CHAPTER 5

FULL-LENGTH mRNA SEQUENCING UNCOVERS A WIDESPREAD COUPLING BETWEEN TRANSCRIPTION AND mRNA PROCESSING

Seyed Yahya Anvar, Eleonora de Klerk,
Martijn Vermaat, Johan T. den Dunnen, Stephen W. Turner,
Peter A.C. 't Hoen.

Manuscript Submitted. 2015.

ABSTRACT

Deciphering the interdependency between transcriptional and posttranscriptional regulatory events acting on the same RNA molecule is key in understanding the regulation of gene expression. The analysis of 7.4 million single-molecule long sequencing reads representing full-length mRNA molecules in MCF-7 human breast cancer cells provides the first comprehensive view of the degree of coordination between alternative transcription initiation, splicing and polyadenylation. In MCF-7 cells, an unforeseen amount of genes undergo vigorous and interdependent preferential selection during transcription and mRNA processing, which occur across the entire mRNA molecules. In particular, alternative polyadenylation sites that are coupled with alternative splicing events are depleted for known polyadenylation signals and enriched for MBNL binding motifs, supporting a dual role of MBNL proteins in regulating splicing and polyadenylation.

Our findings demonstrates that our understanding of transcriptome complexity is far from complete and provides a framework to reveal largely unresolved mechanisms that coordinate transcription and mRNA processing.

INTRODUCTION

The formation of a mature messenger RNA (mRNA) is a multi-step process. In higher eukaryotes, variations in each of these steps, e.g. selection of alternative transcription initiation, alternative exons, and alternative polyadenylation site, change the nature of the mature transcript. Tight regulation and coordination of these processes ensures the production of a set of cell-, tissue- and condition-specific transcript variants to meet variable cellular protein requirements (1-4). The co-transcriptional nature of mRNA processing suggests the presence of yet largely unresolved mechanisms that couple transcription with 5' end capping, splicing, and 3' end formation (reviewed in 5). Thus, resolving full transcript structures and accurate quantification of the abundance of alternative transcripts are important steps towards the delineation of these mechanisms.

RNA sequencing (RNA-Seq) has become a central technology for deciphering the global RNA expression patterns. However, reconstruction and expression level estimation of alternative transcripts using standard RNA-Seq experiments is limited and prone to error due to relatively short read length (typically up to 150 nucleotides) and required amplification steps of second-generation sequencing technologies (6, 7). It is apparent that single-molecule long reads that capture the entire RNA molecule can offer a better understanding of the rich patterns of alternative transcription and mRNA processing events and gene expression in human transcriptome and, hence, the underlying biology.

Despite a number of studies that have pursued long read sequencing to connect different exons or even capture entire transcripts with a rather limited sequencing depth (6, 8-14), the coupling between transcription and mRNA processing has not been extensively studied. Here, we investigate the global pattern of coupling between transcription, splicing and polyadenylation in MCF-7 human breast cancer cell line, which is deeply sequenced using the single-molecule real-time Pacific Biosciences RSII sequencing platform.

We show that transcription and mRNA processing are tightly coupled and that such interdependencies can be found across the entire RNA molecule and across large intra-molecular distances. We demonstrate that transcript identification and understanding of coupling between processes that are involved in the formation of these transcripts is far from complete, even in well-characterized human cell lines such as MCF-7. This study provides an in-depth view of the true complexity of the transcriptome and, for the first time, shows the tight and global interdependency between alternative transcription, splicing and polyadenylation.

RESULTS

Detection of transcript variants and the associated interdependencies between alternative exons

To investigate the genome-wide coupling of transcription and mRNA processing events, full-length mRNAs from MCF-7 human breast cancer cells were sequenced on 119 SMRT cells using Pacific Biosciences RSII platform (Supplementary Table 1). Prior to sequencing, parts of the sequencing library were size selected to allow for capturing rare and longer transcripts. The sequencing depth of our data, consisting of 7.4 million long reads, is equivalent to 70.3 million Illumina paired-end reads. Thus, this data enables reliable quantification of transcript abundance in MCF-7 transcriptome.

Transcript structures were defined by applying the isoform-level clustering algorithm (ICE) on full-length reads, capturing the entire mRNA molecule (containing both 5' and 3' primer sequences). Transcript sequences were further polished using both full-length and partial reads (**Figure 1A**). Our

