**: 457-463.
Pan Q, Shai O, Lee LJ, Frey BJ and Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**: 1413-1415.
Rabani M et al. 2011. Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat Biotechnol* **29**: 436-442.
Royo H and Cavaille J. 2008. Non-coding RNAs in imprinted gene clusters. *Biol Cell* **100**: 149-166.
Runte M, Huttenhofer A, Gross S, Kiefmann M, Horsthemke B and Buiting K. 2001. The IC-SNURF-SNRPN transcript serves as a host for multiple small nucleolar RNA species and as an antisense RNA for UBE3A. *Hum Mol Genet* **10**: 2687-2700.
Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W and Selbach M. 2011. Global quantification of mammalian gene expression control. *Nature* **473**: 337-342.
Shalem O, Dahan O, Levo M, Martinez MR, Furman I, Segal E and Pilpel Y. 2008. Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. *Mol Syst Biol* **4**: 223.
Sims RJ, 3rd, Rojas LA, Beck D, Bonasio R, Schuller R, Drury WJ, 3rd, Eick D and Reinberg D. 2011. The C-terminal domain of RNA polymerase II is modified by site-specific methylation. *Science* **332**: 99-103.
Smit AFA, Hubley R and Green P. 1996-2010. RepeatMasker Open-3.0.
Ule J, Jensen KB, Ruggiu M, Mele A, Ule A and Darnell RB. 2003. CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302**: 1212-1215.
Ule J, Stefani G, Mele A, Ruggiu M, Wang X, Taneri B, Gaasterland T, Blencowe BJ and Darnell RB. 2006. An RNA map predicting Nova-dependent splicing regulation. *Nature* **444**: 580-586.
Vitali P, Royo H, Marty V, Bortolin-Cavaille ML and Cavaille J. 2010. Long nuclear-retained non-coding RNAs and allele-specific higher-order chromatin organization at imprinted snoRNA gene arrays. *J Cell Sci* **123**: 70-83.
Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP and Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470-476.
Weintz G et al. 2010. The phosphoproteome of toll-like receptor-activated macrophages. *Mol Syst Biol* **6**: 371.
Yeo G and Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**: 377-394.


A
4sU−tagged RNA/total RNA [%]
tagging duration [min]
B
Relative frequency
Intron
3' Exon−Intron−Junction
5' Exon−Intron−Junction
Exon−Exon−Junction
Exon
unspliced
5 min
10 min
15 min
20 min
60 min
total
untagged
C
Normalized frequency
Exon
Exon−Exon−Junction
5' Exon−Intron−Junction
3' Exon−Intron−Junction
Intron
D
#reads 60 min/5 min 4sU−RNA
All introns (n=5953)
Present introns (n=5838)
Intron length [nt]
E
Frequency normalized to length


A
Scale
1 kb
hg19
chr17:
27048000
27048500
27049000
27049500
5 min
10 min
15 min
20 min
60 min
total
untagged
RPL23A
SNORD4A
Ensembl Genes
Known retained intron
Spliced Intron
B
Class 1, #introns 3908
Class 2, #introns 580
Class 3, #introns 338
Class 4, #introns 109
Normalized intron/gene ratios
5
10
15
20
60
T
U
C
Class 1
Scale
1 kb
hg19
chr2:
217004000
217005000
217006000
5 min
10 min
15 min
20 min
60 min
total
untagged
XRCC5
Ensembl Genes
D
Class 4
Scale
500 bases
hg19
chr11:
66393500
66394000
5 min
10 min
15 min
20 min
60 min
total
untagged
RBM14
RBM14
RBM14
Ensembl Genes
E
Core exon
Intron retention
Alternative 3' and 5' ends
Normalized exon/gene ratios
5
10
15
20
60
T
U


A
log2(median class/median other classes)
1.0
0.5
0.0
−0.5
−1.0
Intron length
Gene length
Splice site scores
1
2
3
4
Class
B
Intron/gene ratios
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
snoRNA precursor introns
remaining introns
5
10
15
20
60
T
U
C
Scale
100 bases
hg19
chr17:
27050650
27050700
27050750
27050800
27050850
27050900
5 min
10 min
15 min
20 min
60 min
total
untagged
RPL23A
SNORD4B
Ensembl Genes


A
snoRNAs
other genes
Density
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0
500
1000
1500
2000
2500
3000
t1/2 [min] (microarray data)
B
log2(newly−transcribed/total)
2
1
0
−1
−2
−3
snoRNA
0
1000
2000
3000
4000
5000
Uracil count
C
ctrl
Act-D 5μM
FL 0.8μM
c-myc
gapdh
28S EtBr
snoRNA half-life in
microarray data (DG75)
≤ 2 h
2-4 h
4-6 h
> 6 h
SNORD24
SNORD73A
SNORA14B
SNORA8
SNORD76
SNORD50B
SNORD96A
SNORD22
SNORA16A
SNORD32A
SNORA21
SNORA38B
SNORD83B
SNORA73A
SNORD13
SNORD116-29
SNORD116-11
SNORD116-14
D
t1/2 [min] (microarray data)
50
100
200
500
1000
2000
intronic (nc)
intronic (pc)
independent
SNORD116


A
Scale
20 kb
hg19
chr15:
25300000
25305000
25310000
25315000
25320000
25325000
25330000
25335000
25340000
25345000
25350000
5 min
10 min
15 min
20 min
60 min
total
untagged
RefSeq Genes
B
Scale
50 bases
hg19
chr15:
25351700
25351750
5 min
10 min
15 min
20 min
60 min
total
untagged
SNORD116-29
C
Scale
50 bases
hg19
chr20:
47897250
47897300
5 min
10 min
15 min
20 min
60 min
total
untagged
SNORD12