RNA Polymerase II Phosphorylated on CTD Serine 5 Interacts with the Spliceosome during Co-transcriptional Splicing

Graphical Abstract

Highlights

- S5P CTD Pol II associates with the catalytic spliceosome
- Elongating Pol II complexes protect about 60 newly synthesized nucleotides
- Co-transcriptional splicing associated with dominant 5’ ss intermediate
- U12-dependent introns are sequentially spliced in association with Pol II

Authors
Takayuki Nojima, Kenny Rebelo, Tomás Gomes, Ana Rita Grosso, Nicholas J. Proudfoot, Maria Carmo-Fonseca

Correspondence
nicholas.proudfoot@path.ox.ac.uk (N.J.P.), carmo.fonseca@medicina.ulisboa.pt (M.C.-F.)

In Brief
Nojima and colleagues demonstrate widespread co-transcriptional splicing of mammalian genes by complex formation between the active spliceosome and the RNA polymerase II S5P CTD isoform. RNA processing intermediates within this complex are invariably detected with the transcript cleaved at 5’ ss but already spliced to upstream exons, implying a sequential splicing process.

Nojima et al., 2018, Molecular Cell 72, 369–379
October 18, 2018 © 2018 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.molcel.2018.09.004
RNA Polymerase II Phosphorylated on CTD Serine 5 Interacts with the Spliceosome during Co-transcriptional Splicing

Takayuki Nojima,1,4 Kenny Rebelo,2,4 Tomás Gomes,2,3 Ana Rita Grosso,2 Nicholas J. Proudfoot,1,5,* and Maria Carmo-Fonseca2,2*

1Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, United Kingdom
2Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, 1649-028 Lisboa, Portugal
3Wellcome trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom
4These authors contributed equally
5Lead Contact
*Correspondence: nicholas.proudfoot@path.ox.ac.uk (N.J.P.), carmo.fonseca@medicina.ulisboa.pt (M.C.-F.)
https://doi.org/10.1016/j.molcel.2018.09.004

SUMMARY

The highly intronic nature of protein coding genes in mammals necessitates a co-transcriptional splicing mechanism as revealed by mNET-seq analysis. Immunoprecipitation of MNase-digested chromatin with antibodies against RNA polymerase II (Pol II) shows that active spliceosomes (both snRNA and proteins) are complexed to Pol II S5P CTD during elongation and co-transcriptional splicing. Notably, elongating Pol II-spliceosome complexes form strong interactions with nascent transcripts, resulting in footprints of approximately 60 nucleotides. Also, splicing intermediates formed by cleavage at the 5’ splice site are associated with nearly all spliced exons. These spliceosome-bound intermediates are frequently ligated to upstream exons, implying a sequential, constitutive, and U12-dependent splicing process. Finally, lack of detectable spliced products connected to the Pol II active site in human HeLa or murine lymphoid cells suggests that splicing does not occur immediately following 3’ splice site synthesis. Our results imply that most mammalian splicing requires exon definition for completion.

INTRODUCTION

Eukaryotic protein coding genes often contain multiple introns that are removed from primary transcripts of RNA polymerase II (Pol II) by splicing. This is catalyzed by the spliceosome, comprising five snRNAs and more than 200 proteins, a subset of which form functional ribonucleoprotein particles. Exon-intron boundaries are defined by short consensus sequences at the 5’ and 3’ splice sites (ss) that mediate recognition by the spliceosome. Spliceosome assembly also requires recognition of an intronic catalytic adenosine (branch point) and, for mammalian introns, a polypyrimidine tract located between the branch point adenosine and 3’ ss. Within the 3D structure of the spliceosome, these key pre-mRNA sequences are forced into an RNA catalytic center that triggers splicing (Papasaikas and Valcárcel, 2016). This involves two transesterification reactions. First, the 2’ hydroxyl group (OH) of the branch point adenosine carries out a nucleophilic attack on the phosphate group at the 5’ ss; then the exposed 3’ OH of the upstream exon cleaves the lariat structured intron away from the 3’ ss, resulting in covalent joining of the two exons (Patel and Steitz, 2003).

Exactly how two splice sites separated by extensive intronic sequence (in mammals, 1 to 100 kb) are brought together during early stages of spliceosome assembly remains a puzzle. Possibly, the 5’ exon is tethered to elongating Pol II until complete synthesis of the downstream intron has occurred. This would maintain the two splice sites in close proximity, irrespective of intron size (Holland et al., 2016). Co-transcriptional tethering of the 5’ exon may involve the flexible C-terminal domain (CTD) of Pol II’s largest subunit, Rbp1, which acts to coordinate transcription with pre-mRNA processing (Heidemann et al., 2013; Zaborowska et al., 2016). Mammalian CTD comprises 52 repeats of a heptad with the consensus amino acid sequence YSPTSPS. This is extensively modified post-translationally, principally by phosphorylation to S2P, S5P, S7P, T4P, and Y1P (Hintermair et al., 2012; Hsin and Manley, 2012; Mayer et al., 2012). At the promoter, Pol II CTD is largely unphosphorylated, but then, following initial elongation, it is converted to S5P and Y1P. This may facilitate recruitment of the capping enzyme complex. Further elongation correlates with a reduction in S5P but an increase in S2P. Transcription past the polyadenylation site then triggers recruitment of the 3’ end cleavage and polyadenylation machinery to S2P. It is possible that T4P also plays a role here, as it has recently been correlated with Pol II termination regions (Hintermair et al., 2012; Schlackow et al., 2017).