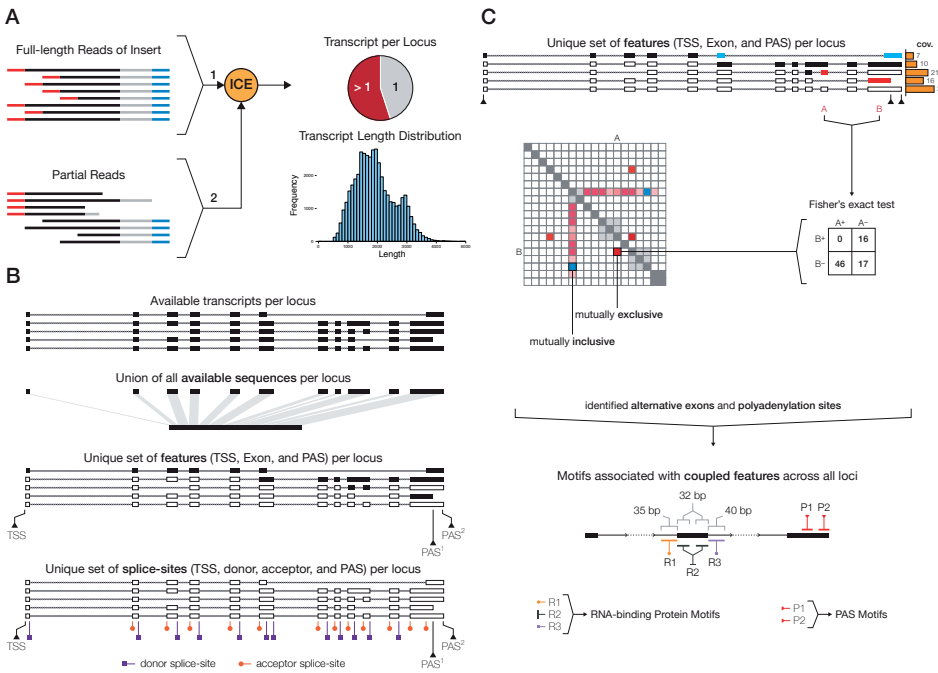

Figure 1. Schematic overview of the approach to characterize the interdependencies between mRNA transcription and processing events. A) Identified full-length reads of inserts are clustered into different transcript structures using the ICE algorithm and further polished using the partial reads. The number of unique transcript structures per locus and the distribution of transcript lengths are assessed. **B)** Based on the available transcripts per locus, the available sequence and unique set of features and splice-sites are identified. The available sequence is the union of all exonic sequences that are observed at each locus. Features are defined as a unique set of transcription start-sites (TSS), alternative exons, and alternative polyadenylation sites (PAS). The unique set of splice sites consists of unique donor and acceptor splice-sites as well as all alternative TSSs and PASs. **C)** The survey of coupling events is done by performing all possible pairwise tests between unique features of expressed and detected genes. The sum of the coverage of all transcripts that support the inclusion or exclusion of each pair is used in a contingency table to perform a Fisher's exact test for statistical significance. The odd ratio (OR) is used to differentiate between mutually inclusive (positive log-transformed OR) and exclusive (negative log-transformed OR) coupling events. Next, for all alternative exons that show significant linkage, a motif search is performed to assess the enrichment of specific RNA-binding protein motifs. For all alternative exons, the 35bp intronic sequences upstream of the acceptor site are defined as R1 domain (depicted in orange), the 32bp exonic sequences downstream of the acceptor site and upstream of the donor site are defined as R2 domain (depicted in dark grey), and 40bp intronic sequences downstream of the donor site are defined as R3 domain (depicted in purple). The 35bp sequence upstream of each PAS (depicted in red) is searched for the presence of canonical and non-canonical poly(A) signals.

analysis pipeline could precisely determine the position of polyadenylation sites (presence of poly(A) tail in the sequence) and intron-exon boundaries, as evident from the presence of the canonical GU motif in 94.9% of donor splice sites and the canonical AG motif in 96.6% of acceptor splice sites. From the 14,385 genes with detectable expression, 49% produced multiple transcript structures (Supplementary Figure 1). A total of 93 candidate fusion genes were identified based on the inter-chromosomal or distant intra-chromosomal split-alignment of transcripts to the human reference genome (Supplementary Table 2). In addition, 42% of identified transcripts in MCF-7 are potentially novel in comparison with the GENCODE annotation (Supplementary Table 3).

To detect and characterize the dependency between transcription and mRNA processing events, we designed the following analysis strategy (**Figure 1**). For each gene, the union of all exonic sequences

was considered as the available sequence and the union of all unique transcription start sites (TSSs), exons (defined as having distinct donor and acceptor splice sites), and polyadenylation sites (PASs) was used as a set of available features (**Figure 1B**). Mutual inclusivity or exclusivity of all possible pairs of features was assessed based on the number of reads that support the inclusion or exclusion of each pair of features. Subsequently, we applied a Fisher's exact test to evaluate statistical significance of the interdependency between a pair of features (**Figure 1C**; also see Methods).

General properties of coupling in human MCF-7 transcriptome

The MCF-7 transcriptome data consist of 14,385 genes containing 1,724,400 combinations of features (TSSs, exons, and PASs). The majority of combinations represent exon-exon pairs as many loci contain only a single TSS or PAS whereas most loci are multi-exonic (Supplementary Figure 2). Since the test is only applicable to genes with multiple transcripts, only 7,008 genes and 1,090,077 pairs of features (TSSs, exons, and PASs) were included in the statistical evaluation. Twenty percent of all feature pairs were significantly coupled (p-value < 4.6e-08, after Bonferroni correction for multiple testing). Generally, we observed large effect sizes for coupled features with the majority (65%) to be mutually inclusive, meaning that features were predominantly present in the same transcripts (Supplementary Figure 3,4). Remarkably, we observed coupling between mRNA features in nearly half of all genes that were evaluated (3,426 out of 7,008; **Figure 2A**; Supplementary Figure 5). We found a substantial amount of interdependencies between all types of features (**Figure 2B**). Of the 3,426 genes with at least one coupling event, 1,212 (35%) showed interdependencies between all classes of features: alternative TSS linked to alternative exons, alternative exon to alternative exon linkage, alternative PAS linked to alternative exons, and alternative TSSs to alternative PASs. Thus, the deep sequencing of full-length mRNAs provided a first image of the large degree of coordination in the usage of alternative TSSs, exons and PASs, mostly restricting the number of produced transcripts given the substantial amount of combinatorial possibilities.

Only 18% of the significantly coupled transcription and mRNA processing features were cataloged in the Ensembl Alternative Splicing Events set, version 75 (**Figure 2C**). These features were almost uniformly distributed across the different categories of alternative transcription or mRNA processing events. The majority of features (75%) that could not be attributed to any of the known categories represented interdependencies between alternatively spliced exons.

The length of individual transcripts was not associated with the likelihood of a significant coupling event in that transcript (Supplementary Figure 6). However, significant coupling events were enriched in genes with larger exonic sequence lengths (**Figure 2D,E**), giving rise to a larger repertoire of possible transcripts and requiring more extensive regulation of the synthesis for transcripts containing different subset of features.

We also examined the effect of the relative position in the gene and the distance between features on the observed degree of coupling. As expected, most TSSs were located at the most 5'-end of genes. The TSSs coupled or not coupled to alternative mRNA processing events showed a similar distribution over the gene (**Figure 3A** and Supplementary Figure 7). Interdependence between alternative TSSs was observed across the entire gene (**Figure 3B;** Supplementary Figure 8). However, alternative TSSs were preferentially coupled to alternative splicing events in relatively close proximity to the TSSs, near the 5'-end **(Figure 3B**). Nevertheless, examples of the coupling of alternative TSS and alternative exon usage across large distances, and spanning multiple exons were also frequently observed (**Figure 3C;** ITGB4). More evidence for interactions across the entire length of genes comes from the significant coupling between TSS and PAS (**Figure 3B,C**; NCAPD2).

