Universidade de Lisboa

Faculdade de Ciências

Departamento de Informática

Genome-wide profiling of RNA polymerase II and associated co-transcriptional processes using advanced NET-seq data

Tomás Pires de Carvalho Gomes

Dissertação

Mestrado em Bioinformática e Biologia Computacional

Especialização em Biologia Computacional

2014

Universidade de Lisboa

Faculdade de Ciências

Departamento de Informática

Genome-wide profiling of RNA polymerase II and associated co-transcriptional processes using advanced NET-seq data

Tomás Pires de Carvalho Gomes

Dissertação

Mestrado em Bioinformática e Biologia Computacional

Especialização em Biologia Computacional

Orientadores:

Doutora Ana Rita Fialho Grosso

Prof. Doutora Lisete Maria Ribeiro De Sousa

2014

Agradecimentos

Este foi um ano preenchido, de trabalho e pessoas.
Quero começar por agradecer à Doutora Ana Rita Grosso pela disponibilidade constante, pelo seu apoio crítico e pelo que me ensinou, não só este ano como também durante a sua cadeira no mestrado. Mas também pelos incentivos e atitude positiva que ofereceu, quer o projeto corresse melhor ou pior.
À professora Maria do Carmo Fonseca, pelas suas ideias, discussão crítica, pelo delinear do projeto, e também pelo entusiasmo constante demonstrado pelos resultados obtidos.
À Ana Paula Leite, que me trouxe inicialmente para este projeto, e que embora me tenha orientado pouco tempo me deu conselhos valiosos.
Agradeço a todos os meus colegas de gabinete, Célia, Bruno, Robert, Paco, Marina, Jorge, pela hospitalidade na introdução no grupo e no mundo científico, através de discussões e reflexões científicas, sociais e por vezes ambas. E estendo o agradecimento a toda a equipa, todos com trabalhos interessantes em diversos temas, o que me abriu vários horizontes de investigação.
Obrigado também à professora Lisete de Sousa, principalmente pelas aulas e tudo o que ensinou, mas também pela coorientação deste trabalho.
Uma nota especial para a comunidade mundial de bioinformáticos/biólogos computacionais, que com a sua grande partilha de problemas e ideias online me ajudou com muitas dúvidas.
Agradeço também a um valioso grupo de amigos. Precisaria de outra tese para indicar um a um quem são e como contribuíram, não só neste ano, mas nos quatro anteriores também, e alguns até antes. Mas os últimos cinco anos foram verdadeiramente de grande mudança de vida, e muito se deveu a vocês.
À minha família, pelo acompanhamento que sempre me deram. Ao meu pai pelo constante interesse e incentivo a continuar a fazer o que gosto; à minha mãe pela constante preocupação e dedicação sobre o meu trabalho; ao meu irmão pela distração no dia-a-dia; aos meus avós pela sua sabedoria e paciência desde sempre; à minha tia Fernanda pelo apoio e encorajamento a procurar fazer mais e melhor. Por tudo, no fundo, desde que me lembro.
Por último, e em jeito de dedicatória, um agradecimento à Hajrabibi, que tem a distinta característica de saber sempre o que dizer quando é preciso. Que eu te possa ajudar tanto ou mais.

Nota Prévia

A presente tese encontra-se escrita em inglês e em formato de publicação. A língua inglesa foi escolhida para ser usada por ser hoje em dia a língua franca da comunidade científica, e ao pretender seguir uma carreira de investigação impõe-se a necessidade de um domínio crescente dessa língua.

Adicionalmente, o inglês é também utilizado devido ao formato de publicação científica desta tese. O projeto desenvolvido ao longo do último ano resultou num artigo científico submetido à revista *Cell*. Como tal, a presente tese inclui não só resultados das análises computacionais realizadas, mas também validações biológicas experimentais complementares ao trabalho desenvolvido, que permitem uma melhor compreensão da questão biológica abordada. No entanto, este relatório pretende salientar o trabalho desenvolvido pelo autor no referido projeto, isto é, a análise de dados de sequenciação de alto rendimento usando software adequado a cada teste, como será descrito mais à frente. Os capítulos desta tese, incluindo não só o Capítulo 2, que contém o artigo, mas também os restantes, estão assim escritos em formato idêntico ao utilizado na submissão de manuscritos à referida publicação, mas com a inclusão de figuras ao longo do texto, a fim de facilitar a leitura e compreensão. Para facilitar a compreensão, foi também dado um estilo diferente aos títulos e subtítulos.

Resumo

A transcrição é o processo, presente em todos os seres vivos, em que a partir de uma cadeia molde de DNA se sintetiza uma cadeia complementar de RNA. A grande maioria dos genes em eucariotas é transcrita pela RNA polimerase II. A cadeia de RNA sintetizada não é, no entanto, o produto final, já que pode ser alvo de vários tipos de processamento, como splicing, poliadenilação ou edição de bases. Estes fenómenos foram já descritos como ocorrendo co ou pós-transcricionalmente. No entanto, não são ainda conhecidos todos os componentes, nem como são regulados estes processos ou qual a sua interação com a RNA polimerase II, em particular com o seu domínio carboxi-terminal (CTD).

Para abordar estes problemas de uma forma não enviesada, optou-se por adaptar uma técnica anteriormente descrita, que abrange todo o genoma, de alto rendimento e precisão, a *native elongating transcript sequencing* (NET-seq); sendo ela modificada de modo a poder detetar qual o estado de fosforilação do domínio carboxi-terminal da polimerase isolada em cada ensaio. Ao novo protocolo chamou-se *advanced NET-seq* (ANET-seq). Para além dos dados gerados por este protocolo, foram também obtidos dados de RNA ligado à fração de cromatina (ChrRNA). Todos os dados foram obtidos de células HeLa, sendo esta a primeira instância em que um estudo de nível genómico com esta precisão de mapeamento foi aplicado em mamíferos.

Análise inicial destes dados revelou uma distribuição das isoformas do CTD nos genes idêntica ao previamente descrito por outras técnicas. Adicionalmente, verificou-se também a captura de precursores do splicing, nomeadamente do 3' do exão upstream, distintamente nos casos em que este é incluído no transcrito final. Estes exões aparecem principalmente associados a polimerase fosforilada na serina 5 do seu CTD. Outra observação curiosa foi a deteção de precursores do processamento de micro RNAs pelo complexo Drosha/DGCR8. Diferenças na deteção destes precursores permitiu postular diferentes dinâmicas para o processamento destes RNAs não codificantes.

Também se obtiveram dados de ANET-seq (com anticorpo para fosforilação da serina 2) e ChrRNA de células HeLa transfetadas com siRNA contra fatores de terminação – Xrn2 - e processamento do terminal 3' do pre-mRNA – CPSF73 e CstF64+CstF64τ. Análise destes dados permitiu concluir que os fatores de processamento, mas não o Xrn2, influenciam significativamente a dinâmica da polimerase na região 3' do gene, no final da transcrição, promovendo a sua pausa e subsequente desassociação do DNA. Constatou-se também que

estes fatores afetam a acumulação de polimerase junto ao promotor dos genes, afetando igualmente a produção de transcritos upstream do promotor (PROMPTs), podendo concluir-se que estes fatores participam na regulação da transcrição não-produtiva.

Os resultados obtidos foram satisfatórios e também surpreendentes. Com este trabalho, é apresentada uma nova forma de estudar, ao nível do genoma, como ocorre a regulação da transcrição pelo CTD. Mostram-se também novas provas sobre processamento co-transcricional do RNA e a sua ligação à fosforilação do CTD. Foram igualmente elucidados os papéis de alguns fatores envolvidos na fase final da transcrição. Finalmente, ficou outra vez demonstrada a importância de estudos abrangentes na área da transcrição, em complemento dos trabalhos moleculares e bioquímicos já desenvolvidos há décadas. Espera-se, de futuro, um aprofundamento das técnicas de alto rendimento, e uma consequente adequação das ferramentas bioinformáticas a estes estudos.

Palavras-chave: transcrição; ANET-seq; sequenciação de RNA; CTD; splicing; micro RNA; clivagem e poliadenilação; terminação.

Abstract

Transcription is a process present in all living beings where, from a DNA template, a complementary RNA strand is synthesized. Most eukaryotic genes are transcribed by RNA polymerase II. The resulting RNA strand is not, however, the final product, since it'll still be subject to various processing steps, such as splicing, polyadenylation or base editing. These modifications have been described as occurring co or post-transcriptionally. Yet, it is still not known how these processes are regulated, nor what all of their interveners are or how do they interact with RNA polymerase II, in particular with its C-terminal domain (CTD).

To address these problems in an unbiased way, a previously described genome-wide and high-precision technique, *native elongating transcript sequencing* (NET-seq), was adapted so it could detect what was the phosphorylation isoform from the isolated polymerase's CTD. The new protocol was called *advanced NET-seq* (ANET-seq). In addition to the data generated by this protocol, RNA associated with the chromatin fraction was also sequenced. All data was obtained from HeLa cells, applying this genome-wide high-resolution technique to a mammalian system.

Initial analysis of ANET-seq data revealed that distribution of CTD isoforms in genes was similar to previously described profiles obtained by other protocols. Additionally, it was also verified the capture of splicing intermediates, in particular the 3' end of the upstream exon, distinctively in cases where it was included in the final transcript. These exons are mainly associated with polymerase phosphorylated in the CTD's Ser5. Another curious observation was the detection of micro RNA precursors, resulting from Drosha/DGCR8 processing. Differences in the detection of these precursors allowed the proposal of different processing dynamics for this type of non-coding RNAs.

ANET-seq data (with a Ser2-directed antibody) and ChrRNA from HeLa cells transfected with siRNA for termination factor Xrn2 and 3' processing factors CPSF73 and CstF64+CstF64τ were also obtained. The analysis of this data showed that 3' processing factors, but not Xrn2, significantly influence Pol II dynamics in the gene's 3' region, at the end of transcription, promoting its pause and dissociation from the DNA template. It was also observed that these factors influence polymerase accumulation near gene's promoters, and equally affect promoter upstream transcripts (PROMPTs), leading to the conclusion that these factors regulate termination of unproductive transcription.

Obtained results were satisfactory and also sometimes surprising. This work presents a novel genome-wide way to study how transcription is regulated by the CTD. New evidence of co-transcriptional RNA processing arouse, as well as their connection with CTD isoforms. There were also new revelations about transcription termination factor's functions. Finally, it was once again demonstrated the importance of genome-wide techniques in transcription study, which complete molecular and biochemical work in the same area that has been developed for decades. In the future, a greater development of high-throughput techniques, and a constant adaptation of bioinformatical tools to these studies is expected.

Keywords: transcription; ANET-seq; RNA sequencing; CTD; splicing; micro RNA; cleavage and polyadenylation; termination.

Table of Contents

Chapter 3

Chapter 1

1. Transcription

1.1 Overview

In 1956, Francis Crick first proposed what he called the "Central Dogma of Molecular Biology". The Dogma not only stated, in his own words, that "Once information has got into a protein it cannot get out again", but also outlined the possible ways this information would be transferred between nucleic acids and proteins. Later, in 1970, Crick developed these ideas, and classified the nine possible ways information could be transferred between the intermediates (DNA, RNA and protein) into three categories: General Transfers, Special Transfers and Unknown Transfers (Francis Crick, 1970) (Figure 1). General Transfers refer to reactions present in all cells, whereas Special Transfer refers to reactions that were postulated to exist, and later identified only in a subset of life forms or *in vitro* (Uzawa et al., 2002). Unknown Transfers are reactions postulated not to exist since they would require very complex machinery. Although great advancements and discoveries have been made in the field of molecular biology, the core message of the dogma still holds, yet it does not address certain details of the described phenomena, such as gene expression regulation or post-translational modifications.