The kinetics of co-transcriptional splicing in budding yeast has been extensively characterized (Wallace and Beggs, 2017). The short size of yeast (S. cerevisiae) introns implies an intron definition model for the splicing mechanism (Ast, 2004; Barrass and Beggs, 2003; Chen and Manley, 2009). In particular, Pol II pausing has been observed over the 3’ ss of downstream exons in the few genes that harbor introns. This may facilitate rapid
A mNET-seq and MS strategies
DNA/RNA digestion
Pol II IP
Sequencing
Nascent RNA profiles
Mass spectrometry
Elongating complex
protein profiles

B mNET-MS
Pol II
spliceosome
SR proteins
others

C mNET-MS
SSP vs S2P
SSP vs T4P

D Western Blot

E Western Blot
DNA/RNA digestion
Glycerol Gradient

F Western Blot
G Western Blot

(legend on next page)
spliceosome recruitment and intron splicing (Alexander et al., 2010). Furthermore, splicing in S. cerevisiae has been shown to occur within a short kinetic window following Pol II elongation through an intron and into the adjoining exon (Oesterreich et al., 2016). In mammals, the prevalence and larger size of introns instead predicts an exon definition mechanism. Binding of regulatory proteins to specific RNA sequences (splicing enhancers and inhibitors) act to select exons for splicing by promoting communication between 3' ss and 5' ss flanking the exon (Astar, 2004; Chen and Manley, 2009). The interplay between these regulatory factors dictates both constitutive and alternative splicing. Notably, transcription kinetics strongly impact splicing decisions. Slow Pol II elongation rates favor splicing by allowing more time for spliceosome assembly (Dujardin et al., 2014; Muñoz et al., 2010). Exons in general display a higher nucleosome density, suggesting that chromatin structure acts as a kinetic barrier to Pol II elongation (Nieto Moreno et al., 2015; Schwartz and Ast, 2010).

We have developed a strategy for native elongation transcript sequencing using mammalian cells (mNET-seq). We found an accumulation of transcripts mapping precisely to the 3' end of exons, as expected for intermediates formed after the first trans-esterification splicing reaction (Nojima et al., 2015). This indicates that splicing must occur within a stable complex formed between the spliceosome and Pol II. Unexpectedly, splicing intermediates were preferentially detected by antibodies specific for SSP CTD, suggesting that splicing occurs in association with this Pol II phospho-isof orm. We now show biochemically and by mass spectroscopy that both protein and snRNA components of the spliceosome associate with Pol II complexes containing SSP CTD. Notably, initial cleavage of the intron creates a dominant 5' ss intermediate that remains embedded in the Pol II- associated spliceosome. In contrast, 3' ss intermediates corresponding to released intron lariats were detected at very low levels, suggesting fast dissociation of the spliceosome from Pol II upon completion of splicing.

RESULTS

Catalytically Active Spliceosome Associates with SSP CTD Pol II

mNET-seq involves immunoprecipitation (IP) of human Pol II elongation complexes from formaldehyde-solubilized by micrococcal nuclease (MNase) digestion with Pol II antibodies specific for different CTD isoforms (Nojima et al., 2015; Schlackow et al., 2017). RNA is isolated from these complexes and sequenced by linker ligation onto the RNA 3' ends derived either from nascent RNA in the Pol II active site or co-transcriptional RNA processing intermediates. Notably, we showed that co-transcriptional splicing is associated with Pol II phosphorylated on the CTD serine 5 position (SSP) (Nojima et al., 2015; Schlackow et al., 2017). We reasoned that, as well as sequencing IPed RNA, we should also be able to establish the protein composition of these specifically IPed Pol II complexes by mass spectroscopy (MS), termed the mNET-MS method (Figure 1A). A label-free quantitative proteomics approach (Hubner et al., 2010; Hubner and Mann, 2011) was used to determine the abundance of proteins enriched in Pol II complexes IPed with antibodies specific for SSP, S2P, and T4P CTD relative to a mock (nonspecific IgG) IP control. The fold increase of spectral intensities was compared to p values determined by t tests, as previously described (Harlen et al., 2016), generating volcano plots (Figure 1B). Significant interactors were taken to be proteins identified as enriched using a false discovery rate of 0.05. We show that complexes precipitated by all three antibodies contained Pol II subunits, transcription factors, histones, chromatin-associated proteins, RNA-binding proteins, and RNA-processing factors (Table S1). Scatterplots with fold-enrichment reveal that Pol II subunits were precipitated with similar efficiency to SUPT6H and PHRF1 (Figure 1C). Remarkably, complexes precipitated by the SSP CTD antibody contained multiple SR and SR-related proteins (SRSF7, 9, 10; SCAF1, 11), as well as spliceosome components. These include Sm proteins, U1 and U2 snRNP-specific proteins, and components of spliceosomal A, B, Bact, and C complexes (Figures 1B and S1; Table S1). This was confirmed by western blot analysis of fractions IPed with SSP compared to S2P antibodies; SSP-specific enrichment was observed for the U5 snRNP protein, SNRP116, the splicing scaffold protein, PRPF8, and the U2 snRNP protein SF3B3 (Figure 1D). There are several possible reasons for this selectivity. Either spliceosomal components interact specifically with SSP CTD or, possibly, SSP and S2P epitopes display differential antibody accessibility. However, we consider this later possibility unlikely, as a recent proteomic analysis in yeast also revealed an enrichment of spliceosomal components in SSP Pol II complexes (Harlen et al., 2016).

As an alternative approach to studying the interaction between Pol II and the spliceosome, chromatin was solubilized by MNase treatment and then fractionated on a 10%–30% glycerol gradient followed by protein extraction from 22 separated fractions (Figure 1E). These were subjected to western blotting with antibodies against total, S2P, and SSP Pol II as well as U5...
snRNP protein SNRP116 and histone H3 (Figure 1F). Significant amounts of S2P, SSP Pol II, and histone H3 were detected in the middle and heavy fractions. However, the heavy fraction contained most of the snRNP116 (Figure 1F). Further analysis with a wider range of antibodies against pooled light, middle, and heavy fractions confirms that U5 snRNP proteins were predominantly associated with the heavy fraction (Figure 1G). In contrast, U2 snRNP proteins appeared similarly distributed in the middle and heavy fractions, and U1 snRNP proteins were predominantly associated with the middle fraction (Figure 1G). The histone mark H3K9me2, which is typically associated with transcriptional repression, was predominantly in the middle fraction, whereas the histone elongation mark H3K36me3 was in both the middle and heavy fractions. These results indicate that the heavy fraction contains spliceosome-associated SSP Pol II-elongating polymers on a chromatin template.