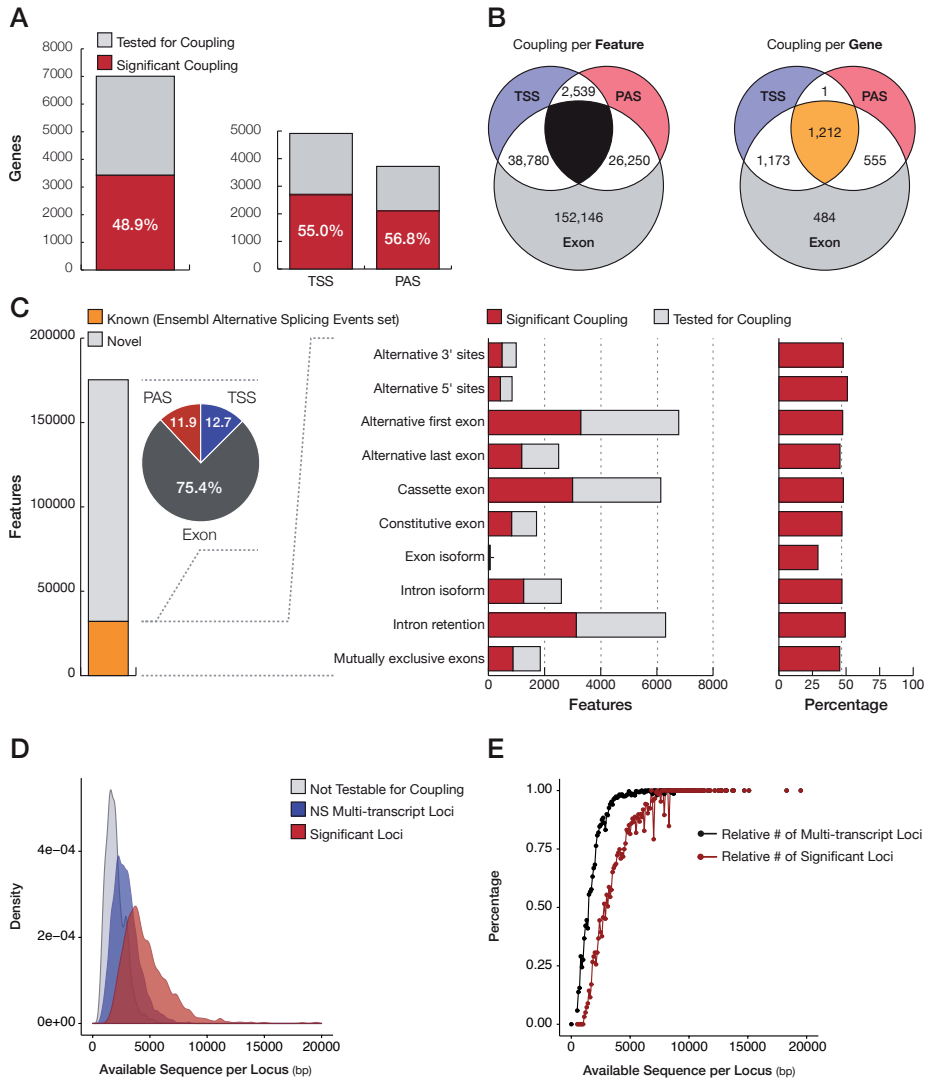

Figure 2. Alternative transcription, splicing, and polyadenylation is highly interdependent. A) The bar and pie charts illustrate the number and proportion of genes that show significant coupling. **B)** Venn diagrams show the number of coupled features based on the type of processes that are tested as well as the number of genes that show significant coupling between different processes. **C)** The number and proportion of known alternative features (TSS, exon, PAS) that are located in genomic regions, associated with Ensembl annotated alternative splicing events. **D)** The distribution of the length of available sequences per locus for loci with only one transcript (not tested; grey), multi-transcript loci with no significant coupling (blue), and all loci that show at least one significant coupling event (red). **E)** The relative number of loci with multiple transcripts (black) and the relative number of multi-transcript loci with significant coupling were plotted against the length of the available sequence. 100bp bins were used to group examined loci by length.

Similarly, coupling events linked to alternative PAS usage were found across the entire gene (Supplementary Figure 8; **Figure 3D**). In concordance with published literature (15-17), alternative PAS usage was preferentially coupled to nearby alternative exons (**Figure 3E**). Nevertheless, a substantial