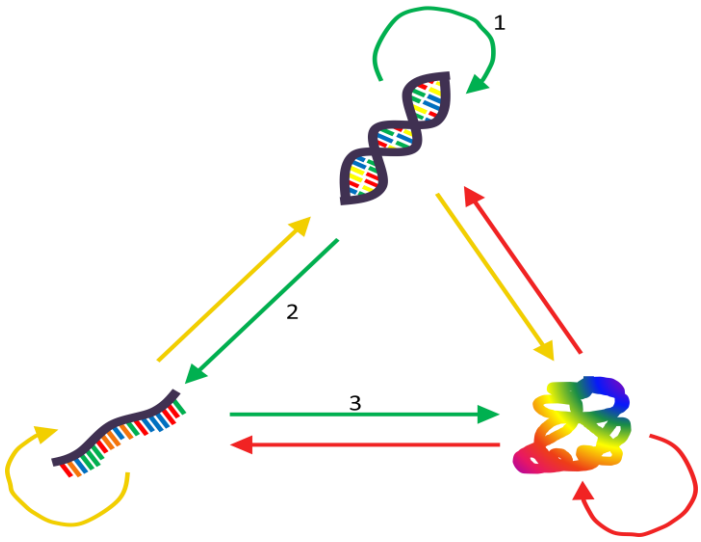


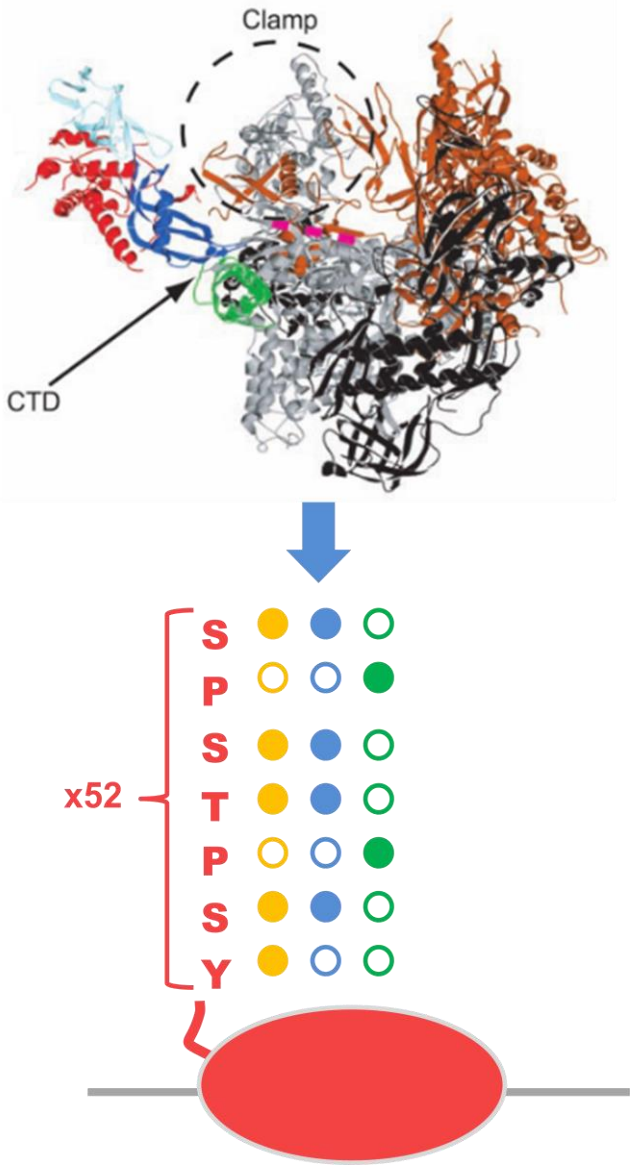
Figure 1: A representation of the Central Dogma of Molecular Biology, including DNA (top), RNA (bottom left) and protein (bottom right), and the possible information transfers between them (arrows). Green, General Transfers; Yellow, Special Transfers; Red, Unknown Transfers. 1 – DNA replication; 2 – Transcription; 3 - Translation

According to the Dogma, information is stored in the DNA nucleotide sequence, and to be effectively used to produce proteins uses an intermediary, RNA. Transcription is the synthesis of RNA using DNA as a template. Although the interveners vary greatly between prokaryotes and eukaryotes, the core process is very similar. The main player in transcription is RNA polymerase, which reads the DNA strand and synthesizes a complementary RNA molecule (Chamberlin and Berg, 1962). Like any complex biochemical reaction, eukaryotic transcription can be divided in several steps. These steps are defined by the different factors that associate with the polymerase and the transcribed gene's sequence in a given moment.

The progression of transcription in each phase, however, depends not only on these factors, but also on chromatin conformation and components, and also the gene sequence (Nojima et al., 2013; Grosso et al., 2012; Peterlin et al., 2006; Jonkers et al., 2014). In addition, eukaryotes possess different RNA polymerases, each responsible for the transcription of a subset of genes. This results in a highly regulated process, allowing the cell to precisely adjust its components' concentration in response to diverse stimuli.

1.2 The RNA Polymerase II and the C-Terminal Domain

As previously mentioned, RNA polymerase is the main agent involved in transcription, synthesizing RNA in a DNA-dependent manner. It is an essential enzyme for all organisms - even virus, which may use the host's polymerase -, but despite that there are many differences, mainly structural but some also functional, between prokaryotes and eukaryotes.



In animals, three RNA polymerases exist (Roeder and Rutter, 1969). RNA polymerase I (Pol I) is responsible for transcription of pre-45S rRNA, which generates all the other mature rRNA except 5S (Jacob, 1995), whereas RNA polymerase III (Pol III) is involved in the production of tRNAs, rRNA 5S, a small subset of micro RNAs and other small RNAs found in the nucleus and cytosol (Weinman and Roeder, 1974; Willis, 1993). As for all the other transcripts,

Figure 2: Top - Back view of the RNA polymerase II complex structure, and a scheme of the CTD sequence. RPB1 in grey, RPB2 in bronze, RPB4 in red, RPB6 in green, RNP domain of RPB7 in blue, C-terminal of RPB7 in light blue and the rest in black. Structure from Bushnell and Kornberg, 2003. **Bottom** – CTD post-translational modifications. Filled circles indicate existence, open circles indicate inexistence. Yellow, phosphorylation; Blue, glycosylation; Green, proline cis-trans isomerization.

including all mRNAs, their production is attributed to RNA polymerase II (Pol II).

Human Pol II is a 550kDa protein complex, composed of 12 different subunits (Acker et al., 1997). The whole complex is highly conserved among eukaryotes, and most of its subunits are interchangeable among species without any prejudice for transcription (Shpakovski et al., 1995). Therefore, many structural and functional studies are conducted using yeast Pol II, considered the archetype for eukaryotic RNA polymerases. Some subunits have a function of their own, whereas others interact to give rise to a function, as is the case for the subunits that constitute the active site (Acker et al., 1997; Woychik and Hampsey, 2002). In addition to the enzyme itself, there are also other components that constitute the RNA polymerase II holoenzyme (Myer and Young, 1998). The holoenzyme is the complex recruited to eukaryotic promoters, including the core enzyme and the proteins that recognize promoters or enhancers, and also factors whose function is to remodel the chromatin, allowing transcription to proceed.

From the 12 subunits that make Pol II, RPB1 is the largest, and, in interaction with others, constitutes part of the enzyme's active site (Cramer et al., 2001). But RPB1 has other important regions, like its C-Terminal Domain (CTD) (Figure 2). This is a structurally disordered region, composed of a repetition of the heptapeptide Tyrosine-Serine-Proline-Threonine-Serine-Proline-Serine (Y-S-P-T-S-P-S). Although the heptapeptide itself is highly conserved among eukaryotes, the number of tandem repetitions varies greatly, from 26 in *Saccaromyces cerevisiae*, to 42 in *Drosophila melanogaster*, 34 in *Arabidopsis thaliana* and 52 in vertebrates. The CTD amino acids serve as targets for reversible post-translational modification of Pol II. These changes are intrinsically linked to the dynamics of transcription and its associated phenomena, yet some modifications assume more preponderant roles than others in the progression of transcription. The most common modifications are the phosphorylation of Ser2 or Ser5 (Phatnani and Greenleaf, 2006), but other residues can be modified in various ways (Figure 2, bottom).

The main, general role of the CTD, along with its modifications, is to act as a scaffold, recruiting different interveners of transcription and RNA modifiers. The post-translational modifications are the cause behind the multitude of CTD interactions, allowing for a fine tune of RNA synthesis and modification. All amino acids of the heptad have modifications associated to them, and although some of these are mutually exclusive (phosphorylation and glycosylation, for example), the number of possibilities allows for a wide range of combinations. In addition, other non-consensus residues may also be modified, as is the case for arginine 1810 methylation (Sims et al., 2011), which regulates CARM1 activity, involved

in some snRNA and snoRNAs expression. Non-consensus lysines were also demonstrated to be the target of an ubiquitin-protein ligase in mice (Li et al., 2007).

From the most studied marks – serine phosphorylations – the first to appear in a gene's transcription is Ser5, highly associated with the promoter, although it can still be found in the rest of the gene. This mark has been particularly linked to 5' capping, H3K4 trimethylation and early termination events (Terzi et al., 2011; Komarnitsky et al., 2000). Ser7 phosphorylation is also an early mark in transcription, but it generally extends further than Ser5, despite their ChIP pattern being very similar (Kim et al., 2009). Ser7 is associated with the Integrator machinery, responsible for snRNA processing (Egloff et al., 2007). However, Ser7 is a less conserved position in the CTD heptad, sometimes replaced by arginine or lysine. Usually after promoter clearance, Ser2 phosphorylation begins to be observable. This does not mean that other marks disappear, since it is well described the double marking Ser2P-Ser5P along the gene body, and is responsible for recruiting SET2, inducing the methylation of H3K36. Phosphorylation of Ser2 is carried out by CDK9, a Ser5P-dependent kinase part of the P-TEFb complex, but only when there's a relation with splicing and termination events (Napolitano et al., 2013). CDK9-driven phosphorylation was once also thought to be related to transition to productive elongation, but it in fact drives such transition by catalyzing the phosphorylation of SPT5, a subunit of the DSIF complex (Garber et al., 2000). It is now believed that CDK12 is the kinase responsible for elongation-associated phosphorylation of the CTD (Bartkowiak et al., 2010).

Other CTD modifications seem to have more specific roles, and consequently they're function is not well known or studied. Thr4 phosphorylation, for example, is known to be involved in histones 3'end processing (Hsin et al., 2011). Glycosylation is also not very well described, but it is postulated to regulate phosphorylation, as the two are mutually exclusive.

The fact that so many transcription-related processes seem to have elements interacting with the CTD repeats of the largest Pol II subunit makes them a prime target for studies in transcription regulation and dynamics. But, in spite of the knowledge gained from genome-wide ChIP studies about where in the gene each CTD isoform appears, it is still fairly misunderstood when in a gene's transcription the phosphorylation/dephosphorylation of some of these amino acids happens, in particular the widely studied Serine 2 and Serine 5 phosphorylations. It can be concluded that a technology that can map these CTD patterns with elevated precision and in a genome wide fashion is essential to reveal their relationship with gene sequence, co-transcriptional processing or regulation events.

1.3 Stages and players of transcription

It is possible to define at least eight steps for the whole transcription process in eukaryotes (Fuda et al., 2009): chromatin opening, pre-initiation complex formation, initiation, promoter clearance, escape from pausing, productive elongation, termination and recycling. These can also be summarized in initiation (comprising the aforementioned initiation, promoter clearance and escape from pausing), elongation and termination, in order to highlight the beginning, development and end of the RNA molecule synthesis. As previously stated, these stages are characterized by specific elements, resulting in a fine regulation of transcription.

Initiation, the first stage of active transcription, depends on the opening of chromatin and pre-initiation complex assembly. Chromatin opening consists on unwinding DNA from nucleosomes, mainly by histone acetylation, a modification very early described to promote RNA synthesis (Allfrey et al., 1964; Hebbes et al., 1988). Conversely, gene silencing is usually promoted by histone methylation (Chen et al., 1999), as is the case for H3K9me3 histone mark (for silenced promoters), but not for H3K4me3 (active promoter, present at the transcription start site), H3K36me3 and H3K79me2 (active gene body), which collectively indicate the presence of an actively transcribed gene (Kouzarides, 2007). After chromatin remodeling, pre-initiation complex assembly – which corresponds to RNA polymerase and general transcription factors – occurs, according to the core promoter elements present, and is regulated by distal and proximal enhancers (Stargell and Struhl, 1996).