Since SSP CTD modified Pol II interacts with protein components of the spliceosome, we also analyzed spliceosomal snRNAs. We reasoned that mNET-seq should detect the free 3’ OH ends of mature spliceosomal snRNA associated with Pol II. Indeed, an accumulation of mNET-seq signal was found mapping to the 3’ ends of U1, U2, U4, and U5 snRNAs (Figure 2A). As expected, no 3’ end peak was detected in U6 snRNA, which contains a 2’,3’ cyclic phosphate terminal group at the 3’ end (Lund and Dahlberg, 1992). Lower mNET-seq peaks were also found at the 3’ ends of the minor spliceosome U11, U12, and U4atac snRNAs, but not in U6atac snRNA (Figure 2B). A weak 3’ end signal was further detected for U7 snRNA, which is associated with histone pre-mRNA 3’ end processing. As expected, no 3’ end peak was detected for U3 snRNA, which is involved in the processing of Pol I-synthesized pre-rRNA (Figures 2C and S2A). To confirm the specificity of these results, we tested whether the spliceosome snRNAs’ 3’ end signal was sensitive to treatment with the strong detergent empigen. We have previously employed empigen to separate RNA processing complexes from Pol II (Schlackow et al., 2017). Notably mNET-seq signals of both major and minor spliceosome snRNAs 3’ ends were substantially diminished (Figure 2D). A more detailed comparison of the association of mature spliceosomal snRNA (based on mNET-seq 3’ end reads) for all phospho CTD isoforms of Pol II revealed higher signal with SSP CTD antibody for U2, U5, and the minor spliceosome snRNA. In contrast, U1 snRNA showed similar read levels for all five phospho CTD isoforms (Figure 2E).

To complete this mNET-seq analysis of spliceosome-associated RNA, we also tested the distribution of the 5’ ss splicing intermediate. We first confirmed that RNA fragments mapping to the 5’ ss are predominantly detected by SSP CTD antibody as opposed to S2P, Y1P, T4P and S7P antibodies (Figure S2B). We also compared mNET-seq profiles with different CTD SSP-specific antibodies; MABI0603 (MBL international), routinely used in our mNET-seq/SSP analysis, and ab5131 (Abcam), often used in Pol II ChiP analysis. Similar profiles of 5’ ss peaks were obtained (Figures S2C and S2D), consistent with our previous ChiP analysis that showed the same specificity for SSP CTD (Nojima et al., 2015). We next generated mNET-seq/SSP and SSP libraries from the heavy and middle fractions of the glycerol gradient of MNase-digested chromatin. Notably, the 5’ ss splicing intermediate was only detected from the heavy SSP library as shown for TARS (Figure S2E) and in a metagene profile of 40,896 aligned exon intron junctions corresponding to spliced events (Figure S2F). These results confirm the existence of a complex between Pol II and the spliceosome in heavy fractions. Finally, since the 5’ ss splicing intermediate detected by mNETseq/SSP is empigen sensitive (Figure S2G), we predict that the 5’ ss RNA is associated with the spliceosome rather than with the Pol II active site.

In summary, these data strongly argue that phosphorylation of CTD serine 5 residues is not restricted to transcription initiation but is also present during elongation and co-transcriptional splicing. It is clear from the range of data presented (Figures 1, 2, S1, and S2) that catalytically active spliceosome stably associates with SSP CTD Pol II.

mNET-Seq Reveals Elongating RNA Pol II Footprints

We next employed mNET-seq to investigate the extent of nascent transcript interaction with both Pol II and the spliceosome. Since this procedure involves MNase digestion before RNA purification and library preparation, it can be used for Pol II and spliceosome footprinting analysis akin to ribosome footprinting. Previously, we size selected RNA of 30–100 nt for library preparation (Nojima et al., 2015; Schlackow et al., 2017). To explore a larger RNA size range, we prepared new mNETseq/SSP libraries using long-sized (60–160 nt) or short-sized (20–60 nt) RNA fractions with a read length of 150 bp and 75 bp, respectively (Figure 3A). To analyze Pol II footprints independently of spliceosome assembly, we examined the length of reads mapping to a region from the TSS to +150 nucleotides in first exons that were at least 200 nt long. Most reads (> 85%) obtained from the short-sized RNA fractions were 18–40 nt long with a peak at 27 nt, whereas those for the long-sized RNA fractions were 40–100 nt long with a peak at 60 nt (Figure 3B). We infer that Pol II complexes protect a total of either 27 or 60 nascent transcript nucleotides. Previous structural data indicate that nascent RNA must be at least 15 nt in length to reach the Pol II surface (Andrecka et al., 2009; Kettenberger et al., 2004) and 17–23 nt to reach the active site of the capping enzyme, which binds Pol II at the end of the RNA exit tunnel (Martinez-Rucobo et al., 2015). To determine whether binding of the capping enzyme to Pol II explains the 20–30 nt footprint, we compared the length of reads mapping to the same exons on the region from the TSS to +50 nts and from +100 to +150 nts (Figure S3A). A similar footprint was obtained irrespective of read position with mNET-seq libraries prepared after IP with either SSP and S2P antibodies (Figure S3A). This suggests that once nascent transcripts emerge from the polymerase exit channel, they are protected from MNase by RNA-binding proteins such as the capping enzyme. The larger 60-nt footprint obtained from long-sized RNA fractions (Figure 3B) was further characterized. Analysis of mNET-seq profiles obtained from the long-sized RNA fractions no longer shows the characteristic peaks corresponding to promoter proximal transcriptional pausing that typically occurs within the first 50 bp of the transcription start site (TSS). This is shown both for the TSS region of DAP3 (Figure 3C) and a metagene profile of long and short mNET-seq/SSP TSS reads (Figure 3D). We conclude that the mNET-seq protocol distinguishes two types of elongating Pol II complexes, a smaller one corresponding to a nascent RNA footprint of 27 nt and a larger one protecting approximately 60 nt of nascent transcripts (Figure 3E).
Next, we examined the read length of splicing intermediates in mNETseq/S5P and S2P libraries prepared from the short and long-sized RNA fractions (Figure S3B). The results show that the most frequent length of both 5' ss intermediates (i.e., exons that have been cleaved at the 5' ss but not yet ligated) and 3' ss intermediates (i.e., released intron lariats) is 30–32 nt in the short libraries and 57–60 nt in the long libraries. As splicing intermediates are spliceosome-bound, we conclude that these fragments result from the footprinting of splicing complexes associated with S5P and S2P CTD Pol II. Notably, a very recent analysis of RNA co-purifying with late-stage spliceosomes similarly revealed that splicing complexes protect a total of approximately 56 nucleotides (Chen et al., 2018).
mNET-Seq Does Not Detect Spliced Nascent Transcripts