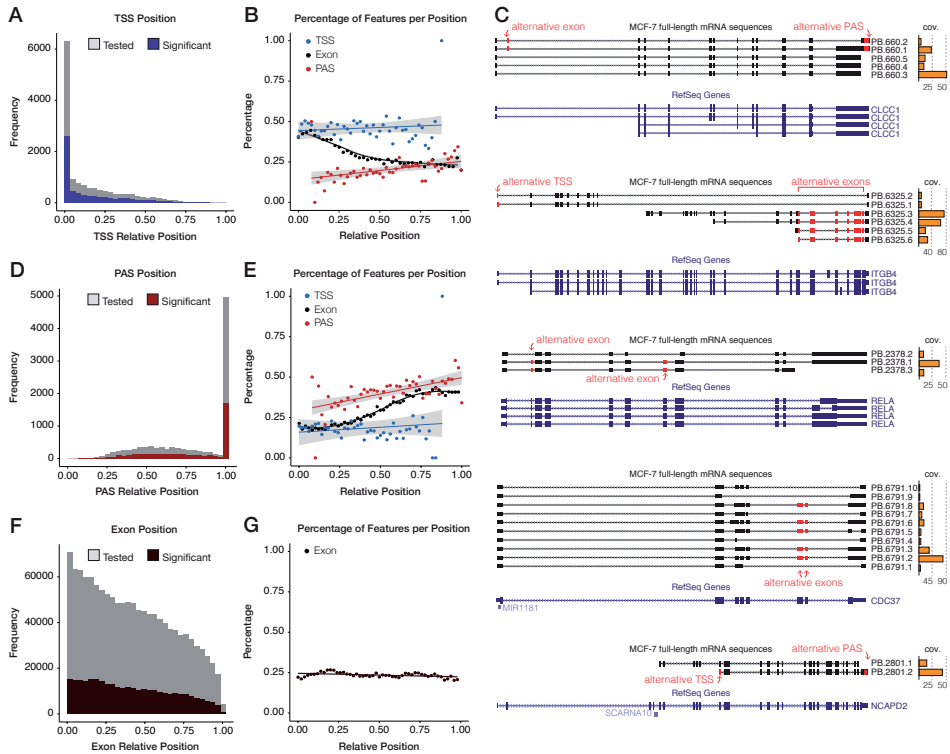

Figure 3. The interdependence of transcription and mRNA processing events can range across large distances. **A)** Histogram of the relative positions of transcription start sites (TSS) with (blue) and without (grey) significant coupling to mRNA processing events. Relative positions are calculated based on the length of the total exonic sequence at each locus. **B)** The fraction of significantly coupled TSSs (blue) to alternative exons (black) and PASs (red) is plotted at each relative position. **C)** Examples of genes that show evidence of coupling between transcription and mRNA processing events across the entire length of the gene. CLCC1 gene shows an example of long-range mutual inclusivity of the alternative second exon and the usage of the most distal PAS (depicted in red). ITGB4 gene presents an example of coupling between TSS and mutually exclusion of a cassette of alternative exons. RELA transcripts support the mutual inclusion of non-consecutive alternative exons. CDC37 gene provides evidence of mutual inclusion of consecutive exons during alternative splicing. NCAPD2 gene shows an example of coupling transcription with alternative polyadenylation. The number of supporting reads for each transcript is shown in orange bars. RefSeq annotated transcript structures are presented. **D)** Histogram of the relative positions of polyadenylation sites (PAS) with (red) and without (grey) significant coupling to alternative transcription and splicing events. **E)** The fraction of significantly coupled PASs (red) to alternative TSSs (blue) and exons (black) is plotted at each relative position. **F)** Histogram of the relative positions of alternative exons with (brown) and without (grey) significant coupling to other exons. **G)** The fraction of significantly coupled exons to other exons is plotted at each relative position. For plots depicting the percentage of linked features per position, the bin size of 0.02 was used.

proportion of PASs was coupled to alternative exons in more 5'-regions of genes (see CLCC1, **Figure 3C**). The proportion of alternative exons was higher at the 5'-end; however, dependencies between multiple alternative splicing events were uniformly observed across the entire gene (**Figure 3F, 3G**). In spite of this uniform distribution of exon-exon coupling events and the presence of distant coupling events (**Figure 3C**, RELA), the majority of independent alternative splicing events was between nearby or neighboring exons (Supplementary Figure 9; **Figure 3C,** CDC37).

We performed pathway enrichment analysis to analyze whether the coupling between alternative

transcription and mRNA processing events was associated with the molecular function of the proteins encoded by the transcripts. A number of pathways associated with mRNA processing and protein degradation such as spliceosome, proteasome, and ubiquitin mediated proteolysis, were enriched in transcripts demonstrating significant coupling (Supplementary Table 4).

Poly(A) signal usage for coupled polyadenylation sites

The majority of the alternative PASs in MCF-7 cells was found in different exons. From 3,719 genes that contain alternative PASs, we identified only 200 tandem PASs in the same 3' UTR. From these, only 56 loci (28%) included PASs that were significantly coupled with alternative exons. The low number of tandem 3' UTRs, in both coupled and uncoupled PAS-exon pairs (3.2% and 2.8%, respectively), has been previously explained by a general shortening of 3' UTRs in MCF-7 cell line (18). Thus, the majority of coupling events between alternative PASs and inclusion or exclusion of alternative exons are due to the use of exonic and intronic PASs, leading to the formation of new 3' UTRs.

To assess whether certain poly(A) signals are preferentially associated with alternative transcription and splicing, we searched for canonical (AATAAA and ATTAAA) and eleven known non-canonical poly(A) signals in the 35bp sequences upstream of the identified PASs. Canonical poly(A) signals could be found in the 35bp sequences upstream of 51.5% of all PASs (**Figure 4A**; Supplementary Figure 10, 11). This percentage is lower than what is generally reported and is most likely due to a global shortening of the 3' UTRs in MCF-7 cell line (18). Interestingly, the proportion of PASs that could be associated with canonical poly(A) signals was significantly lower (40.7%) for those that were coupled with TSSs or alternative exons. PASs that were linked with TSSs showed an even lower proportion of canonical poly(A) signals (34.7%). This decrease was not accompanied by an increase in known non-canonical poly(A) signals, but was mainly due to the use of PASs for which no known poly(A) signal could be identified (**Figure 4B**). This suggests that a novel poly(A) signal and alternative mechanisms may be involved in transcription- or splicing-coupled polyadenylation in MCF-7 cells. Thus, we screened for enriched motifs in the 35bp sequences upstream of PASs that are not associated with known poly(A) signals. While we did not observe any enrichment for PASs that were coupled with alternative TSSs, the ones that were coupled with alternative splicing were enriched for ASCCTG and GYGACA motifs. Interestingly, the ASCCTG motif could be associated with the binding site of muscleblind-like (MBNL) protein family, known to play a dual role in the regulation of splicing and polyadenylation (19, 20). Each MBNL isoform can bind to slightly different motifs (20) and a few motifs have been associated with MBNL proteins (20-22). Although all three MBNL proteins are expressed in MCF-7 cells, the enrichment of de novo identified ASCCTG and the recently reported CWGCMWKS motifs that can be recognized by MBNL3 protein20 were more prominent. Notably, previously identified binding motifs for MBNL1 (CTSCYB21 and RSCWTGSK20) and MBNL2 (TGCYTSYY20) were also enriched in sequences upstream of the PASs without a known poly(A) signal (**Table 1**). However, these motifs were not found to be associated with PASs that were coupled with alternative TSSs or alternative exons. Together, these results support an important role of MBNL proteins in the coupling between alternative polyadenylation and alternative splicing.

Identification of binding motifs for RNA-binding proteins potentially involved in coupling

We investigated the potential involvement of RNA-binding proteins (RBPs) in the coordination of alternative transcription and mRNA processing events by enrichment analysis of their binding motifs in

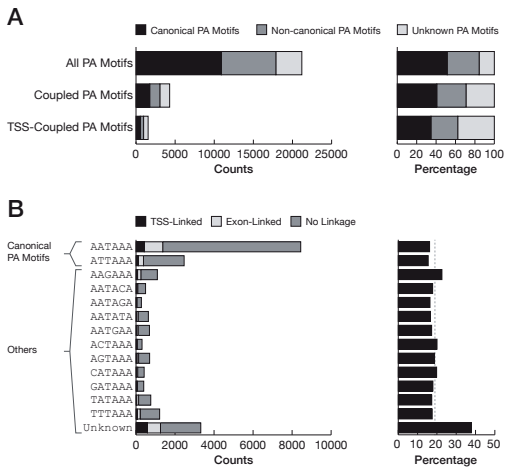

Figure 4. Alternative transcription start sites and exons are significantly associated with non-canonical poly(A) signals. A) Bar charts of the number and relative proportion of PASs that are associated with canonical or non-canonical poly(A) signals for all PASs, PAS with significant coupling, and TSS-linked PASs. **B)** The number and relative proportion of poly(A) signals for TSS-linked, exon-linked, or non-significantly coupled PASs.