Although pre-initiation complex assembly makes Pol II essentially ready to start transcribing, it won't always occur. Many times initiation is a rate-limiting step in transcription, resulting in an accumulation of RNA polymerase at the transcription start site (TSS). Initiation can be regulated in the open complex formation step, promoter clearance by dethatching from pre-initiation complex factors, or escape from promoter-proximal pausing (Saunders et al., 2006). In particular, promoter-proximal pausing is responsible for most of the accumulation of polymerase in the TSS region. This stage is known to be regulated by P-TEFb, a complex that includes a Ser2 kinase for the C-terminal domain (CTD) of RNA polymerase II (Pol II), and that enables its transition to productive elongation (Ni et al., 2008). P-TEFb not only phosphorylates Pol II, but also some factors that contribute negatively to elongation, such as DRB Sensitivity Inducing Factor (DSIF). Shortly, DSIF interaction with the Negative Elongatin Factor (NELF) and Pol II is disrupted by the kinase activity of P-TEFb, thus allowing for the polymerase to advance, with a hyperphosphorylated CTD (Yamaguchi et al., 1999). This promoter-pausing regulation mechanism allows for a fast

response to environmental changes in terms of gene expression, as was attested by the description of this mechanism in *Drosophila melanogaster* hsp70 gene (Boehm et al., 2003).

During elongation, fewer factors seem to be involved in transcription regulation. However, polymerase progression is not constant. Productive elongation requires chromatin remodeling by removing nucleosomes out of the way (Orphanides et al., 1998; Belotserkovskaya et al., 2003). More recently, it has been demonstrated that pausing is highly correlated with nucleosomes and sequence (Chruchman and Weissman, 2011, Grosso et al., 2012), and also that elongation rates are correlated to GC content, DNA methylation and exon density, suggesting a connection to splicing (Jonkers et al., 2014).

Transcription termination is the hardest phase of transcription to study, due to its many interveners, its variability between genes and the difficulty to establish an *in vitro* system that replicates it. Nevertheless, it has been described that transcription proceeds after the polyadenylation (pA) site, peaking on average about 1.5kb after this sequence (Core et al., 2008). However, evidence also shows that this is not a general feature, and depends on the gene's transcription rate and magnitude (Grosso et al., 2012). A more detailed and mechanistic description will be presented next.

2. Pre-mRNA Processing

2.1 Splicing

When the Human Genome Project started in 1990, no one would still believe that the human genome contained the 6.7 million genes proposed in 1964 by Friedrich Vogel. However, the estimate at the time would still be about 5 times larger than the most recent number of about 19000 (Pertea and Salzberg, 2010; Ezkurdia et al., 2014). More so, being this value very close to the predicted number of genes in other invertebrates (and, in general, less complex life forms), there was a realization that phenotype diversity wasn't that much dependent of protein-coding genes number. However, protein diversity – which accounts for part of phenotype diversity, together with expression regulation - can be achieved by other means, such as post-translational modification and alternative splicing. It has been shown that alternative splicing patterns divergence has a relevant role in determining differences between vertebrate species (Barbosa-Morais et al., 2012). Exon skipping seems to be the most prevalent form of alternative splicing in this clade – especially in humans –, and more

relevant than in invertebrates (Kim et al., 2007), but there can are also be other types, like alternative donor or acceptor sites and retained introns (Sammeth et al., 2008).

But not all introns are subjected to alternative splicing, meaning that generating diversity is not the sole reason why introns exist. While their origin is highly debatable, they are maintained in large genomes because the organism can support their energetic cost and because they are not very disadvantageous, even when suffering insertions or deletions (Lynch and Conery, 2003). Introns can then be made useful to genomes, as is the case with alternative splicing previously explained. Introns are also associated with many non-coding RNAs (ncRNA), which can be excised after splicing occurs (Rearick et al., 2011), and with regulatory roles (Jonsson et al., 1992; Hughson and Schedl, 1999).

Figure 3 shows how the known types of splicing can be organized. Most introns depend on the spliceosome for their excision, but some are capable of it by themselves, through more or less similar mechanisms. Trans-splicing is a distinct case in spliceosome-dependent splicing, occurring only in a restricted number of species, and it involves splicing together two exons from different genes (Bonen, 1993).

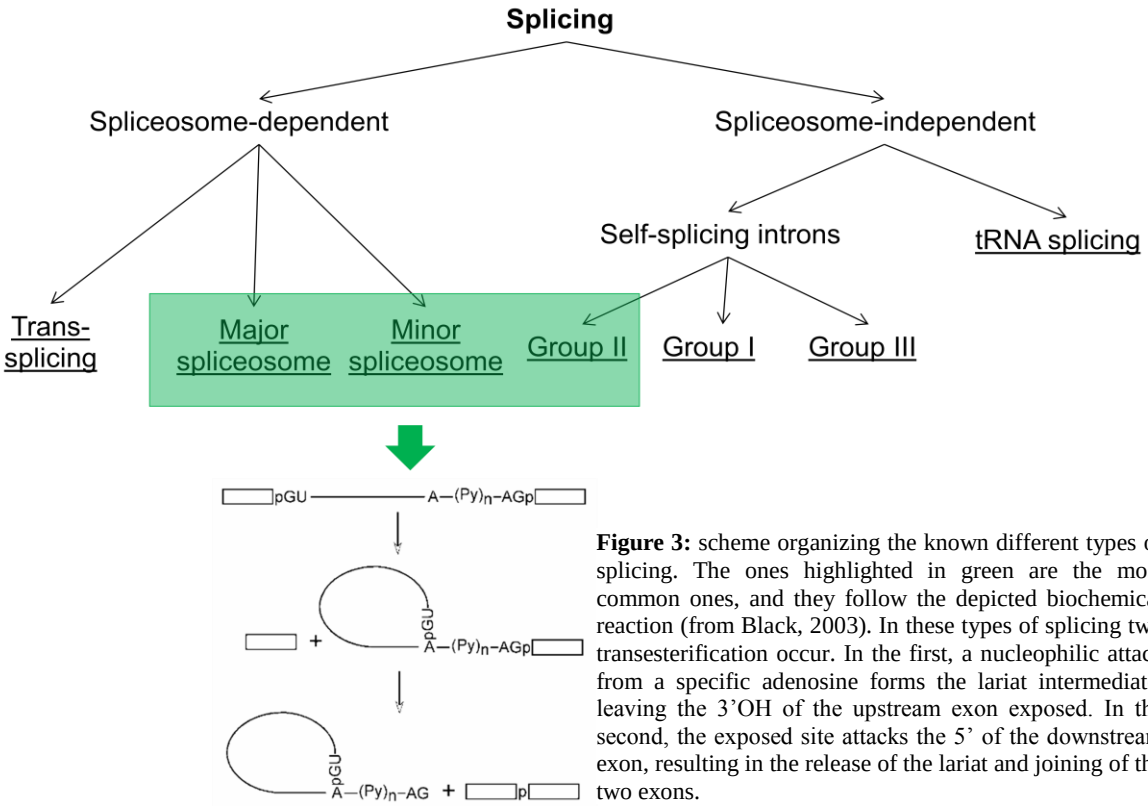


Figure 3: scheme organizing the known different types of splicing. The ones highlighted in green are the most common ones, and they follow the depicted biochemical reaction (from Black, 2003). In these types of splicing two transesterification occur. In the first, a nucleophilic attack from a specific adenosine forms the lariat intermediate, leaving the 3'OH of the upstream exon exposed. In the second, the exposed site attacks the 5' of the downstream exon, resulting in the release of the lariat and joining of the two exons.

Most introns undergo splicing through the major spliceosome pathway. The major spliceosome is composed of several ribonucleoproteins (RNPs), consisting in associations of one or two small nuclear RNA (snRNA) and proteins. The snRNA that are part of this structure are U1, U2, U4, U5 and U6. U4 and U6 are assembled together in the same RNP,

whilst the others are each on a different RNP. The association between RNPs and the intron varies during splicing, as well as the conformation of the snRNA and proteins (Will and Lührmann, 2011) (Figure 4). This process depends on the conservation of the components of the spliceosome, but also sequence components of the intron. A mutation in the branch point or any of the other conserved sequences will result in defective splicing (Reed and Maniatis, 1988; Talerico and Berget, 1990). This is also one of the keys to alternative splicing, as certain factors may enhance the detection of weaker splice sites, i.e., sequences that are only partially similar to the canonical splice sequences (Guiner et al, 2001).

The minor spliceosome is responsible for the splicing of only about 1 in every 300

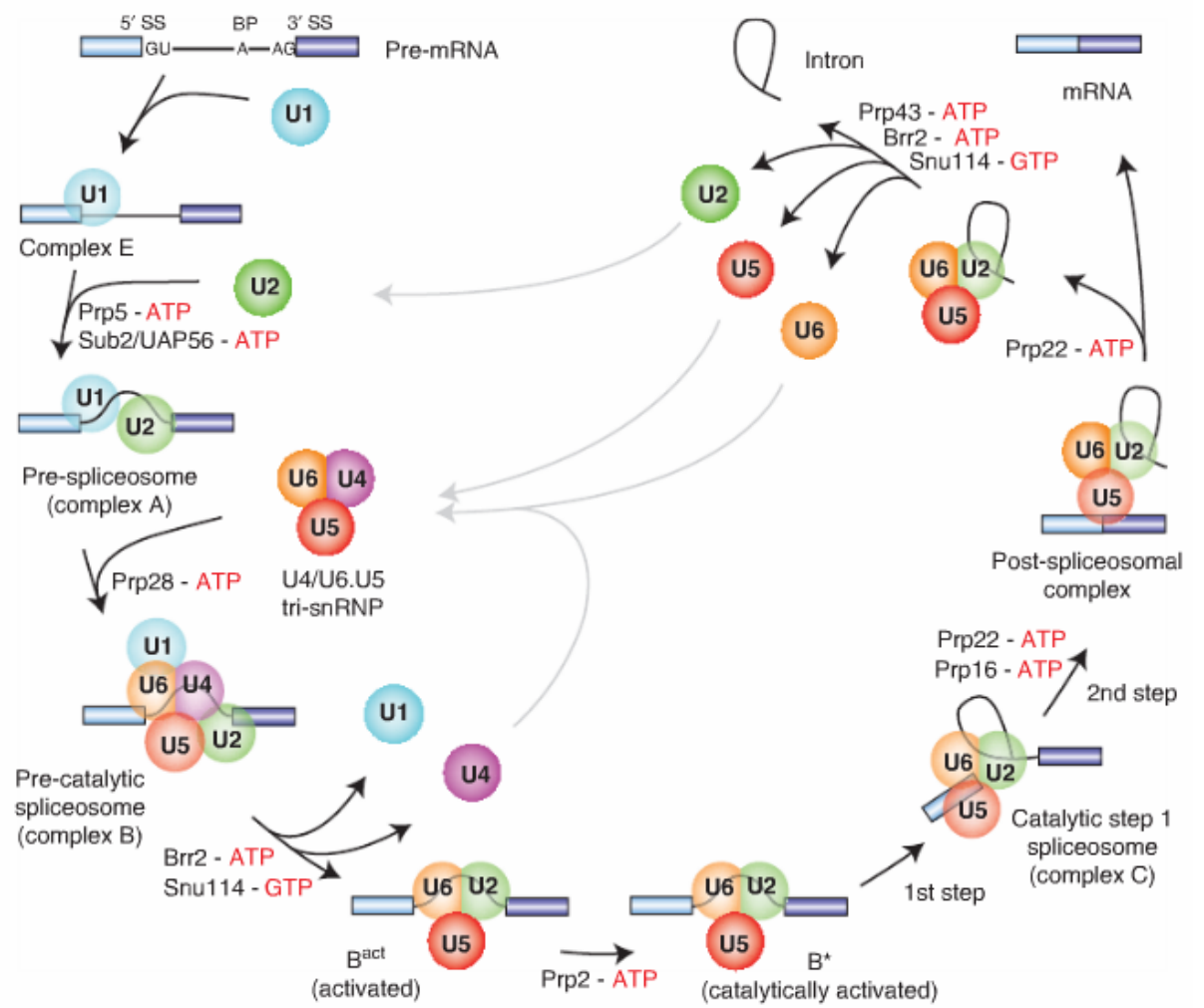


Figure 4: Assembly dynamics of the major spliceosome during intron splicing, highlighting interactions between snRNPs and the pre-mRNA sequence. From Will and Lührmann, 2011

introns (Steitz et al., 2008). This spliceosome contains the unique snRNA U11, U12, U4atac and U6atac, and also shares the U5 snRNA with the major spliceosome (Tarn and Steltz, 1996). This pathway is also characterized by splicing AT-AC introns, which have different conserved sequences. Despite the differences, the mechanism employed is very similar to the major spliceosome. There is a functional equivalence between U1 and U11, U2 and U12, U4

and U4atac, and U6 and U6atac (Will and Lührmann, 2005). However, the minor spliceosome includes some proteins not involved in RNP complexes.