We show above (Figures 1 and 2) that the spliceosome forms a complex with Pol II. We therefore tested if splicing can occur immediately after the 3' ss in a nascent transcript emerges at the Pol II surface. We reasoned that, if this was the case, then mNET-seq should detect spliced nascent transcripts. Analysis of mNET-seq libraries prepared from the short RNA fraction revealed spliced exon reads corresponding to approximately 300 splicing events, while over 20,000 splicing events were detected in libraries prepared from the long RNA fraction (Figure 3F). Most spliced events detected by spliced exon reads were predominantly observed in mNET-seq/S5P libraries, consistent with our data indicating that complexes of Pol II and associated catalytically active spliceosomes are preferentially IPed by S5P CTD antibodies (Figures 1, 2, S1, and S2). At this stage in our analysis, we added additional mNET-seq/S5P libraries for the mouse lymphoid cell line TAP to generalize our results on co-transcriptional splicing to a different mammal and cell type. We also performed heat-map analysis of mNET-seq signals and detected more S5P signals on spliced exons than other phospho CTD isoforms (Figures 1D and 2D).
used in next-generation sequencing can internally prime on spliceosomes and murine libraries led us to appreciate that cDNA primers but not from the Pol II active site. The combined analysis of human and murine libraries demonstrated that even though numerous spliced read transcripts are specifically detectable in our mNET-seq/S5P libraries, they do not reflect immediate splicing of transcripts synthesized by a polymerase that has elongated just a few nts past the 3′ end as seen in S. cerevisiae (Oesterreich et al., 2016). Although spliced reads may correspond to contaminating fully spliced mRNA that is detected due to mis-priming, it is unlikely that contaminants would be specifically enriched by S5P antibody. Rather, these spliced reads likely derive from fragments of spliced products connected to Pol-II-associated spliceosomes.

We further reasoned that if a spliceosome assembles onto nascent pre-mRNA as soon as the 3′ ss is exposed at the Pol II surface, akin to the binding of the capping machinery, this should result in extended nascent transcripts fragments protected from MNase digestion. We therefore analyzed the length distribution of reads mapping to included or skipped exons, but we obtained similar footprint sizes (Figure S3J). Taken together, the lack of detectable spliced read transcripts connected to the Pol II active site and the absence of combined Pol II-spliceosome footprints argue that most splicing occurs when Pol II has elongated for a significant distance beyond the 3′ ss. These observations are consistent with the view that most mammalian splicing requires exon definition for successful completion and therefore, unlike yeast, mammalian splicing only occurs following full exon synthesis.

Co-Transcriptional Splicing Is Globally Associated with 5′ ss Intermediates

The ability of mNET-seq to map splicing intermediates enables in-depth investigation of the extent of co-transcriptional splicing in vivo. Because the first step of the splicing reaction (i.e., cleavage at the 5′ ss) results in multiple mNET-seq reads mapping to the very last nucleotide of exons, we sought to identify the proportion of exons with 5′ ss peaks. For this we adapted a previously described algorithm that finds nucleotides where the read density is at least three standard deviations above the mean in a defined window (Churchman and Weissman, 2011). Upon analyzing replicates of mNET-seq/S5P long and short RNA libraries, we found substantial overlap of approximately 25,000 exons showing 5′ ss peaks between three independent datasets (Figure S4A). In contrast, 3′ ss peaks that correspond to released intron lariats were only detected at very low levels (up to 170 cases) (Figure 4A). Like 5′ ss peaks, these few cases of 3′ ss peaks were again predominantly detected by the S5P antibody (Figure S4B). We also investigated the intensity of the signal at the end of exons in the long and short mNET-seq/S5P libraries. As shown for FUS (Figure S4C) and in a meta gene analysis (Figure S4D), higher 5′ ss peaks were seen for the long library. We next ascertained the proportion of splicing events in the HeLa cell transcriptome that show a 5′ ss intermediate by performing a global analysis of the percent splicing index (PSI). This measures the inclusion level of internal exons based on nucleoplasm-derived RNA-seq (Nojima et al., 2015) and shows that nearly all exons with 5′ ss peaks are fully included in mature mRNAs (PSI > 0.9). Remarkably, 88% of efficiently spliced exons show a 5′ ss mNET-seq peak, indicating that this splicing intermediate is a dominant splicing feature (Figure 4B).