Motif s	Source	Total	Random Set	P-value *	Coupled PAS	Not Coupled PAS	P-value §
ASCCTG	DREME	2232	102	0	710 (26.0%)	1522 (20.5%)	4.6E-09
CTSCYB	Masuda, 2012 [21]	970	650	5.8E-17	237 (8.7%)	733 (9.9%)	9.7E-01
YGCY	Purcell, 2012 [22]	2499	3485	1.0E-00	611 (22.3%)	1888 (25.5%)	9.9E-01
RSCWTGSK	Batra, 2014 [20]– MBNL1	195	99	9.1E-09	59 (2.2%)	136 (1.8%)	1.7E-01
TGCYTSYY	Batra, 2014 [20]– MBNL2	92	51	3.7E-04	16 (0.6%)	76 (1.0%)	9.9E-01
CWGCMWKS	Batra, 2014 [20]– MBNL3	1894	129	0	590 (21.6%)	1304 (17.6%)	3.9E-06
	Total PASs	10146	10146	–	2736	7410	–

* The enrichment of binding motifs in sequences upstream of PASs without a known poly(A) signal were calculated by Fisher's exact test (one-sided). A randomly generated set was used as a background for enrichment analysis.

§ PASs without significant coupling were used as the background set to identify a binding site that is enriched in the coupled PAS s without a known poly(A) signal.

Table 1. Enrichment of MBNL binding site motifs in sequences upstream of alternative PAS with unknown poly(A) signal that are coupled with alternative TSS or alternative exons.

coupled versus non-coupled exons. We screened three genomic regions relative to donor and acceptor splice-sites of coupled exons for enriched sequence motifs (**Figure 1C**; also see Methods): the 35bp intronic sequences upstream of the acceptor site (R1), the 32bp exonic sequences downstream of the acceptor sites and upstream of the donor sites (R2), and the 40bp intronic sequences downstream of the donor sites (R3).

For exons linked to alternative TSSs, the sequences from the R1 and R3 domains (upstream of the acceptor and downstream of the donor splice-sites, respectively) were both enriched for motifs (**Table 2**) that can be recognized by the splicing modulator RBM14 protein, known to play a dual role in regulating transcription and splicing (23, 24). In addition, the R2 sequences were enriched for binding sites for PPRC1, RBM8A, and TRA2 proteins. RBM8A has been shown to couple pre- and post-mRNA splicing events (25, 26) and TRA2 is also associated with regulating pre-mRNA splicing (27, 28). R3 regions immediately downstream of exons that are linked to alternative PASs were enriched for an A-rich motif, which can be recognized by a number of poly(A) binding proteins (**Table 2**), suggesting a competitive binding to these R3 sequences and genuine poly(A) tails.

Discussion

Short-read RNA sequencing has become central in assessing the global RNA expression patterns. However, as a result of the complexity of human transcriptome, these approaches disappoint in precise reconstruction and reliable expression estimation of transcript variants (6, 7, 29), owing to the short length of sequencing reads. In contrast, single-molecule long-read sequencing provides a unique opportunity to reveal the true complexity of the transcriptome as it can determine the full structure of individual transcripts by single-pass and full-length sequencing.

Here, we have analyzed the deepest and longest transcriptome data so far to better understand the extent of interdependencies between transcription and mRNA processing. Notably, full-length mRNA sequencing and de novo identification of high-quality sequence of transcript variants uncovered an unprecedented amount of potentially novel transcripts. The majority of alternative mRNA processing events could not be attributed to those that are cataloged in the latest Ensembl Alternative Splicing Events database. Our findings not only unravel a higher level of alternative transcription, splicing and polyadenylation in MCF-7 transcriptome than previously appreciated, but also provide valuable information on the preferential selection and interdependency between these processes.

We showed that transcription initiation, splicing and 3' end formation are tightly coupled in nearly 50% of genes with multiple transcripts and such interdependencies can be found across the entire length of the mRNA molecule. Notably, we report an unforeseen and unprecedented number of genes that undergo a vigorous preferential selection during transcription and mRNA processing as the choice of transcription initiation subsequently influences both alternative splicing of exons and the usage of alternative poly(A) site. These genes were enriched in mRNA processing and protein degradation pathway that may be in line with the previously observed auto-regulation of mRNA processing factors (30) and feedback loops between protein degradation and mRNA synthesis.

Ample evidence points at the critical role for RNA Pol II in the coordination between transcription and mRNA processing (reviewed in 5, 31-33). It has been shown that RNA Pol II initiation, pausing, and elongation rate can influence alternative splicing and polyadenylation of transcripts (34-37). Moreover, the C-terminal domain of RNA Pol II likely acts as a scaffold for regulatory factors that are involved in

TSS -coupled exons						
Domain	**Motif**	**Length Restriction**	**E-value**	**Fold Enrichment**	**RBP Motif**	**RBP**
R1	SGCGSGC	7 nt.	4.2E-02	1.44	RNCMPT00113	RBM14
R2	BCGCG	5 nt.	2.1E-02	1.18	RNCMPT00045	PPRC1
					RNCMPT00056	RBM8A
	GAWGARG	5 nt., 7 nt.	1.8E-02	1.16	RNCMPT00078	TRA2
R3	CGCSG	-	6.7E-09	1.35	RNCMPT00052	RBM14
Exon -Exon Coupling						
Domain	**Motif**	**Length Restriction**	**E-value**	**Fold Enrichment**	**RBP Motif**	**RBP**
R1	-	-	-	-	-	-
R2	RAAGAAG	7 nt.	1.8E-02	1.15	RNCMPT00078	TRA2
R3	-	-	-	-	-	-
PAS -coupled exons						
Domain	**Motif**	**Length Restriction**	**E-value**	**Fold Enrichment**	**RBP Motif**	**RBP**
R1	-	-	-	-	-	-
R2	-	-	-	-	-	-
R3	AAAARH	-	3.3E-56	1.33	RNCMPT00043	PABPC4
	AAAAAAVB	7 nt.	3.4E-55	1.64	RNCMPT00043	PABPC4

Table 2. The RNA-binding protein motifs associated with alternative exons that are coupled to TSS, other alternative exons, or PAS.

splicing and polyadenylation (reviewed in 33). Concordantly, we found an enrichment of coupling events in larger genes that seem to undergo a more extensive regulation during mRNA synthesis. However, the exact mechanisms by which the coordination is achieved remain largely unclear.