It has been shown that splicing can occur not only post-transcriptionally, but also co-transcriptionally (Beyer and Osheim, 1988). Although there is much evidence that this phenomenon is widespread, and that the spliceosome actually co-localizes with nascent transcripts (Lacadie et al., 2006), it has proven difficult to establish how it is regulated. Like other transcript modifications (such as capping and 3' end processing) there is a known correlation between co-transcriptional splicing and CTD modification (Fong and Bentley, 2001), but the exact interaction with the spliceosome is not known. Some evidence, although not definite, points to the recruitment of spliceosome RNPs to the nascent RNA by interaction of these with newly synthesized splicing signals and by elongation factors, this last point explaining the correlation with the CTD dynamics (Neugebauer, 2002). Evidence also points to there being no distinction between splicing of constitutive or alternative exons happening co or post transcriptionally, although there seems to be some lag in transcription and splicing (Johnson et al., 2000; Pandya-Jones and Black, 2009). In another recent study, spliced intermediates were sequenced together with nascent transcripts associated with polymerase in yeast, allowing for a new way of studying these events (Churchman and Weissman, 2011). Novel approaches are now needed to understand transcription and splicing's mutual influence.

2.2 3' end processing and transcription termination

Compared with transcription initiation, less is known about termination. This is in part due to difficulties in studying termination, since it requires handling nascent transcripts, and also because of some neglect for being a process happening downstream of the encoding region, and hence it could be concluded to not have any role in gene expression regulation. Termination is characterized by detachment of Pol II from the DNA template, after transcription of the polyadenylation (pA) site. Since it was discovered that an intact pA site was essential for transcription termination (Connelly and Manly, 1988), evidence for a connection between termination and pre-mRNA 3' end processing - which includes cleavage and polyadenylation (CPA) – has been increasing (Proudfoot et al., 2002).

3' end processing mechanisms depend on large protein complexes (Shi et al., 2009), with elements involved in binding to specific RNA sequences, cleaving the RNA and polyadenylation of the new transcript's generated end. Recognition of the AAUAAA motif, which has long been proven to be required for cleavage and polyadenylation (Zarkower et al.,

1986), is performed by the Cleavage and Polyadenylation Specificity Factor (CPSF), a complex with five subunits, including the endonuclease CPSF73. Downstream of this motif, the Cleavage stimulation Factor, specifically its subunit CstF64, will bind to a U/GU rich motif. These two factors interact with each other through various other elements of the complex, ultimately promoting the endonucleolytic activity of CPSF73 between the two motifs, 10 to 30 bases downstream of the pA site (Liu et al., 2007; Mandel et al., 2006; Lutz, 2008). This process and its components are highly conserved in eukaryotes but, despite being functionally very relevant, it is interesting to point out that CstF64 has a redundant role, since another protein, CstF64τ, is capable of performing the same function. Both proteins are very conserved, and seem to have different affinities with their interaction partners, but the biological reason for this functional duplication is not still fully understood (Yao et al., 2013).

In the late 1980's, two models emerged to explain transcription termination. The allosteric model (Logan et al., 1987) postulates that transcription of the pA site leads to conformational changes in Pol II or associated elongation factors, which causes dissociation of said factors and/or the association of termination factors, leading to 3' end processing and downstream pausing of the polymerase after the release of the downstream transcript. The torpedo model (Connely and Manly, 1988) advocates for a termination-dependent degradation of the transcript downstream of the cleavage site by an exonuclease (the "torpedo"), later revealed to be Xrn2 (West et al., 2004). This enzyme's activity requires a 5' entry point, which is generated by co-transcriptional cleavage (CoTC), a process in which RNA cleaves itself once it is transcribed (Teixeira et al., 2004). However, CoTC activity was only so far identified in a small subset of genes (Nojima et al., 2013), making it hard to generalize this mechanism for now. But importantly, there is evidence that the two mechanisms may act together, since it has been described that cleavage after the pA can happen with Pol II still bound to the template, and that degradation by Xrn2 precedes the polymerase release from the template (West et al., 2008). This implies that pA site recognition is needed for the success of termination, hence pointing to a mixed model. Finally, it is also worth referring that pausing after the pA site might also play a role in termination by slowing down the polymerase.

The above descriptions of transcription termination and 3' end processing refer to protein-coding genes in general, but replication-dependent histones are a notable exception. Histones are a highly conserved class of proteins responsible for chromatin packing in nucleosomes. They are subject to modifications, leading to very diverse roles in transcription regulation. Their genes are especially upregulated at the start of the S phase (Stein et al., 2006) because of DNA synthesis. Genes coding for replication-dependent histones are typically less than

2000 base pairs and intronless. The promoter region of H4 histone gene – the most studied – has regulatory sequences unique to other protein-coding genes (Ramsey-Erwing et al., 1994), but studies in *Drosophila* suggest that not all genes from this family are regulated by the same factors (Isogai et al., 2007). Similarly to other genes, histone mRNA has a 5' 7-methylguanosine cap. However, these transcripts are not polyadenylated, relying instead on an exclusive 3' end processing mechanism. It involves an RNA hairpin formed in the 3' untranslated region (UTR), where a hairpin-binding protein (HBP) binds, so that CPSF73 can cleave the pre-mRNA (Dominski et al., 2005). Recognition of the cleavage site and positioning of the nuclease is thought to be made by U7 snRNP, which also helps in degradation of the 3' cleaved portion (Cotten et al., 1988). These specificities, and the relevance of histones in the cellular context, may translate into particular Pol II dynamics and profiles in the synthesis of histone mRNA, not observed in other protein-coding genes.

The complexity of 3' processing and termination mechanisms, allied to the co-transcriptionality and diversity of functions of its components - as shown for CPSF73, but also Xrn2, that has a role in premature termination (Brannan et al., 2012) – hints at a link with transcription regulatory mechanisms. It can be postulated that Ser2 phosphorylation of the CTD, a mark often found at the end of genes, may be related to these processes. An in-depth study tracking the polymerase in CPA factors-depleted cells can certainly shine a light on mRNA 3' end determination and effects of the downstream processing events in transcription.

3. Micro RNAs

3.1 Overview

Micro RNAs (miRNA) are RNA strands of about 22 nucleotides that are involved in gene expression regulation by transcriptional silencing. They were first discovered in *C. elegans* in 1993, when it was described that the gene lin-4 produced a short non-coding RNA with an almost complementary sequence to the 3 'end of the mRNA of lin-14 (Lee et al., 1993). Since then, miRNA have been discovered in all superior eukaryotic organisms (Maxwell et al., 2012). miRNA derive from pre-miRNA, which is an RNA hairpin.

Regulation by miRNA is preformed through the RNA interference (RNAi) pathway. Their function is accomplished together with other proteins in the RNA-induced silencing complex (RISC). In the RNAi pathway, pre-miRNA is exported to the cytoplasm via the Exportin 5-

RanGTP complex (Lund et al., 2004). Upon arriving, the hairpin is cleaved by Dicer, a type III RNase, generating a double-strand RNA (dsRNA) of about 22nt (Bernstein et al., 2001). Dicer is one of the elements of RISC, the others being HIV-1 transactivation responsive element (TAR) RNA-binding protein (TRBP), PACT and proteins of the Argonaut family (Rana, 2007). The complex then selects only one of the strands of the dsRNA to be used. The selection is not yet fully comprehended, but some evidence points to a selection based on the strand thermodynamic stability, discarding the most stable strand (Siomi and Siomi, 2009). The complete ribonucleoprotein complex is called miRISC. Finally, the miRNA incorporated in RISC will find its target and bind to it. Binding can be partial (only some nucleotides pair with the target) or complete. The second is more common in plants, although it can also happen in animals. These two mechanisms are functionally different, since incomplete pairing only leads to translational silencing (which can be transient), whereas complete pairing leads to target mRNA degradation (by the C-terminal PIWI domain of Argonaut proteins), although it might not always be the case (Bartel et al., 2004). Many studies point to the pairing of some positions having more effect in the mRNA's fate, rather than the whole miRNA. Additionally, in cases of degradation, it has been shown that miRNA can proceed to a different target to fulfill the same function (Hutvágner and Zamore, 2002). As for silenced mRNAs, in some instances they are clustered in sub-cellular regions called Processing bodies (P-bodies), where other interveners may eventually, but not certainly, mediate RNA turnover, generally by decapping mRNA (Brengues et al., 2005).

3.2 Micro RNA transcription and processing

As more miRNAs were discovered, it became possible to classify them into distinct categories according to their gene structure. Some miRNA are intragenic (Figure 5A), whereas others are intergenic (Figures 5B and C). Intragenic micro RNAs are found in introns, and usually have the same orientation than the host gene. They can be found in introns of protein-coding and long non-coding RNA genes (He et al., 2008). Intergenic miRNA can be near or far from other genes. They can be classified into clustered miRNA (Figure 5B) or single miRNA (Figure 5C). Micro RNAs belonging to the same cluster can have similar functions or related targets, with some clusters being associated with tumors (Mendell, 2008). It is thought that all intergenic miRNA have a larger associated transcript called primary-miRNA (pri-miRNA), and some were already described, as seen in Figure 5B (Lee et al., 2002). Like their intragenic counterparts, most intergenic miRNA are transcribed

by Pol II (Lee et al., 2004), although some miRNA clusters, because of their close association with Alu elements, are transcribed by RNA polymerase III (Borchert et al., 2006).

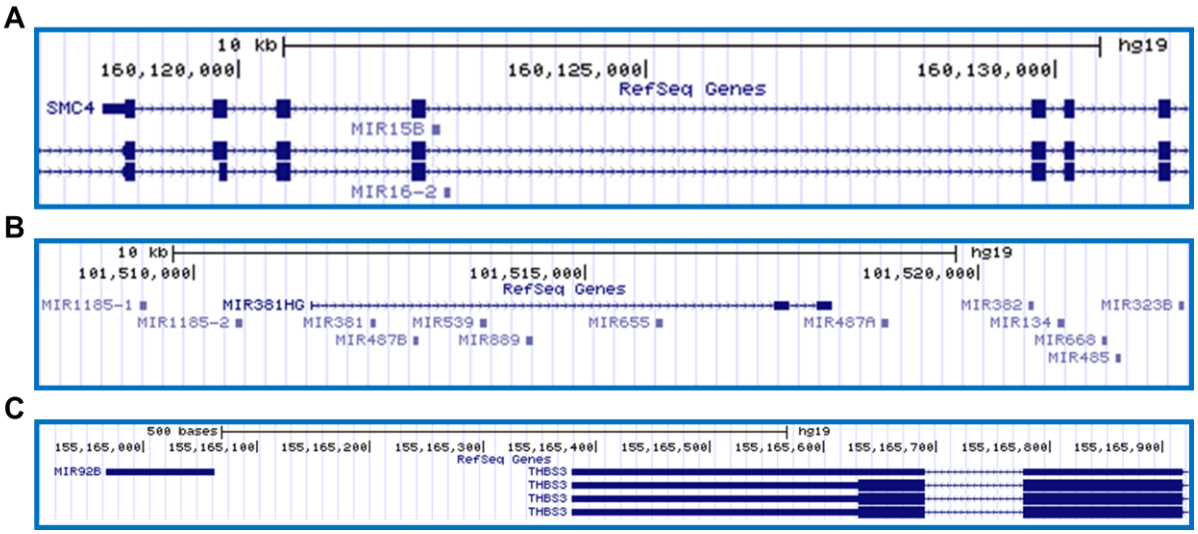


Figure 5: Three distinct types of miRNA genes. A – intronic miRNA; B – Intergenic clustered miRNA; C – intergenic single miRNA

In order to obtain the pre-miRNA, introns (after being debranched) and pri-miRNA have to be cleaved so that the hairpin sequence can be obtained. The nuclear RNase III Drosha is responsible for the primary transcript cleavage (Lee et al., 2003), acting together with DGCR8 (Yeom et al., 2006). However, the mechanism Drosha uses for recognizing the cleavage site is still very debated (Zeng et al., 2004; Ma et al., 2013), and so prediction of these sites has a large value in understanding the mechanism and recognizing novel miRNA (Hu et al., 2013). In addition to this, some micro RNAs differ from their reference in very few nucleotides. These are called isomiRs (Morin et al., 2008), and although their biogenesis is still poorly understood, there is evidence pointing to some of them owing their variability to multiple cleavage by Drosha (Ma et al., 2013).