mNET-Seq Reveals Sequential Co-transcriptional Splicing

We suggest above that the majority of mammalian splicing occurs in association with Pol II but with a lag between synthesis of the 3′ ss and assembly of the spliceosome onto the nascent transcript. We next investigated the prevalence of upstream spliced exons covalently connected to 5′ ss splicing intermediates, as this would be indicative of sequential co-transcriptional splicing. In HeLa long-read mNET-seq/S5P libraries, we detected over 4,000 exons with spliced reads mapping to splicing intermediates (Figure 4C). Although these were more frequently detected in mNET-seq/S5P, both S5P and S2P libraries contained approximately one quarter of all spliced exon reads associated with splicing intermediates (Figure S4E). We also detect 5′ ss intermediates associated with spliced reads in TAP long-read mNET-seq/S5P libraries, indicating that both cell types display a predominance of spliced relative to unspliced reads associated with the 5′ ss intermediate (Figure 4D). Thus, when these exons are cleaved at the 5′ ss, their upstream intron has already been spliced (Figure 4E). However, sequential splicing is not an obligatory rule, as we also observe 5′ ss intermediates associated with unspliced reads, particularly in the case of alternatively spliced exons (Figure 4F). Remarkably, the rarer examples of minor U12-type introns also show a predominance for spliced over unspliced 5′ ss intermediates in both HeLa and TAP cells (Figures 4G and S4F). Although previous studies suggested that U12-associated splicing is less efficient than splicing of U2-type introns (Turunen et al., 2013), our results demonstrate that U12-type introns have been excised by the time a spliceosome initiates splicing of the downstream exons.

**DISCUSSION**

We have previously described a dominant splicing intermediate generated by the first catalytic step in splicing, which is associated with S5P CTD-modified Pol II (Nojima et al., 2015). This suggested that cleaved upstream exons are retained on the
elongating Pol II complex within a coIPed spliceosome. Our new results now establish this hypothesis (Figure 4H). Thus, both spliceosome-associated proteins (Figure 1) and snRNAs (Figure 2) selectively associate with S5P CTD Pol II. In particular, the observed enrichment for U5 snRNP as well as most protein components of spliceosomal Bact and C complexes (Figures 1, 2, and S1) shows that the catalytically active spliceosome associates with S5P CTD Pol II during co-transcriptional splicing. Interestingly, we detect lower levels of U1 snRNP in our datasets. Even though U1 snRNP is a known Pol II interactor (Das et al., 2007; Harlen et al., 2016), its association is likely transient in nature, possibly because it acts as a splice site selector factor rather than playing a direct role in spliceosome-mediated catalysis (Papasaikas and Valcárcel, 2016). Furthermore, it also acts to block the recognition of cryptic polyA signals often found within large mammalian introns (Oh et al., 2017). We consistently observed that spliceosomal proteins and snRNAs were much more abundant in Pol II complexes IPed with antibodies specific for S5P CTD than antibodies specific for S2P and T4P CTD. In the case of T4P, ChIP-seq analysis suggests that this

Figure 4. mNET-Seq Reveals Sequential Co-transcriptional Splicing
(A) Number of exons with a peak of mNET-seq/S5P signal detected at either the 5′ ss or the 3′ ss (only peaks detected in all three libraries prepared with long and short RNA fragments were considered).
(B) Percent spliced in (PSI) score for all exons containing a peak of mNET-seq/S5P signal at 5′ ss (aggregated data from three libraries). Pie chart depicts the proportion of spliced exons (exons with PSI > 90% estimated from nucleoplasmic RNA-seq data) that contain a peak of mNET-seq/S5P signal at the 5′ ss (aggregated data from three libraries).
(C) Venn diagram comparing spliced 5′ ss intermediates detected in three independent mNET-seq/S5P and mNET-seq/S2P libraries prepared with long RNA fragments.
(D) Ratio between spliced and unspliced reads for all splicing events associated with at least one spliced read mapping to the 5′ ss intermediate. Graphs depict frequency of events with indicated proportions of spliced reads in HeLa and TAP cells.
(E–G) Diagrams showing spliced and unspliced 5′ ss intermediates associated with a constitutive exon (E), an alternative exon (F) and a U12-type intron (G).
(H) Model showing sequential co-transcriptional splicing associated with S5P CTD of Pol II.
modification is associated with transcription termination (Hintermair et al., 2012). Its lack of spliceosomal association is therefore predictable. In contrast, numerous ChIP-seq studies revealed that Pol II complexes containing SSP-modified CTD are enriched at the promoter proximal region and diminish thereafter (for a comprehensive review, see Eick and Geyer, 2013). However, a splicing-related accumulation of SSP Pol II along gene bodies was found both in HeLa cells (Batsché et al., 2006) and in yeast (Alexander et al., 2010; Chathoth et al., 2014). Moreover, a recent proteomic analysis of phospho-specific Pol II complexes in yeast revealed that the interactome of SSP-modified Pol II is enriched for splicing-related components (Harlen et al., 2016). Our results extend these observations by showing that a SSP Pol II-associated, catalytically active spliceosome exists as a stable entity in human cells (Figure 4H). We suggest that SSP CTD is not cated, catalytically active spliceosome exists as a stable entity in...
REFERENCES

Alexander, R.D., Innocente, S.A., Barrass, J.D., and Beggs, J.D. (2010). Splicing-dependent RNA polymerase pausing in yeast. Mol. Cell 40, 582–593.

Alioto, T.S. (2007). U12DB: a database of orthologous U12-type splicedosomal introns. Nucleic Acids Res. 35, D110–D115.

Andreacka, J., Treutlein, B., Arcusa, M.A., Muschielok, A., Lewis, R., Cheung, A.C., Cramer, P., and Michaelis, J. (2009). Nano positioning system reveals the course of upstream and nontemplate DNA within the RNA polymerase II elongation complex. Nucleic Acids Res. 37, 5803–5809.

Ast, G. (2004). How did alternative splicing evolve? Nat. Rev. Genet. 5, 773–782.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., et al. (2012). The evolutionary landscape of alternative splicing in vertebrate species. Science 338, 1587–1593.

Barrass, J.D., and Beggs, J.D. (2003). Splicing goes global. Trends Genet. 19, 295–298.

Barikowiaq, B., Liu, P., Phatnani, H.P., Fuda, N.J., Cooper, J.J., Price, D.H., Adelman, K., Lis, J.T., and Greenleaf, A.L. (2010). CDK12 is a transcription factor that regulates the RNA polymerase II elongation complex. Nucleic Acids Res. 37, 369–379, October 18, 2018.