From previous studies it became clear that polyadenylation couples with splicing machinery to influence the removal or inclusion of the last intron (15, 38, 39). We now show that (i) the interdependencies between splicing and polyadenylation are not necessarily restricted to the final introns, (ii) that they can also involve introns that are far from the poly(A) site and (iii) that the coupling between splicing and alternative polyadenylation is not restricted to tandem 3' UTRs. The exact mechanisms by which these coupling events are achieved fall beyond the scope of this study. Previously, it has been shown that spliceosome components are also part of the human pre-mRNA 3'-end processing complex (40). Moreover, it is likely that there are RNA-binding proteins with a dual role in alternative splicing and polyadenylation in order to coordinate mRNA processing events. hnRNP H17, CstF6439, MBNL1 and ELAV1 (HuR) (19, 41-43) are a few examples of such proteins. We found MBNL binding motifs enriched in the sequences upstream of polyadenylation sites coupled with alternatively spliced exons. Interestingly, these regions often lacked canonical or non-canonical poly(A) signals. This suggests that MBNL proteins mark alternative poly(A) sites and play a dual and possibly coordinating role in splicing and polyadenylation. This is in line with previous studies in MBNL1-deficient cells where both splicing and polyadenylation were shown to be disrupted (19, 20).

Based on the reported sequence preference of MBNL proteins (20), MBNL3 is the most likely candidate of the MBNL family responsible for the coordination between alternative splicing and polyadenylation of transcripts in MCF-7 cells. However, it is not clear to what extent these findings can be extrapolated to other cell lines and cell types. In MCF-7 cells, the balance between alternative poly(A) site usage is shifted to more proximal poly(A) sites (18, 44). The absence of binding sites for regulatory proteins and miRNAs can enhance the tumorigenic activity of MCF-7 cells by allowing transcripts to escape from inhibition (18). Our findings mostly relate to the use of alternative polyadenylation by utilizing different 3' UTRs and not tandem polyadenylation sites that are in the same 3' UTR region. It is not clear whether MBNL-mediated polyadenylation, coupled with transcription initiation and splicing, is achieved through direct recruitment of RNA processing machinery or via alteration of secondary structure and formation of RNA molecules that, in turn, affect the choice for poly(A) site usage. Our analysis also identified a few more candidates with dual roles in mRNA processing, notably RBM14 (23, 24), RBM8A (25, 26, 45) and TRA22 (7, 28), which warrant further investigations by performing additional functional assays.

This study demonstrates that our understanding of transcript structures and coordinating mechanisms that regulate transcription and mRNA processing is far from complete, even in well-characterized human cell lines such as MCF-7. Single-molecule full-length RNA sequencing of other human tissues and cell-lines can provide a comprehensive view of the true complexity of the human transcriptome. Moreover, although it has been shown that single-nucleotide variants can alter the inclusion of exons in transcripts (9), it is of interest to identify variants that can affect allele-specific coupling between transcription and mRNA processing. Together, these can offer a better understanding of the mechanisms that control transcription and mRNA processing.

Methods

RNA sample preparation, library preparation, and sequencing

The methodologies and experimental settings for RNA preparation, cDNA synthesis, library preparation,

and sequencing are described at: http://blog.pacificbiosciences.com/2013/12/data-release-human-mcf-7-transcriptome.html.

Annotation of transcripts using isoform-level clustering algorithm (ICE)

The identification, polishing, and annotation of transcripts were previously carried out using the ICE algorithm and made public by Pacific Biosciences. To find transcript clusters, ICE performs a pairwise alignment and reiterative assignment of full-length reads to clusters based on likelihood. This process is followed by consensus calling and further polishing of the sequence to reduce the redundancy and increase the overall accuracy of sequences for identified transcript variants. For further information on the methodology and experimental settings visit: https://github.com/PacificBiosciences/cDNA_primer/wiki.

Comparison to the GENCODE annotation

We used GENCODE annotated transcripts (version 19) as reference to compare with the identified transcripts in the human MCF-7 transcriptome data. The comparison was carried out using cuffcompare from the Cufflinks suite (46).

Definition of transcription start site, polyadenylation site, and donor and acceptor splice sites

In this study, by processing the GFF file that contains the annotation of all identified transcripts and exon/intron boundaries (defined by the genomic position and strand on the hg19 reference sequence), a list of all transcription and mRNA processing events is produced. Transcription start sites (TSSs) are defined as the first genomic position of each transcript structure. Polyadenylation sites (PASs) are defined as the last genomic position of each transcript. The most upstream and downstream genomic positions of exons were used to define donor and acceptor splice-sites, respectively. However, for the first exon only the donor site is described as the first position is defined as transcription start site. Likewise, the last exon does not contain a donor splice site as the position is defined as polyadenylation site. If multiple transcripts share the same feature, then only one copy is kept in the unique set of features at each locus. Furthermore, the union of all unique exons is defined as the available sequence at each locus. This is also illustrated in **Figure 1B**.

Alignment and quantification of supporting reads for each transcript

The number of reads aligned to each transcript was used as the supporting evidence for each transcript structure. To identify the number of supporting reads, the polished sequences of all unique transcripts were used as a reference for the unique alignment of raw reads using BLASR47. Other parameters were set default and according to the Pacific Biosciences guidelines.

Statistical analysis

After defining unique features (transcription start sites, exons, and polyadenylation sites) and identifying the number of supporting reads for transcripts at each locus, all possible pairwise comparisons between features were made. To do this, the sum of all reads that support the presence of the two selected features in all observed transcripts is reported in a two-by-two contingency table. The table describes the number of times two features are observed in the same transcript

or exclusively, as well as the sum of reads that are mapped to transcripts that do not support the presence of either features (**Figure 1C**). A significant linkage between two features is assessed using the Fisher's exact test. The mutual inclusivity or exclusivity of coupled features are defined using their log-transformed odd-ratio. All p-values are adjusted using Bonferroni multiple testing correction. Many aspects of this analysis were carried out in Python and R.