Just like every type of gene, micro RNAs must also have proper nomenclature for organization purposes. In earlier times, naming was more similar to genes, and the standardization to include “mir” on their name (still following the formatting for a species gene’s name) was only included later (Ambros et al., 2003). Also, miRNAs from the same hairpin used to be distinguished by their expression level, but nowadays that is done based on strand. Naming also depends on homology with miRNA found in other organisms, on the genomic locus they’re included, on base differences between two sequences and on the order of discovery (Griffiths-Jones, 2004). Most of the current information about micro RNAs is on miRBase (current version is v21, most recent description in Kozomara and Griffiths-Jones, 2014). This database includes deep sequencing datasets, genomic coordinates, sequence and

biological information, among other relevant details. In this database, miRNA are named like has-mir-17-5p, where the first three letters describe the species, and 5p tells the strand. When new miRNAs are discovered they can be added to the database and the attributed name can be then included in the publication.

4. Genome-wide study of transcription

4.1 High-throughput sequencing approaches

Understanding transcription is historically associated with molecular, genetic or biochemical studies of single genes or components. The first change in this paradigm came with the introduction of microarrays (Schena et al., 1995), allowing for the quantification of the RNA from several genes at the same time. With development of the technology, it became possible to identify alternative splice sites (Johnson et al., 2003) and to correlate co-expression of genes with promoter motifs (Veerla and Höglund, 2006). In spite of the huge breakthroughs they allowed in gene expression studies, microarrays have some inherent bias when it comes to coverage and amplification (Boelens et al., 2007). The development of next generation sequencing (NGS) technologies allowed for the development of less biased, albeit more expensive, techniques, with greater coverage. RNA-seq was capable of providing the same type of information as microarrays, but in an even larger scale, driving the discovery of novel transcripts and isoforms. Analyzing cell-fraction RNA-seq data also revealed a generalized presence of co-transcriptional splicing in protein-coding genes, but not so much in lncRNAs (Tilgner et al., 2012). But because it is a dynamical process, transcription ought to be studied using methodologies that focus not only on the end product, the mRNA, but also on the intermediates in its synthesis – nascent RNA and Pol II.

Chromatin immunoprecipitation (ChIP) protocols have been developed to assess interactions between proteins and DNA (Kim and Ren, 2006). The binding sites in DNA were initially analyzed using microarrays, but quickly became evident that the coverage offered by next generation high-throughput sequencing would provide more accurate and robust results, which led to the development of ChIP-seq protocols (Johnson et al., 2007; Robertson et al., 2007). Later, ChIP-seq was used to assess Pol II distribution across genes (Baugh et al., 2007), which attested the large accumulation of stalled Pol II in a promoter proximal region previously described in microarray studies (Kim et al., 2005; Guenther et al., 2007). ChIP-seq

assays were also performed using antibodies that specifically target different phosphorylation patterns, thus showing a broad genome-wide picture of where in genes each phosphorylation mark could be found (Rahl et al., 2010, Grosso et al., 2012).

While ChIP-seq profiles rendered a good image of gene occupancy by Pol II, resolution was still low, there was no distinction of whether the enzyme was actively transcribing or paused, and the focus in the relevant part of transcription – RNA synthesis – was not being captured. This resulted in the development of four techniques that aimed to solve these problems, all of them targeting Pol II-associated RNA (Figure 6).

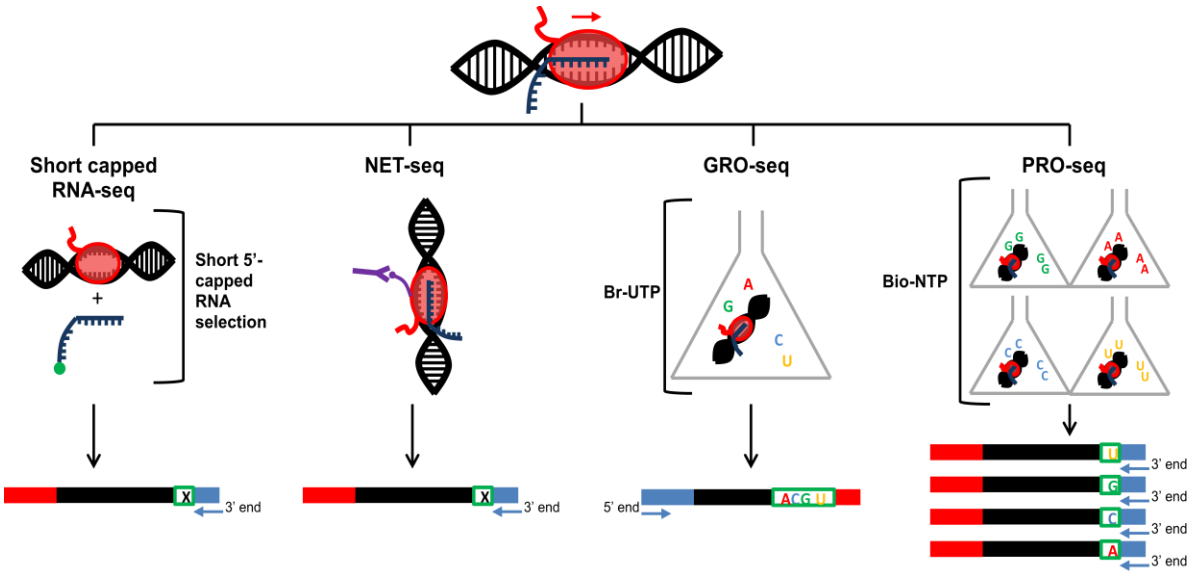


Figure 6: RNA-based genome-wide transcription tracking protocols. Short-capped RNA sequencing selects the target RNAs by the presence of a 7-methylguanosine cap; NET-seq selects the RNA attached to the polymerase by immunoprecipitaton targeting a synthetic flag; and GRO-seq and PRO-seq are based on *in vitro* run on reactions using labeled nucleotides. Description of each technique in the main text.

Short capped RNA-seq (Nechaev et al., 2010) was introduced to study promoter-proximal pausing. It captures RNAs with a 7-methylguanosine cap, selecting those with between 25 and 120 bp. Those short RNAs are then sequenced, and it is possible to determine the first base transcribed - by using a 5' sequencing primer – or the last base – using a 3' sequencing primer. The last base sequenced is the last base incorporated by the polymerase, so the technique allows determination of paused Pol II position at a single-nucleotide resolution by aligning the whole reads and then extracting that base. However, this protocol can only provide knowledge on the position of paused polymerases in early elongation, because of the size and cap selection steps, and it cannot also accuratly distinguish between Pol II associated nascent transcripts and released short transcripts.

The sequencing of native elongating transcripts (NET-seq) associated with Pol II (Churchmand and Weissman, 2011) was able to expand the single-nucleotide resolution of polymerase tracking to the whole gene. This method is based on the immunoprecipitation of a

flag-tagged Pol II. Due to the stability of the trenary complex, it is possible to extract the RNA associated with the polymerase. Aligning the reads and extracting the position of the last base will tell which was the last nucleotide incorporated. The method was also designed as strand-specific, giving information about sense and antisense nascent transcripts. This is an important distinction as it reduces noise in the data and allows the discovery of new transcriptional units. Finally, NET-seq also captures splicing intermedates from co-transcriptional splicing. While this has to be considered while analyzing the data, it also allows the study of splicing, and perhaps other co-transcriptional events that generate intermediates, if they're captured by the protocol.

Both of the methods described above unbiasedly capture Pol II-associated transcripts. But they make no distinction of which polymerases are paused or actively engaged in transcription. Acquiring the position of engaged Pol II can be achieved through nuclear run-on reactions coupled to sequencing, as it is done in global run-on sequencing (GRO-seq) and precise run-on sequencing (PRO-seq) (Core et al., 2008; Kwak et al., 2013). The first method uses Br-UTP to mark the run-on synthesized transcripts so they can be isolated and then sequenced. Yet this method lacks precision when compared to the others described, so it was upgraded to PRO-seq. In this protocol, four libraries are prepared from four run-on reactions, each of them using only one nucleotide as substrate. The nucleotides used are biotinylated so that the transcripts can be selected. The libraries are then sequenced and merged, and the last base of each read is considered the position where the engaged Pol II is. This method does not guarantee single nucleotide resolution as it is possible to find the same nucleotide repeated, which would drive the incorporation of two nucleotides in the nuclear run-on reaction. Also, the added manipulation required for isolation of nuclei and run-on may also disrupt some features of transcription.

Even though these genome-wide protocols show data for virtually all the genes in the genome, there are still biases that need to be considered when presenting the results for these experiments. First of all, the protocols are designed to be preformed in a collection of cells. While this captures the biological variability between individuals, it does not express exactly how a single individual behaves. Specifically, when looking at a single-nucleotide resolution gene profile, we may identify two consecutive bases with signal, yet it is physically impossible to have two polymerases in consecutive bases. Development of single-cell protocols will bring this individuality to the analysis, but conversely sacrificing the patterns only observable when looking to many subjects. Second, it is customary to look at metagene profiles (i.e. a global average profile for all genes) when interpreting this kind of data. While

it may be informative of general features, it may be necessary to subset the genes by some feature, and ultimately look at many individual genes, so as to identify differences between them that may be lost while averaging the full set. Finally, a constant adaptation of computational biology's protocols and tools is needed, as more advanced and singular arise. While some steps in the pipelines are well established, especially alignments, other steps are specifically adapted to the protocol in question and may therefore not be fully optimized.

4.2 Data Analysis

The exponential increase of available biological datasets from high-throughput techniques – and, in particular, NGS protocols – has stimulated the development of bioinformatical tools that can accurately, but also efficiently, process the large volumes of data produced. These datasets are not only abundant, but also diverse, as all biological research fields have realized the importance of more comprehensive data, rather than focusing on individual components of the system studied. This led to the development of many protocols, like the ones described in the previous section, to which the data analysis must adapt.

Although the output of sequencing platforms can differ between them, the most common type of files – and the one used in this project - is FASTQ (Cock et al., 2010). In this kind of files, each read is named uniquely and accompanied by the quality of each base. It is important to know what is the encoding for the quality scores, since some tools do not automatically recognize it. Read quality should also be assessed before starting any analysis, to account for possible biases. A commonly used tool that collects relevant quality statistics is fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, last accessed on August 23rd 2014). This allows the identification of overrepresented sequences, such as adapters or RNA sequences that either hinder the analysis or introduce bias, leading to wrong conclusions.

There is a plethora of adaptor trimming tools to choose from. The choice varies with read type and size, as well as features offered by each software. Cutadapt (Martin, 2011) is one of these tools, and it allows the user to trim reads from both ends and considering as many adapter sequences as needed. Besides, it does not require the complete identification of the adaptor to perform trimming, allowing the user to set a minimum of nucleotides corresponding to the adapter to be trimmed with a certain error rate. When sequencing short nucleotide strands using a pair-end protocol, it is possible that the read size is greater than the actual length of what is being sequenced. In this case, a variable number of bases

corresponding to the opposite strand's adaptor might be sequenced, resulting in a read that cannot be aligned with the reference genome. The referred features offered by Cutadapt present a way to solving this issue.

Aligning the obtained reads to a reference genome is arguably the most important step in a sequencing analysis. However, it is usually also the most computationally demanding stage, even more considering the need of high coverage data. This happens because of the elevated number of comparisons that would have to be made between the bases of millions of reads and the billions of bases in a genome, and even more considering the quality of each base. Luckily, it is possible to implement some heuristics in order to speed up the algorithms. Many aligners have been developed over the years, with the purpose of finding the best balance between an accurate alignment and a fast execution. TopHat2 (Kim et al., 2013) is the latest version of a widely-used tool for aligning RNA-derived reads to reference genomes. TopHat2 resorts to Bowtie2, a tool from the same lab (Langmead and Salzberg, 2012), to align the reads, and then performs splice junction finding. Many parameters are customizable so that the user can fit them to the data. It is usually important to get uniquely aligned reads, as multiple aligned RNA reads cannot be interpreted as a product of a single DNA sequence transcription.