Batschê, E., Yaniv, M., and Muchardt, C. (2006). The human SWI/SNF subunit Rbr is a regulator of alternative splicing. Nat. Struct. Mol. Biol. 13, 22–29.

Chathoth, K.T., Barrass, J.D., Webb, S., and Beggs, J.D. (2014). A splicing-dependent transcriptional checkpoint associated with prespliceosomal formation. Mol. Cell 53, 779–790.

Chen, M., and Manley, J.L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nat. Rev. Mol. Cell Biol. 10, 741–754.

Chen, W., Moore, J., Ozadam, H., Shulha, H.P., Rhind, N., Weng, Z., and Moore, J.M. (2018). Transcriptome-wide interrogation of the Functional Inntron by Spliceosome Profiling. Cell 173, 1031–1044.e1013.

Churchman, L.S., and Weissman, J.S. (2011). Finding the active genes in deep RNA-seq gene expression studies. Nat. Biotechnol. 29, 2784–2797.

Churchman, L.S. (2016). Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue. Cell Rep. 11, 345–355.

Churchman, L.S., Dujardin, G., Lafaille, C., de la Mata, M., Marasco, L.E., Munoz, M.J., Le Jossic-Corcos, C., Corcos, L., and Kornblitt, A.R. (2014). How slow RNA polymerase II elongation favors alternative exon skipping. Mol. Cell 54, 683–690.

Eick, D., and Geyer, M. (2013). The RNA polymerase II carboxy-terminal domain (CTD) code. Chem. Rev. 113, 8456–8490.

Harlen, K.M., Trotta, K.L., Smith, E.E., Mosheab, M.M., Fuchs, S.M., and Churchman, L.S. (2016). Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue. Cell Rep. 15, 2147–2158.

Hart, T., Komori, H.K., LaMere, S., Podshivalova, K., and Salomon, D.R. (2013). Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics 14, 778.

Hollander, D., Naftelberg, S., Lev-Maor, G., Kornblitt, A.R., and Ast, G. (2016). How Are Short Exons Flanked by Long Introns Defined and Committed to Splicing? Trends Genet. 32, 596–606.

Hsin, J.P., and Manley, J.L. (2012). The RNA polymerase II CTD coordinates transcription and RNA processing. Genes Dev. 26, 2119–2137.

Hubner, N.C., and Mann, M. (2011). Extracting gene function from protein-protein interactions using Quantitative BAC Interactomics (QUICB). Methods 53, 453–459.

Hubner, N.C., Bird, A.W., Cox, J., Splettstoesser, B., Bandilla, P., Poser, I., Hyman, A., and Mann, M. (2010). Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. J. Cell Biol. 189, 739–754.

Kohd, Y.L., Rodriguez, J., Abruzzo, K.C., Tang, C.H., Marr, M.T., 2nd, and Rosbash, M. (2011). Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. Genes Dev. 25, 2502–2512.

Kohd, Y.L., Menet, J.S., Tolan, M., and Rosbash, M. (2012). Cotranscriptional splicing efficiency differs dramatically between Drosophila and mouse. RNA 18, 2174–2186.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36.

Lund, E., and Dahlberg, J.E. (1992). Cyclic 2',3'-phosphates and nontemplated nucleotides at the 3' end of spliceosomal U6 small nuclear RNA’s. Science 257, 327–330.

Martin, M. (2011). Cutadiprem adapts sequencer sequences from high-throughput sequencing reads. EMNetJournal. https://doi.org/10.14806/efj.17.1.200.

Martinez-Rucobo, F.W., Kohler, R., van de Waterbeemd, M., Heck, A.J., Hamann, M., Herzog, F., Stark, H., and Cramer, P. (2015). Molecular Basis of Transcription-Coupled Pre-mRNA Capping. Mol. Cell 58, 1079–1089.

Mayer, A., Heidemann, M., Lidschreiber, M., Schreieck, A., Sun, M., Hintermair, C., Kremmer, E., Eick, D., and Cramer, P. (2012). CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. Science 336, 1723–1725.

Muñoz, M.J., de la Mata, M., and Kornblitt, A.R. (2010). The carboxy terminal domain of RNA polymerase II and alternative splicing. Trends Biochem. Sci. 35, 497–504.

Nieto Moreno, N., Giono, L.E., Cambindo Botto, A.E., Munoz, M.J., and Kornblitt, A.R. (2015). Mammalian NET-Seq reveals Genome-wide Nascent Transcription Coupled to RNA Processing. Cell 161, 526–540.

Nojima, T., Gomes, T., Carmo-Fonseca, M., and Proudfoot, N.J. (2015). Mammalian NET-Seq analysis defines nascent RNA profiles and associated RNA processing genome-wide. Nat. Protoc. 11, 413–428.

Nojima, T., Gomes, T., Carmo-Fonseca, M., and Proudfoot, N.J. (2016). Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide. Nat. Protoc. 11, 413–428.
STAR METHODS

KEY RESOURCES TABLE

| REAGENT OR RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| **Antibodies**      |        |            |
| Mouse monoclonal anti-Pol II CTD, Total | MBL international | Cat# MABI0601; RRID: AB_2728735 |
| Mouse monoclonal anti-Pol II CTD, phospho Ser2 | MBL international | Cat# MABI0602; RRID: AB_2747403 |
| Mouse monoclonal anti-Pol II CTD, phospho Ser5 | MBL international | Cat# MABI0603; RRID: AB_2728736 |
| Rabbit polyclonal anti-U1A | Abcam | Cat# ab155054; RRID: AB_2747404 |
| Rabbit polyclonal anti-U1C | Bethyl | Cat# A303-947A; RRID: AB_2620296 |
| Rabbit polyclonal anti-SNRP200 | Bethyl | Cat# A303-454A-T; RRID: AB_10953802 |
| Rabbit polyclonal anti-SNRP116 | Bethyl | Cat# 300-957A; RRID: AB_805780 |
| Rabbit polyclonal anti-Spt6 | Novus Biologicals | Cat#NB100-2582; RRID: AB_2196402 |
| Rabbit polyclonal anti-eIF4A3 | Bethyl | Cat# 302-981A-T; RRID: AB_10748369 |
| Rabbit polyclonal anti-SF3B3 | Bethyl | Cat# 302-508A; RRID: AB_1966103 |
| Mouse monoclonal anti-H3 | MBL international | Cat# MAB0301; RRID: AB_11142498 |
| Mouse monoclonal anti-trymethyl H3 (Lys36) | MBL international | Cat# MAB0333; RRID: AB_11126731 |
| Mouse monoclonal anti-trymethyl H3 (Lys9) | GeneTex | Cat# MAB0308 (GTX50900); RRID: AB_111242491 |