Pathway analysis

This analysis was performed on a subset of genes that contain at least one coupling event and separated based on the type of coupling between features: TSS-exon, TSS-PAS, exon-exon, and exon-PAS. A list of all genes that could be detected in this study and subsequently annotated using GENCODE v19 (10,673 ENSEMBL gene IDs) was used as a background. Prior to the analysis, official gene symbols were converted to DAVID IDs. KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis was performed using DAVID Functional Annotation Tool (48).

Annotation of alternative exons

The genomic region of significantly coupled transcription start sites, alternative exons, polyadenylation sites are compared to the Ensembl Alternative Splicing Events annotation to characterize regions that have already been associated with one of the following ten classes:

1) Cassette exon
2) Intron retention
3) Mutually exclusive exons
4) Constitutive exon
5) Exon isoform
6) Intron isoform
7) Alternative 3' site
8) Alternative 5' site
9) Alternative first exon
10) Alternative last exon

To assess the enrichment of different categories of alternative splicing events in the Ensembl annotation, all the transcription start sites, exons, and polyadenylation sites that are present in the MCF-7 transcriptome data were also attributed to this annotation to serve as a background quantification.

Sequence motif analysis relative to polyadenylation sites

For each detected locus, we reported the last nucleotide as polyadenylation site. Each genomic location was converted into a BED format. Strand specific genomic sequences located up to 35 nucleotides upstream each unique polyadenylation site were extracted, in a FASTA format, using UCSC Table Browser (GRCh37/hg19). FASTA files were parsed using a custom bash script to count the number of sequences containing specific 6-mer motifs: one of the two canonical polyadenylation signals AATAAA and ATTAAA, or one of the eleven non-canonical polyadenylation signals (AAGAAA, AATACA, AATAGA, AATATA, AATGAA, ACTAAA, AGTAAA, CATAAA, GATAAA, TATAAA, TTTAAA). Subsequently, the same 6-mer motifs were counted for each unique PAS significantly coupled to TSSs or exons and for each unique PAS that did not show a significant coupling.

For PASs that could not be attributed to known poly(A) signals, we ran DREME (49) (v. 4.9.1) to

identify enriched motifs. Randomly shuffled set of sequences was generated from the original sequences of the examined PASs and used as a background set. In addition, the sequences of known recognition motifs for MBNL proteins (20-22) were counted for each set using a custom script. Subsequently, the enrichment of each motif was assessed by Fisher's exact test.

Tandem 3' UTR analysis

This analysis was performed to identify loci that contain tandem 3' UTRs (loci that contain more multiple PASs located in the same last exon). Custom scripts were used to identify loci that contain at least two PASs that share the same coordinates of the last exon start. The number of loci with tandem 3' UTRs was calculated for those in which PAS was significantly coupled to alternative exons and for those that did not show any significant interdependencies between alternative exons and the PAS usage.

Sequence motif analysis relative to acceptor and donor sites

For each detected locus, we reported the first and last nucleotide of each exon as acceptor splice site and donor splice site, respectively. Each unique genomic position was converted into a BED format and the strand specific sequences of 2 nucleotides length were extracted using UCSC Table Browser (GRCh37/hg19) for both acceptor and donor splice sites. A custom bash script was used to count the number of dinucleotide sequences containing 'GT' and/or 'AG'.

RNA binding motif analysis

We used MEME suite tools to identify enriched sequence motifs present in exons significantly coupled with TSSs, PASs or other alternative exons. For each unique exon, three regions were considered: R1 (containing up to 35 nucleotides upstream the acceptor splice site), R2 (containing 32 nucleotides downstream the acceptor splice site and 32 nucleotides upstream the donor splice site), and R3 (containing up to 40 nucleotides downstream the donor splice site). R1, R2 and R3 regions were obtained by extracting strand specific FASTA sequences using UCSC Table Browser (GRCh37/hg19).

We locally ran DREME (49) (v. 4.9.1) for each region separately, and performed a motif search using a negative background (R1, R2 and R3 regions from exons that were not significantly coupled). We ran DREME in two modes, one without any limitation for the motifs' width, and one with limiting the search to a minimum width of 5 or 7 nucleotides. In each case, a maximum of 10 motifs with E-values < 0.05 was reported. The remaining parameters were kept as default. We then compared each motif found by DREME against the human RNA-binding motifs database CISBP-RNA using TOMTOM Motif Comparison tool (50). We ran the analysis by setting the Pearson correlation coefficient as comparison function and considered only matches with a minimum false discovery rate (q-values) < 0.05.