TopHat2 outputs a BAM file with information about where the reads aligned to the reference. It is a useful way of storing the alignment, but cannot be directly worked with. To work with BAM files (and their non-binary counterpart, SAM files), SAMtools is available (Li et al., 2009). This collection of tools enables visualization, filtering, sorting and indexing of these files, among other features. They can also be used together with other tools to find spliced reads, or to isolate the last nucleotide as single-nucleotide resolution techniques require. However, BAM and SAM formats do not supply appropriated data that can be worked on, such as read counts in genome intervals, and also don't provide effective means to make operations in the data, like intersections or subtractions. For these cases, BEDTools (Quinlan and Hall, 2010) come in handy. This toolkit, that relies mostly on the BED format (Kent et al., 2002), provides the user with a framework for intersecting datasets, separating data by location, calculating local or genome-wide coverage, and can be articulated with other tools if needed. The output BED format files are fairly more readable than the SAM format ones, and information can be easily extracted to be processed. BEDTools is also compliant with other file formats, such as GTF of BAM files, which makes the analysis much more agile since there is no need of converting these files to another format. For instance, a BAM file output by an aligner may be directly used to assess coverage of desired features.

Visualization is of great importance when dealing with genomic data. The UCSC Genome Browser (Kent et al., 2002) is a useful online resource for visualizing data in a genomic context. It provides several tracks of annotations or data from other sources, so as to facilitate interpretation. It is possible to see, gene by gene, the presence of RNA-seq reads or polymerase distribution. But to get a sense of what is happening in the average gene, data can be compiled in metagene profiles. These profiles divide genes in windows, and plot the mean read counts - or a normalized version like reads per kilobase per million reads (RPKM) – of every window of all genes. This results in an average profile that makes it possible to identify features present in many genes, for instance the accumulation of polymerase in the promoter region (Core et al., 2008; Nechaev et al., 2010). However, construction and interpretation of these must be done carefully. First, not all genes can be included in a profile. Many genes overlap each other, and that can give rise to features that may not exist. Other sets of genes may also have their unique features - for example, replication-dependent histones, which are shorter than average genes and have unique Pol II occupation profiles -, and so should be removed from the analysis. Second, genes can also have unique features that are not displayed or that interfere with the profile, and therefore some manual selection of the genes included has to be preformed, and individual examples should always be shown. Other metrics can be presented to highlight differences between sets of genes or conditions. A logarithm of the quotient of the number of reads in two windows is a useful means for making such comparisons, and a distribution of the values for the genes can also be shown as a boxplot. Further statistical testing can be applied to confirm those differences.

From an RNA-seq experiment we can also identify which transcript isoforms are more or less expressed. The Cufflinks software (Trapnell et al., 2010) is widely used as a tool to attribute read counts to genes and isoforms, allowing inferences to be made about their expression. But it is also possible to identify which exons are alternatively being expressed. MISO (Mixture-of-Isoforms) (Katz et al., 2010) uses RNA-seq data to quantitatively predict which alternative splicing phenomena occur in a sample or between samples. This program uses its own database of alternative splicing events, and follows a Bayesian framework to attribute a read either to one isoform or another. The database are divided by their type of event (skipped exons, alternative 3'/5' splice sites, mutually exclusive exons, tandem 3' UTRs, retained introns, and alternative first or last exons), and processing results in a value that indicates whether one event or the other is selected. For instance, in a larger scale, this allows seeing differences between included or skipped exons in polymerase occupation, but these and other features can also be shown individually.

NGS technologies opened the door to large genomic studies, and to high coverage sequencing data. Yet, constant adaptation of methodologies is required for capturing not only the broad patterns but also the fine details provided by these approaches.

5. Objectives

Considering the complexities of RNA transcription and processing, this work aims to elucidate more about the interactions between these two processes in a genome-wide scope. The focus of this study is the Pol II CTD, since it is one of the key regulators of interactions involving Pol II during transcription. To achieve this, the NET-seq protocol (Churchman and Weissman, 2012) was modified to enable usage of CTD isoform-specific antibodies. An additional step was also included where the chromatin fraction from which Pol II was precipitated was treated with micrococcal nuclease (MNase), in order to degrade exposed RNA sequences, thus increasing the specificity for RNA protected by Pol II. This new protocol was termed "advanced NET-seq" (ANET-seq), and allows for the first time for a single-nucleotide resolution mapping of CTD isoforms in the genome. To further support ANET-seq results and interpretation, chromatin-fraction RNA (ChrRNA) was sequenced. This would show some unstable RNAs, such as promoter upstream transcripts (PROMPTs) introns, and transcription downstream of the 3' end. All samples were sequenced in Illumina HiSeq 2000 or 2500 sequencers.

Two sets of data were produced for this project. The first includes one ChrRNA sample and ANET-seq samples generated using antibodies for unphosphorylated Pol II CTD, Ser2-phosphorylated CTD, Ser5-phosphorylated CTD and all CTD isoforms. These aimed at showing differences between CTD isoforms in transcription. The second includes ChrRNA and ANET data from three 3' end processing and termination factors knock-downs and a control. These are meant to show the effects that each factor has on Pol II transcription dynamics and changes in newly synthesized transcripts. The ANET-seq from this set used an antibody targeting Ser2-phosphorylated CTD, the predominant isoform at the end of genes.

The main objectives of this project were:

1. Define an analysis pipeline for ANET-seq data.
2. Comprehend the roles of different CTD isoforms in during transcription.

3. Examine the CTD phosphorylation dynamics associated with splicing and miRNA biogenesis.

4. Elucidate the different roles of the termination factors Xrn2, CPSF73 and CstF64+CstF64τ during transcription.

Being a frontier discipline, computational biology requires the input and collaboration specialists from different fields. It is from the combination of these different skill sets that complex and relevant new discoveries can be made. In this work, I preformed all of the sequencing data analysis, such as trimming and aligning reads, making average gene and exon profiles, plotting single gene profiles or calculating Escaping Indices. The experimental work was performed by other scientists specialized in those protocols and they are duly credited in the final manuscript.

The funding for the work here presented was granted by Wellcome Trust Programme, ERC Advanced Grants and Fundação Ciência e Tecnologia.

Chapter 2

Human genome-wide profiles of nascent transcription and co-transcriptional processing: advanced NET-seq technology

Takayuki Nojima[1§], Tomás Gomes[2§], Ana Rita Fialho Grosso[2], Hiroshi Kimura[3], Michael J. Dye[1], Somdutta Dhir[1], Maria Carmo-Fonseca[2]* and Nicholas J. Proudfoot[1]*

[1]Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, United Kingdom.

[2]Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, 1649-028 Lisboa, Portugal

[3] Department of Biological Sciences, Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, Yokohama, Japan

[§]These authors have contributed equally to the work.

* Joint communicating authors

1. Summary

RNA polymerase II (Pol II) transcribes nascent RNA throughout the mammalian genome. However, many aspects of nascent RNA metabolism are poorly understood due to RNA instability and technical limitations. We have employed high throughput sequencing at single-nucleotide resolution to characterize nascent transcription in HeLa cells; advanced native elongating transcript-sequencing (ANET-seq). This provides precise maps of nascent RNA within the Pol II elongation complex that correlate with the C-terminal domain (CTD) phosphorylation state of the Pol II large subunit. We detect substantial Pol II bidirectional pausing at transcription start sites (TSS). We also demonstrate exon tethering to the CTD Ser^5 phosphorylated Pol II complex and co-transcriptional pre-miRNA biogenesis. Depletion of cleavage and polyadenylation (CPA) factors causes termination defects, reducing Pol II pausing at transcription end site (TES). Additionally the 3' end termination machinery plays a promoter role by restricting non-productive RNA synthesis at the TSS in both sense and antisense directions.
(150 words)

2. Highlights

- ANET-seq monitors nascent RNA within the mammalian Pol II complex.
- Pol II pausing at TSS and TES with different Pol II CTD phosphorylation states.
- Exon tethering during co-transcriptional splicing links CTD S5P to 5’SS cleavage.
- Diverse kinetics of co-transcriptional pre-miRNA biogenesis.
- CPA factors are associated with Pol II pausing at TES.
- CPA factors and Xrn2 regulate sense and antisense premature termination at TSS.

3. Introduction

The global analysis of nascent RNA has been achieved by genome-wide nuclear run on-sequencing (GRO-seq) and precision nuclear run on-sequencing (PRO-seq) using modified nucleotides (Core et al., 2008; Kwak et al., 2013). These approaches have provided high resolution maps of Pol II nascent transcription in mammals and flies. In both cases, Pol II accumulation was detected at promoters where it acts as a major regulatory block in the transition into productive transcriptional elongation (Core et al., 2008; Hah et al., 2011; Min et al., 2011; Rahl et al., 2010; Saunders et al., 2006). The precise maps of PRO-seq reads identified two different types of Pol II pausing at the transcription start site (TSS), referred to as proximal and distal TSS pausing. PRO-seq additionally showed Pol II accumulation near 3' splice sites (SS) which is likely to be important for the selection of active exons (Kwak et al., 2013). The GRO-seq approach has also provided a correlation between Pol II density and nucleosome occupancy as observed at the 3' end of many genes (TES, transcription end site), suggesting a connection with transcription termination (Grosso et al., 2012).

Precise maps of Pol II nascent RNA have also been generated by the native elongating transcript-sequencing (NET-seq) method in yeast (Churchman and Weissman, 2011). Here endogenous Pol II was flag tagged by genomic integration which allows the Pol II nascent RNA complexes to be immuno-precipitated with flag antibody. This method revealed that Pol II back-tracks during elongation, based on single nucleotide resolution of nascent RNA profiles. However, the relationship between Pol II CTD modifications and nascent RNA could not be determined. We now report the establishment of a modified mammalian NET-seq technique using a selection of CTD modification specific Pol II antibodies. We use this technology to monitor genome-wide nascent RNA profiles in HeLa cells and call this technology advanced NET-seq (ANET-seq). Importantly, we correlate different Pol II CTD modifications with specific patterns of nascent transcription and coupled RNA processing. Our extensive ANET-seq datasets (obtained using different CTD modification specific Pol II antibodies) provide a “treasure trove” of detailed information on co-transcriptional RNA processing in mammalian cells. In this study we have focused on protein coding gene transcripts. Future analysis will turn to intergenic non coding (nc) RNA transcription.

It is widely known that Pol II CTD is differentially phosphorylated during the transcription cycle. CTD comprises a 52 repeated heptapeptide (Tyr^{1}-Ser^{2}-Pro^{3}-Thr^{4}-Ser^{5}-Pro^{6}-$Ser^{7,}$ YSPTSPS) which is highly phosphorylated during productive transcription. Based

on chromatin immuno-precipitation (ChIP), Ser^5 phosphorylation (S5P) accumulates at active promoters while Ser^2 phosphorylation (S2P) is involved in co-transcriptional processing events in the gene body, such as splicing and 3' cleavage and polyadenylation (CPA) (Brookes and Pombo, 2009; Egloff et al., 2012a; Heidemann et al., 2013). We used unphosphorylated (unph), S2P, S5P and total (unph+ph) CTD antibodies to analyze CTD phosphorylation-specific nascent RNA profiles across the human genome. ANET-seq analysis reveals that unph CTD Pol II-nascent RNAs are accumulated over the TSS while S2P Pol II nascent RNA are spread throughout the gene body and TES, demonstrating that this method provides differential maps of CTD phosphorylation-specific nascent RNA. Interestingly, high CTD S5P Pol II associated signals are detected at 3' ends of functional exons. We have also characterized co-transcriptional microprocessing of pre-miRNA in the introns of protein coding genes and describe new features of this mechanism. Although Pol II pausing at TES is well established (Davidson et al., 2014; Proudfoot, 2011; Skourti-Stathaki et al., 2011), CPA factors are also recruited co-transcriptionally onto chromatin (Glover-Cutter et al., 2008), but their effect on Pol II pausing has not been widely characterized. Our ANET-seq data show that depletion of CPA factors such as CPSF73 and CstF-64+CstF-64 tau proteins cause a substantial reduction of Pol II pausing downstream of TES. In contrast 5'-3' exonuclease Xrn2 knockdown did not affect TES pausing. Surprisingly depletion of all of these 3' end termination factors increases promoter-associated CTD S2P Pol II pausing on both mRNA and promoter upstream transcript (PROMPT) strands.

It is abundantly clear that ANET-seq can be used to generate precise maps of Pol II phosphorylation-dependent nascent RNA profiles across the human genome. We predict that ANET-seq will be a powerful tool to demystify the complexities of Pol II pausing and co-transcriptional RNA processing.