| Deposited Data | | |
|----------------|------------------|-------------------|
| Raw sequencing data | This paper | GEO: GSE106881 |
| Raw mass spectrometry data | This paper | ProteomeXchange: PXD008197 |
| Raw deposited images | Mendeley | https://doi.org/10.17632/cy8mj7sxp9.1 |
| Re-analyzed mNET-seq data | (Nojima et al., 2015; Schlackow et al., 2017) | GEO: GSE60358 and GSE81662 |
| Re-analyzed ChrRNA-seq and NpRNA-seq data | (Nojima et al., 2015; Schlackow et al., 2017) | GEO: GSE60358 and GSE81662 |

| Experimental Models: Cell Lines | | |
| HeLa (human) | Proudfoot Lab | N/A |
| TAP (mouse) | Carmo-Fonseca Lab | N/A |

| Other | | |
| TruSeq small RNA library prep kit | Illumina | Cat# 15016911 |
| NEBNext small RNA library prep kit for Illumina | NEB | Cat# E7300S |

| Software and Algorithms | | |
| Cutadapt(v1.8.3) | (Martin, 2011) | http://cutadapt.readthedocs.io/en/stable/installation.html |
| Tophat(v2.0.13) | (Kim et al., 2013; Trapnell et al., 2010) | http://ccb.jhu.edu/software/tophat/index.shtml |
| Cufflinks(v2.2.0) | (Trapnell et al., 2010) | http://cole-trapnell-lab.github.io/cufflinks/getting_started/ |
| bedtools (v2.25.0) | (Quinlan and Hall, 2010) | http://bedtools.readthedocs.io/en/latest/content/installation.html |

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to the lead contact, Nicholas Proudfoot (nicholas.proudfoot@path.ox.ac.uk).
HeLa cells and TAP cells were maintained in Dulbecco’s Modified Eagle’s Medium (DMEM) and RPMI, respectively, with 10% fetal bovine serum (FBS). Empigen treatment and in vivo splicing inhibition with Pla-B were as previously described (Nojima et al., 2015; Schlackow et al., 2017).

METHOD DETAILS

Short and long RNAs mNET-seq
Four 15 cm plates of HeLa cells (3.2x10^7 cells) were used for mNET-seq of fractionated RNAs. Immunoprecipitated samples with Pol II antibodies (MABI0602 and MABI0603) from solubilized chromatin of HeLa cells were washed eight times with 1 mL of NET-2 buffer (50 mM Tris-HCl pH7.4, 150 mM NaCl and 0.05% NP-40). 5’ phosphate is added to IPed RNAs using T4 PNK for 6 min as described before (Nojima et al., 2016). The 5’ p-RNAs were purified according to a protocol of TRI reagent (Sigma-Aldrich). The purified RNAs were run on denaturing gel (6% Novex TBE-Urea gel, Thermo-Fisher Scientific) at 300 mA for 10 min. The two RNA fractions (20-60nt and 60-160nt) were size-selected on the gel. The RNAs were put into a 0.5mL tube with three small holes that were made by 26G needle and the 0.5mL tube was put onto a 1.5mL tube. The gels were slushed by centrifuge at 16,000 g for 1 min. The slushed RNAs were incubated with 400 μL RNA elution buffer (1M NaOAc and 1mM EDTA) at 25 °C for 1 hour. The solution with gel was put into Spin X column (Coaster #8160) with 2 glass filters (Whatman #1823-010) and the column was centrifuged at 16,000 g for 1 min. The flow through (400 μL) was purified with ethanol precipitation. Finally, RNA libraries were prepared using Illumina TruSeq small RNA library prep kit (Illumina) or NEBNext small RNA library prep kit for Illumina (NEB) with minimum cycles of PCR (12~15 cycles). The expected sizes of PCR products from short and long RNAs are 150-180bp and 180-280bp, respectively. The libraries were sequenced using Hiseq4000 (Illumina) at Wellcome Trust Centre of Human Genetics, Oxford.

Mass spectrometry with filter-aid sample preparation
Four 15 cm plates of HeLa cells (3.2x10^7 cells) were used for mass spectrometry experiment. Immunoprecipitated samples with Pol II antibodies from solubilized chromatin of HeLa cells were washed eight times with 1 mL of NET-2 buffer (50 mM Tris-HCl pH7.4, 150 mM NaCl and 0.05% NP-40) and then once with 500 μL of NET buffer (50 mM Tris-HCl pH7.4 and 150 mM NaCl). After NET buffer was removed completely, washed beads were incubated in 300 μL of lysis buffer twice by centrifugation at 16,000 g for 15 min at room temperature. The samples were incubated in 200 μL of NET buffer twice by centrifugation at 16,000 g for 15 min at room temperature. The samples were incubated in 200 μL of reduction buffer (8 M Urea, 100 mM TEAB pH 8.0 and 20 mM Tris (2-carboxyethyl)phosphine hydrochloride solution pH 7.0) on the filter for 30 min, followed by centrifugation at 16,000 g for 15 min at room temperature. Next, 30 min incubation with 200 μL of alklylation buffer (8 M Urea, 100 mM TEAB pH 8.0 and 20 mM chloroiodoacacetamide) was performed in the dark place. The reaction solution was removed by centrifugation at 16,000 g for 15 min at room temperature. The samples were washed with wash buffer (1 M Urea, 50 mM TEAB pH 8.0) twice and then digested with 1 μg trypsin in 100 μL lysis buffer for 30 min at 37 °C on the filter. The trypsinization was repeated once more for over-night at 37 °C. The samples were collected by 16,000 g for 15 min at room temperature as the follow through and desalted with a C18 column. Finally, the purified peptides were analyzed by electrospray tandem mass spectrometry at the central proteomics facility in University of Oxford.