REFERENCES

1. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature genetics 40, 1413-1415 (2008).
2. Barash, Y. et al. Deciphering the splicing code. Nature 465, 53-59 (2010).
3. Wang, E.T. et al. Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470-476 (2008).
4. Auboeuf, D. et al. A subset of nuclear receptor coregulators act as coupling proteins during synthesis and maturation of RNA transcripts. Molecular and cellular biology 25, 5307-5316 (2005).
5. Bentley, D.L. Coupling mRNA processing with transcription in time and space. Nature reviews. Genetics 15, 163-175 (2014).
6. Tilgner, H. et al. Accurate identification and analysis of human mRNA isoforms using deep long read sequencing. G3 3, 387-397 (2013).
7. Steijger, T. et al. Assessment of transcript reconstruction methods for RNA-seq. Nature methods 10, 1177-1184 (2013).
8. Koren, S. et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nature biotechnology 30, 693-700 (2012).
9. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M.P. Defining a personal, allele-specific, and single-molecule long-read transcriptome. Proceedings of the National Academy of Sciences of the United States of America 111, 9869-9874 (2014).
10. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. Nature biotechnology 31, 1009-1014 (2013).
11. Au, K.F. et al. Characterization of the human ESC transcriptome by hybrid sequencing. Proceedings of the National Academy of Sciences of the United States of America 110, E4821-4830 (2013).
12. Thomas, S., Underwood, J.G., Tseng, E., Holloway, A.K. & Bench To Basinet Cv, D.C.I.S. Long-read sequencing of chicken transcripts and identification of new transcript isoforms. PloS one 9, e94650 (2014).
13. Treutlein, B., Gokce, O., Quake, S.R. & Sudhof, T.C. Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. Proceedings of the National Academy of Sciences of the United States of America 111, E1291-1299 (2014).
14. Schreiner, D. et al. Targeted Combinatorial Alternative Splicing Generates Brain Region-Specific Repertoires of Neurexins. Neuron (2014).
15. Berget, S.M. Exon recognition in vertebrate splicing. The Journal of biological chemistry 270, 2411-2414 (1995).
16. Martinson, H.G. An active role for splicing in 3'-end formation. Wiley interdisciplinary reviews. RNA 2, 459-470 (2011).
17. Katz, Y., Wang, E.T., Airoldi, E.M. & Burge, C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature methods 7, 1009-1015 (2010).
18. Fu, Y. et al. Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. Genome Res 21, 741-747 (2011).
19. Wang, E.T. et al. Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. Cell 150, 710-724 (2012).
20. Batra, R. et al. Loss of MBNL Leads to Disruption of Developmentally Regulated Alternative Polyadenylation in RNA-Mediated Disease. Molecular cell (2014).
21. Masuda, A. et al. CUGBP1 and MBNL1 preferentially bind to 3' UTRs and facilitate mRNA decay. Scientific reports 2, 209 (2012).
22. Purcell, J., Oddo, J.C., Wang, E.T. & Berglund, J.A. Combinatorial mutagenesis of MBNL1 zinc fingers elucidates distinct classes of regulatory events. Molecular and cellular biology 32, 4155-4167 (2012).
23. Auboeuf, D. et al. CoAA, a nuclear receptor coactivator protein at the interface of transcriptional coactivation and RNA splicing. Molecular and cellular biology 24, 442-453 (2004).
24. Kang, Y.K. et al. Dual roles for coactivator activator and its counterbalancing isoform coactivator modulator

in human kidney cell tumorigenesis. Cancer research 68, 7887-7896 (2008).

25. Kataoka, N. et al. Specific Y14 domains mediate its nucleo-cytoplasmic shuttling and association with spliced mRNA. Scientific reports 1, 92 (2011).
26. Kataoka, N. et al. Pre-mRNA splicing imprints mRNA in the nucleus with a novel RNA-binding protein that persists in the cytoplasm. Molecular cell 6, 673-682 (2000).
27. Dauwalder, B., Amaya-Manzanares, F. & Mattox, W. A human homologue of the Drosophila sex determination factor transformer-2 has conserved splicing regulatory functions. Proceedings of the National Academy of Sciences of the United States of America 93, 9004-9009 (1996).
28. Tacke, R., Tohyama, M., Ogawa, S. & Manley, J.L. Human Tra2 proteins are sequence-specific activators of pre-mRNA splicing. Cell 93, 139-148 (1998).
29. Engstrom, P.G. et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nature methods 10, 1185-1191 (2013).
30. Lewis, B.P., Green, R.E. & Brenner, S.E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proceedings of the National Academy of Sciences of the United States of America 100, 189-192 (2003).
31. Kornblihtt, A.R. et al. Alternative splicing: a pivotal step between eukaryotic transcription and translation. Nature reviews. Molecular cell biology 14, 153-165 (2013).
32. Schor, I.E., Gomez Acuna, L.I. & Kornblihtt, A.R. Coupling between transcription and alternative splicing. Cancer treatment and research 158, 1-24 (2013).
33. Hsin, J.P. & Manley, J.L. The RNA polymerase II CTD coordinates transcription and RNA processing. Genes & development 26, 2119-2137 (2012).
34. Danko, C.G. et al. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. Molecular cell 50, 212-222 (2013).
35. de la Mata, M. et al. A slow RNA polymerase II affects alternative splicing in vivo. Molecular cell 12, 525-532 (2003).
36. Nogues, G., Kadener, S., Cramer, P., Bentley, D. & Kornblihtt, A.R. Transcriptional activators differ in their abilities to control alternative splicing. The Journal of biological chemistry 277, 43110-43114 (2002).
37. Pinto, P.A. et al. RNA polymerase II kinetics in polo polyadenylation signal selection. The EMBO journal 30, 2431-2444 (2011).
38. Cooke, C., Hans, H. & Alwine, J.C. Utilization of splicing elements and polyadenylation signal elements in the coupling of polyadenylation and last-intron removal. Molecular and cellular biology 19, 4971-4979 (1999).
39. Movassat, M., Crabb, T., Busch, A., Shi, Y. & Hertel, K. Coupling between alternative polyadenylation and alternative splicing is limited to terminal introns (560.2). The FASEB Journal 28 (2014).
40. Shi, Y. et al. Molecular architecture of the human pre-mRNA 3' processing complex. Molecular cell 33, 365-376 (2009).
41. Lebedeva, S. et al. Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. Molecular cell 43, 340-352 (2011).
42. Mukherjee, N. et al. Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. Molecular cell 43, 327-339 (2011).
43. Barnhart, M.D., Moon, S.L., Emch, A.W., Wilusz, C.J. & Wilusz, J. Changes in cellular mRNA stability, splicing, and polyadenylation through HuR protein sequestration by a cytoplasmic RNA virus. Cell reports 5, 909-917 (2013).
44. Mayr, C. & Bartel, D.P. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell 138, 673-684 (2009).
45. Ishigaki, Y. et al. Depletion of RNA-binding protein RBM8A (Y14) causes cell cycle deficiency and apoptosis in human cells. Experimental biology and medicine 238, 889-897 (2013).
46. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology 28, 511-515 (2010).
47. Chaisson, M.J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC bioinformatics 13, 238 (2012).

48. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols 4, 44-57 (2009).

49. Bailey, T.L. DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics 27, 1653-1659 (2011).

50. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L. & Noble, W.S. Quantifying similarity between motifs. Genome biology 8, R24 (2007).

SUPPORTING INFORMATION

Supplementary Figures and Tables are available upon request.

S**upplementary Text is available online at** http://nbviewer.ipython.org/urls/git.lumc.nl/mcf7/full-length-rna-coupling/raw/master/2015%20-%20Coupling%20between%20transcription%20and%20mRNA%20processing%20events.ipynb#section%200