4. Results

4.1 ANET-seq strategy

As a starting point to enrich for unstable nascent RNA across the human genome, we isolated a nuclear chromatin fraction which is enriched in the transcriptionally active Pol II isoform (Pol IIo) and associated nascent RNA (Figure S1; (Nojima et al., 2013; West et al., 2008). This chromatin-bound RNA was directly sequenced (ChrRNA-seq) as follows. RNA was fragmented to 150~200 nt and ligated to adaptors for strand-specific paired end deep sequencing (Figure 1A top and Experimental Procedures). ChrRNA-seq detects unstable RNA such as promoter upstream transcripts (PROMPTs), introns and read through transcripts (Figure 1D). For ANET-seq, the chromatin fraction was independently subjected to Pol II immuno-precipitation (IP) so that nascent RNA could be correlated with Pol II genic distribution. In detail chromatin was first digested with micrococcal nuclease (MNase) prior to Pol II IP. Note that accessible RNA will also be digested by MNase treatment (Figure 1A bottom and Figure S2). To confirm that Pol II is effectively released from the insoluble chromatin fraction, western blot analysis was carried out on the supernatant fraction using Pol II 8WG16 antibody. Both phosphorylated (Pol IIo) and unphosphorylated (Pol IIa) forms were detected in a MNase dose-dependent manner (Figure 1B). IP was then carried out on the supernatant derived from MNase-digested chromatin using Pol II 8WG16 antibody (Figure 1C). To check nascent RNA distribution after the cell fractionation and MNase digestion, we initially used nuclei that had been subjected to nuclear run on (NRO) labeling with [α-^{32}P] UTP. In the nucleoplasmic (Np) fraction, radiolabeled long RNA (over 600 nt) was detected. After MNase incubation, a smear of RNA (10-600 nt) was detected in the chromatin pellet (P), but a shorter RNA smear (10-200 nt) in the chromatin supernatant (S). As expected, these shorter RNAs were efficiently precipitated by Pol II 8WG16 antibody. Although the predominant size of the immuno-precipitated RNA was 20-45 nt, we selected a longer RNA fraction (35-100 nt) to obtain uniquely mapable reads on the human genome following deep sequencing. In this method, the Pol II complex will protect nascent RNA from MNase digestion. The hydroxylated 3' end (3’OH) of the nascent RNA corresponds to the terminal nucleotide synthesized by Pol II (shown by an asterisk in Figure 1A). The 5' end of the cleaved Pol II-associated RNA will also be hydroxylated after MNase digestion. To achieve strand-specific RNA sequencing we carried out a kinase reaction on the IP beads to phosphorylate all nascent RNA 5’ ends but leaving the 3’OH intact (Figure S2). We then ligated Illumina adapters to gel purified RNAs and performed Illumina High throughput

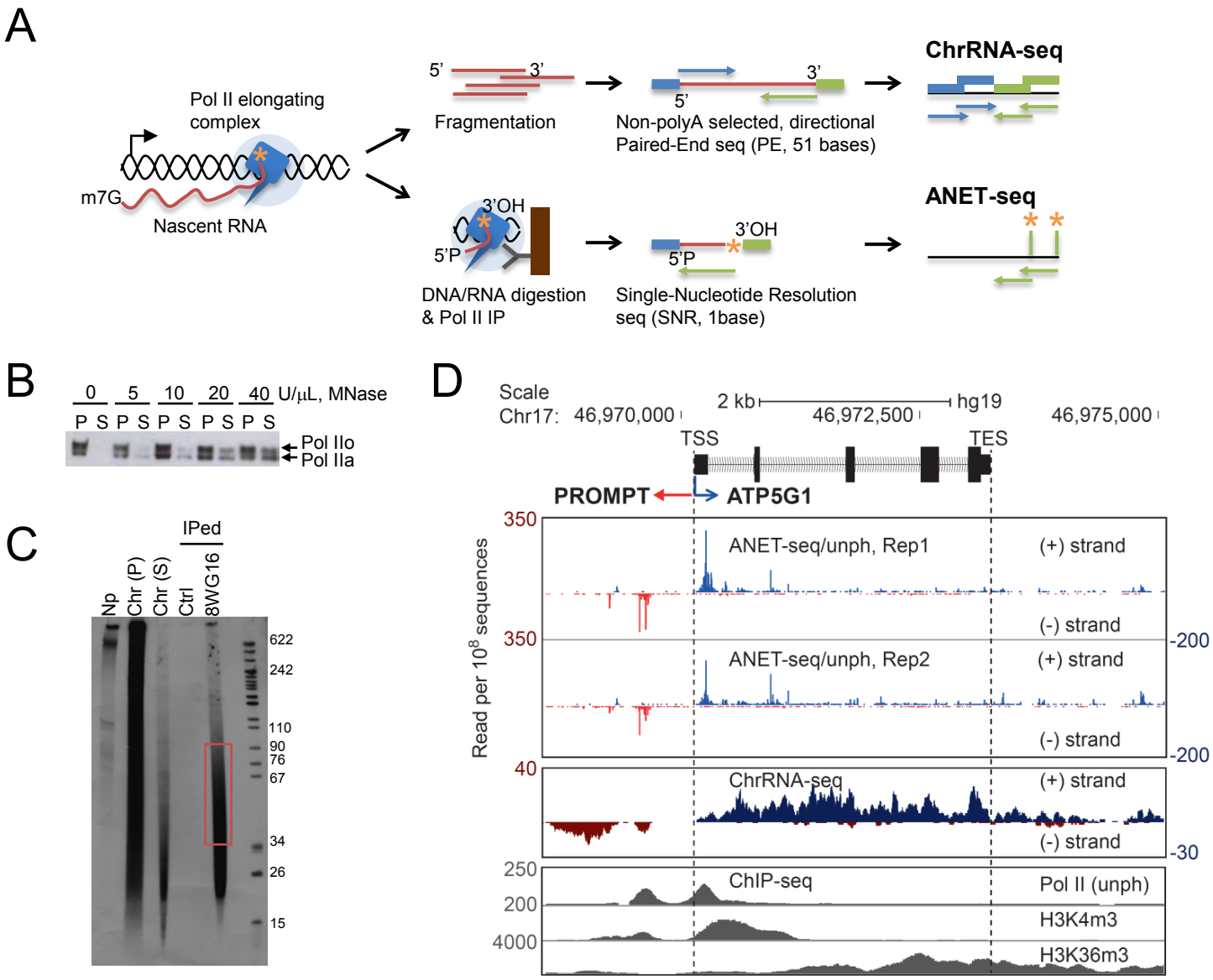


Figure 1. ANET-seq methodology

(A) ChrRNA-seq and ANET-seq strategies. Pol II (dark blue) elongation complex (light blue circle) and associated nascent RNA (red line) was purified from chromatin for ChrRNA-seq (top). Orange asterisk shows catalytic site in Pol II. Fragmented nascent RNA was subjected to directional paired-end deep sequencing. For ANET-seq (bottom), DNA and RNA were digested with MNase and the Pol II-nascent RNA complex was precipitated with different Pol II antibodies. Isolated RNA was 3' end deep sequenced and the 3' end nucleotide uniquely mapped on the human genome (green bars).

(B) Pol II release from insoluble chromatin DNA. Chromatin DNA was digested with indicated concentration of MNase. Western blot was carried out using 8WG16 Pol II antibody. P; pellet, S; supernatant.

(C) Nascent RNA distribution in ANET-seq method. Nascent RNAs were 32P-labeled by NRO reaction. Fractionated nascent RNA were from Nucleoplasm (Np), Chromatin pellet (Chr (P)) and supernatant (Chr (S)). IP was with 8WG16 Pol II antibody. 35-100 nt RNA purified from gel (red box).

(D) ATP5G1 gene ANET-seq. Two biological replicates of ANET-seq/unph using 8WG16 Pol II antibody. ChrRNA-seq shown as ANET-seq input. ChIP-seq (Pol II (unph), H3K4m3 and H3K36m3) data are from ENCODE project datasets (Consortium et al., 2012).

paired-end sequencing which generated ~10^8 reads for each ANET-seq sample. For library construction we omitted the NRO step since the NRO reaction disturbs the native Pol II distribution (data not shown). The above Pol II IP from MNase treated chromatin, isolation and sequencing of the associated RNA constitutes a refined mammalian NET-seq protocol that we term ANET-seq.

Finally, libraries were prepared from two biological replicates of HeLa native chromatin after Pol II 8WG16 IP. Deep sequencing was conducted using a reverse sequence primer to read the 3' ends of the RNA insert which corresponds to the RNA synthesis site in the Pol II active site (Figure 1A). ANET-seq data aligned to the human genome (hg19) was compared to 8WG16 Chromatin IP (ChIP)-seq and ChrRNA-seq in either transcriptionally active or inactive genes (Figure 1D). Modifications of Histone H3, H3K4m3 and H3K36m3, reflect active promoters and gene bodies, respectively. Strand-specific transcription activity was revealed by ChrRNA-seq. As expected, both replicates of ANET-seq/8WG16 (unph) display strong peaks at the active TSS consistent with the ChIP-seq/8WG16 (unph) profile. Additionally, ANET-seq data revealed both sense and antisense transcription on active genes, as previously shown by GRO-seq and PRO-seq (Core et al., 2008; Kwak et al., 2013). Note that ChIP-seq is not able to distinguish the strand-specific Pol II distribution.

4.2 Pol II CTD phosphorylation-specific nascent RNA profiles at TSS and TES

A major benefit of ANET-seq is that it allows the use of different Pol II antibodies to precipitate modified Pol II-associated nascent transcripts. We therefore used specific monoclonal antibodies to detect CTD phosphorylation-dependent nascent RNA profiles. The newly described CMA302, CMA303 and CMA301 mouse monoclonal antibodies are specific for CTD S2P, CTD S5P and all CTD isoforms respectively (Stasevich et al., 2014). 8WG16 is known to be relatively selective for unphosphorylated CTD. By way of confirmation we show IP Pol II western blots using these antibodies under ANET-seq conditions (Figure 2A). Although CMA302 antibody is able to precipitate some Pol IIa, both of CMA302 and CMA303 antibodies mainly recognize Pol IIo (Figure 2B). As expected 8WG16 antibody precipitated mainly Pol IIa. We also performed Pol II ChIP analysis on three specific genes using these monoclonal antibodies and compared them to the commercial polyclonal antibodies (ab5095 (S2P) and ab5131 (S5P), respectively) which are widely used for ChIP-seq assay (Perez-Lluch et al., 2011) (Figure S3). Notably very similar ChIP profiles were observed for the different S2P and S5P specific antibodies.