QUANTIFICATION AND STATISTICAL ANALYSIS

mNET-seq and RNA-seq data processing
Data pre-processing was performed as previously described (Nojima et al., 2016). Read Coverage, gene elements comparison and manipulation were performed using bedtools (v2.25.0) (Quinlan and Hall, 2010). Meta-profiles, splice junction average profiles and individual profiles were generated as described (Nojima et al., 2015).
Nucleoplasmic RNA-seq data analysis
Strand-specific nucleoplasmic RNA-seq data was analyzed with Tophat(v2.0.13)/Cufflinks(v2.2.0) (Trapnell et al., 2010) using Ensembl Homo sapiens GRCh37.74 release as template for quantification. zFPKM values for each transcript were converted to log2 and their distribution plotted. The threshold value chosen to identify expressed genes was the local minimum between the peak of high-expressed genes and low-expressed genes as described in (Hart et al., 2013).

Annotation
Annotations for TSS, introns, exons and alternative splicing events were retrieved from Ensembl for hg19 and mm10 genome version (Karolchik et al., 2004; Rosenbloom et al., 2015). Constitutive events were retrieved from Ensembl and U12-type spliceosomal introns retrieved from U12DB through a liftOver tool to bring all genetical information to the same reference build (Alioto, 2007).

mNET-Mass Spectrometry data analysis
Spectral counts for each specific IP along with the respective mock control were used for analysis. Missing values were imputed, and data were normalized by median subtraction as described (Harlen et al., 2016). Enriched proteins were defined using a false discovery rate (FDR) of 0.05.

snRNA read counts
A list of all ENSEMBL genes was filtered to select all genes containing U1, U2, U3, U4, U5, U6 or U7 in the respective gene symbol, grouping the genes by each snRNA category. The obtained list was then manually curated to remove incorrectly classified genes. Afterward, the number of nucleoplasm-fraction RNA reads was counted for non-overlapping snRNA genes, and the ones with a log2(RPM) equal or greater than 1 for this sample were considered expressed, constituting the gene set used in the analysis. For this final snRNA set, reads were counted and normalized to RPM for every tested sample, adding 5 bases to the annotated coordinates to account for possible annotation uncertainties, and RPM values for genes belonging to the same snRNA category were added.

mNET-seq peak finding analysis
To identify spikes in the density of 3’ ends of nascent transcripts we used an algorithm that finds nucleotides where the read density is at least three standard deviations above the mean in a local region (Churchman and Weissman, 2011). Only positions with coverage of 4 or more reads were considered. To identify peaks at the 3’ end of exons and introns indicative of splicing intermediates, the number of reads at these positions was compared to the mean read density across the corresponding exon or intron. To identify peaks corresponding to Pol II pause positions along exons and introns avoiding contamination by 5’ ss splicing intermediates, we removed reads aligning to the last 3 nucleotides of exons and the first 3 nucleotides of introns. The number of reads at each nucleotide position along the exon was then compared to the mean read density across the entire exon. Peaks were identified in the annotated exons and introns of expressed genes. Exons that intersected other isoform exons were discarded and the same principle was applied for introns.

Percent-spliced-in (PSI) calculation
To obtain a measure of splicing completeness for any given exon, the percent-spliced-in (PSI) metric (Barbosa-Morais et al., 2012) was applied to nucleoplasmic RNA-seq datasets. When calculating this index, exons that have alternative 3’ or 5’ splice sites were excluded from the analysis, as well as exons overlapping other exons on the same strand and exons with less than 100bp, or which the preceding intron had less than 100bp. Finally, only junctions that had a total of at least 5 reads in both regions were included in the analysis.

Identification of spliced events
Detection of spliced events was based on spliced reads obtained in mNET-seq/SSP experiments. Reads that aligned uniquely and with no mismatch across two exons were classified as spliced reads. To reduce the possibility of fortuitous alignments, only reads that map to more than 8 nucleotides across any junction were considered. A splicing event was considered if the 3’ SS was covered by 3 or more reads.

Splicing associated with 5’ ss peak
Merely spliced or unspliced mNET-seq/SSP reads that specifically align to the last nucleotide of exons were taken into consideration. Calculation of splicing percentage was performed, on exons with a total between spliced and unspliced reads equal or above 10, in the following manner:

\[\text{spliced\%} = \frac{SR}{SR + UR} \times 100\]

“SR” stands for spliced reads and “UR” unspliced reads. Significance of the differences between the mean and distribution of the percentage of splicing for skipping and constitutive exons was calculated through a Wilcoxon signed-rank test and Kolmogorov-Smirnov tests, respectively.
**mNET-seq read density analysis**

For each expressed exon and intron, that do not intersect with another exon or intron isoform respectively, read average density was calculated as follows:

\[
RAD = \frac{TR - 3nt}{EL - 3nt}
\]

“TR” stands for the total number of reads along the exon and “EL” to the exon length. The last 3 nucleotides of exons and the first 3 nucleotides of introns are removed because of the high read densities associated with 5’ ss splicing intermediates. For each group of exons considered, exons without reads were not considered and the signal originating from spliced reads was removed.

**DATA AND SOFTWARE AVAILABILITY**

All the sequencing data and mass spectrometry data used in this work is submitted at GEO: GSE106881 and ProteomeXchange: PXD008197. Original images of western blot are available at Mendeley data [https://doi.org/10.17632/cy8mj7sxp9.1](https://doi.org/10.17632/cy8mj7sxp9.1). Software’s used in this work are all publicly available, with the links to them in the above tables.