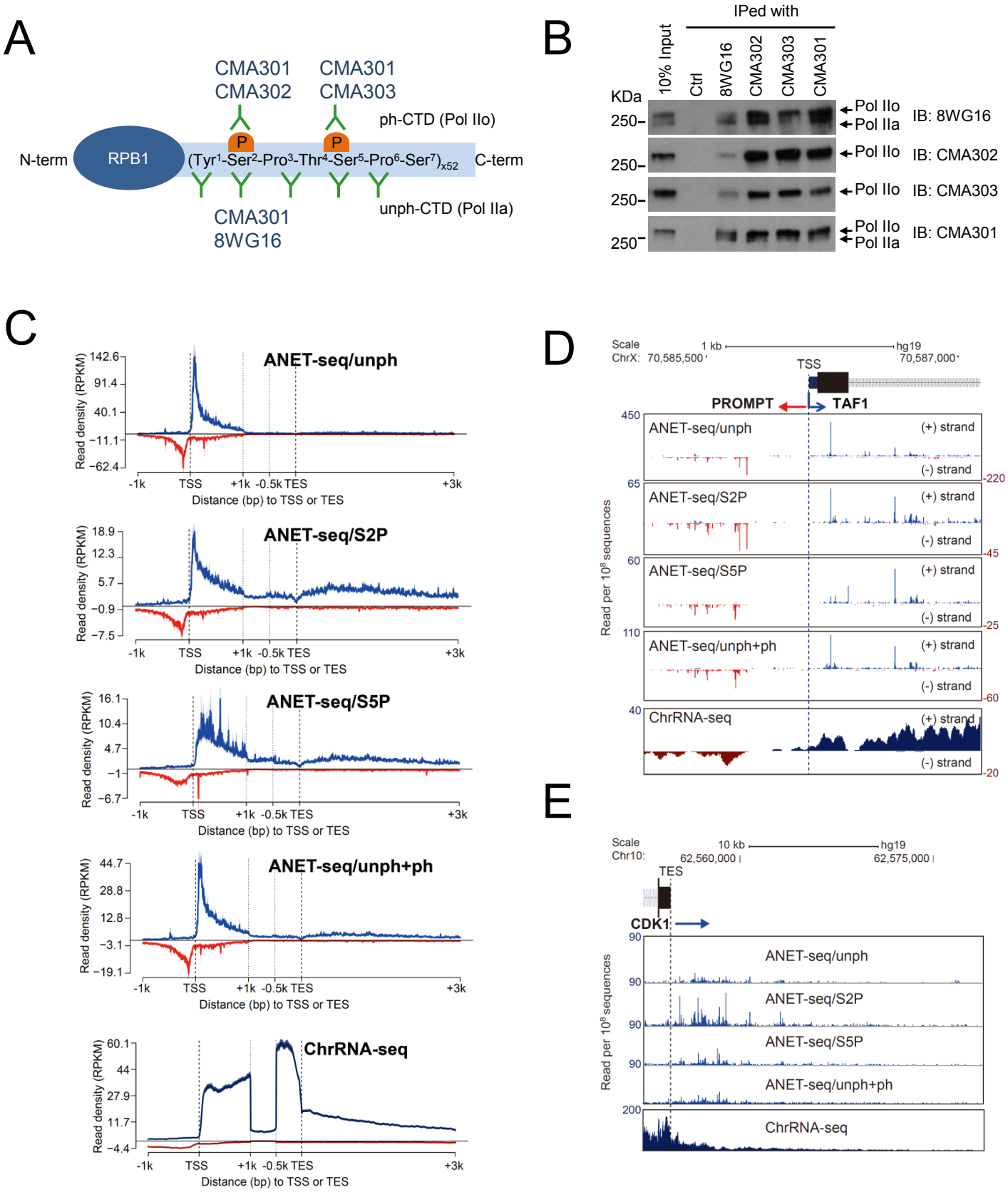

A
CMA301
CMA302
CMA301
CMA303
ph-CTD (Pol IIo)
N-term
RPB1
(Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7)x52
C-term
unph-CTD (Pol IIa)
CMA301
8WG16
B
IPed with
10% Input
Ctrl
8WG16
CMA302
CMA303
CMA301
KDa
250
Pol IIo
Pol IIa
IB: 8WG16
IB: CMA302
IB: CMA303
IB: CMA301
C
ANET-seq/unph
ANET-seq/S2P
ANET-seq/S5P
ANET-seq/unph+ph
ChrRNA-seq
Read density (RPKM)
Distance (bp) to TSS or TES
-1k
TSS
+1k
-0.5k
TES
+3k
D
Scale
ChrX: 70,585,500
1 kb
hg19
70,587,000
TSS
PROMPT
TAF1
ANET-seq/unph
ANET-seq/S2P
ANET-seq/S5P
ANET-seq/unph+ph
ChrRNA-seq
(+) strand
(-) strand
Read per 10^8 sequences
E
Scale
Chr10:
62,560,000
10 kb
hg19
62,575,000
TES
CDK1
ANET-seq/unph
ANET-seq/S2P
ANET-seq/S5P
ANET-seq/unph+ph
ChrRNA-seq
Read per 10^8 sequences

Figure 2. ANET-seq with different Pol II modifications

(A) Diagram showing different Pol II antibody epitopes on CTD (Stasevich et al., 2014).

(B) Pol II precipitated from cell extracts with indicated antibodies detected by western blot using each antibody.

(C) Metagene analyses of ANET-seq on TSS and TES. Read density (RPKM) of ANET-seq databases was plotted around TSS (+/- 1 kb) and TES (-0.5k~+3 kb). Metagene of ChrRNA-seq is shown as input. Each metagene has different scales on y-axis. Data are represented as mean +/- SE from 1,647 genes.

(D) ANET-seq profiles on TSS of TAF1 gene. Read density; read per 10^8 sequences. Each ANET-seq has different y-axis scale.

(E) CDK1 gene ANET-seq profiles at TES. All ANET-seq data are shown on same y-axis scale.

Based on previously published RNA-seq data (Lacoste et al., 2014) , we found 11,560 (45%) of RefSeq genes are actively transcribed in our HeLa cell line. However to avoid noise caused by over-represented sequences from ncRNA (such as rRNA, tRNA, snoRNA and snRNA) in the ANET-seq metagene analysis, we excluded genes that overlap with these sequences. We also excluded overlapping gene transcription units as these might give bioinformatic bias such as pseudo-antisense transcripts from neighboring genes in the TES (Figure S4A). Finally we selected only isolated genes that have no other transcription unit within -1 kb of the TSS or +3 kb of TES (Figure S4B). We were therefore left with 1,647 protein-coding genes to study in metagene analyses of our ANET-seq data (Figure S4C). These data reveal striking differences between the four different antibody IPs used in our ANET-seq analysis. 8WG16 and CMA301 (ANET-seq/unph and ANET-seq/unph+ph, respectively) display substantial bidirectional (sense and antisense) peaks at the TSS. However CMA302 and CMA303 (ANET-seq/S2P and ANET-seq/S5P, respectively) show lower TSS peaks (Figure 2C and 2D). In contrast ANET-seq/S2P gives more signal at TES than ANET-seq/unph and ANET-seq/S5P (Figure 2C and 2E), consistent with the gene specific ChIP profiles (Figure S3). Unexpectedly, ANET-seq/S5P does not display a major TSS peak in contrast to previously published Pol II S5P ChIP profiles (Heidemann et al., 2013).

4.3 Exon tethering to Pol II S5P for co-transcriptional splicing

The coupling of Pol II transcription to splicing is now well established (David and Manley, 2011; Moore and Proudfoot, 2009). Thus phosphorylated Pol II CTD (S2P) recruits splicing factors to enhance pre-mRNA splicing efficiency (Ahn et al., 2004; Hirose and Manley, 1998). Also altered Pol II elongation speed can affect alternative splicing patterns (Ip et al.,

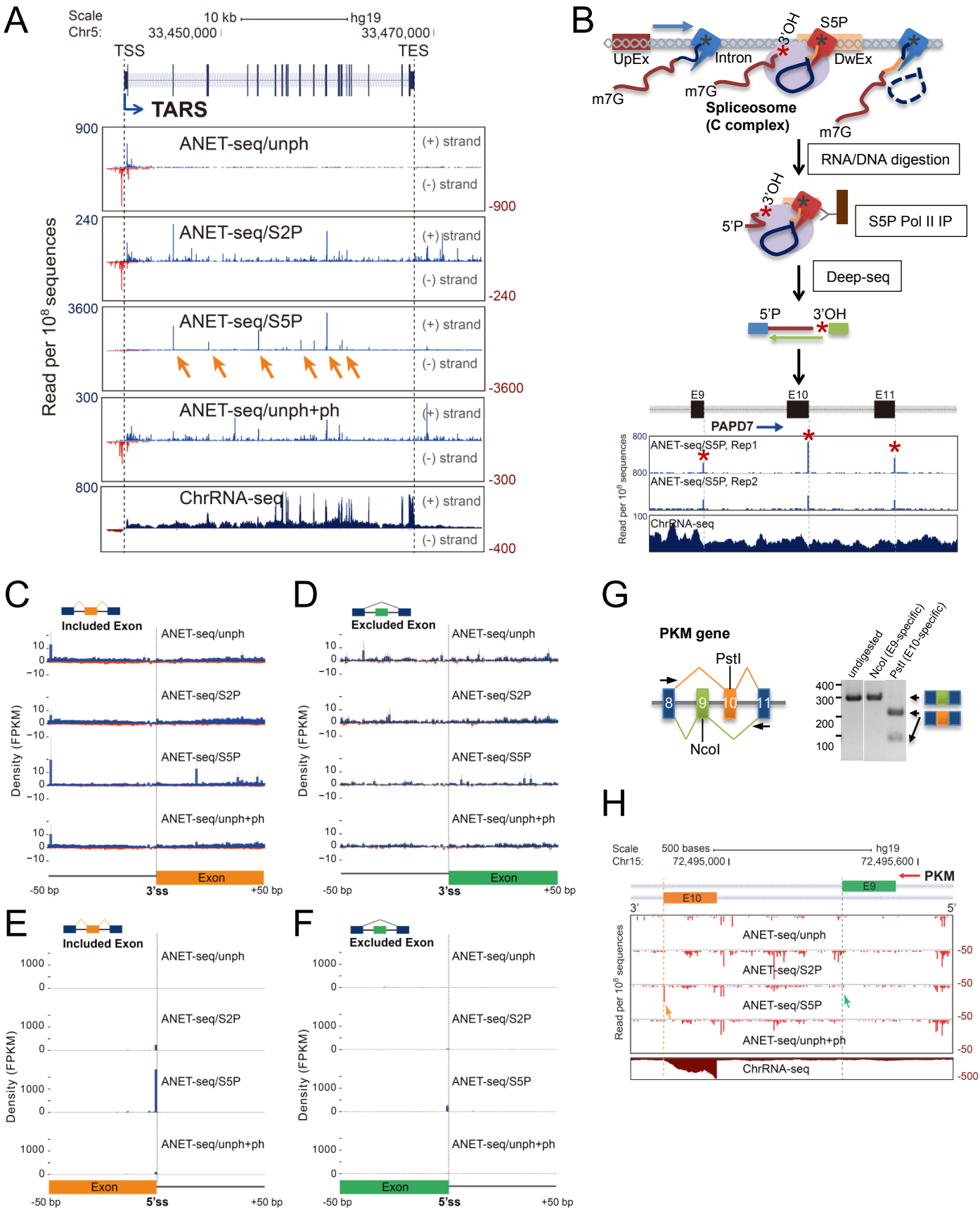

A
Scale
Chr5:
10 kb
hg19
33,450,000
33,470,000
TSS
TES
TARS
ANET-seq/unph
ANET-seq/S2P
ANET-seq/S5P
ANET-seq/unph+ph
ChrRNA-seq
(+) strand
(-) strand
Read per 10^8 sequences
B
UpEx
Intron
DwEx
S5P
3'OH
m7G
Spliceosome
(C complex)
RNA/DNA digestion
5'P
S5P Pol II IP
Deep-seq
E9
E10
E11
PAPD7
ANET-seq/S5P, Rep1
ANET-seq/S5P, Rep2
ChrRNA-seq
C
Included Exon
D
Excluded Exon
Density (FPKM)
Exon
-50 bp
3'ss
+50 bp
E
F
5'ss
G
PKM gene
PstI
NcoI
undigested
NcoI (E9-specific)
PstI (E10-specific)
H
Scale
Chr15:
500 bases
72,495,000
72,495,600
PKM

Figure 3. Exon tethering to Ser5-phosphorylated Pol II complex

(A) TARS ANET-seq profile with different antibodies. S5P-dominant peaks are indicated by orange arrows.

(B) Co-transcriptional splicing model. 3'OH of upstream exon (UpEx, dark red). RNA and catalytic site in Pol II are shown as red and black asterisks. 3' OH of the UpEX RNA is protected in S5P Pol II (red)-spliceosome C complex (purple circle) and mapped at 3' ends of PAPD7 exons 9, 10 and 11 in two independent replicates of ANET-seq/S5P data.

(C and D) Metagene of ANET-seq data over 5' ends (3'SS) of included (spliced) exons (C, orange rectangle) and excluded exons (D, green rectangle).

(E and F) Metagene of ANET-seq data around 3' ends (5' SS) of included exons (C, orange rectangle) and excluded exons (D, green rectangle). Data are mean +/- SE from 3,115 and 304 genes for excluded and included exons.

(G) PKM exons 8-11 are illustrated. Exon 9 (green) and exon10 (orange) are mutually exclusive. PCR primers indicated. RT-PCR products were digested with indicated exon-specific restriction enzyme (NcoI or PstI).

(H) ANET-seq data around mutually exclusive exons 9 and 10 of PKM. ANET-seq/S5P signals at 3' end of exon 9 and exon 10 are shown by green and orange arrows.

2011; Kornblihtt et al., 2004; Munoz et al., 2009). This is taken to indicate that Pol II slows down near splice sites (SS) to promote spliceosome assembly. In particular genome-wide analysis of nascent RNA by high-resolution tiling arrays in yeast has shown that Pol II is paused over terminal exons, but only for co-transcriptionally spliced genes (Carrillo Oesterreich et al., 2010). Additionally, precisely timed ChIP analysis in yeast showed that phosphorylated Pol II (S5P CTD) accumulates over the 3' SS of intron containing genes. Furthermore this splicing-dependent Pol II pausing requires pre-spliceosome assembly (Alexander et al., 2010; Chathoth et al., 2014).

We were interested to determine if our ANET-seq profiles reflect the co-transcriptionality of splicing but observed unexpected patterns. First we present the ANET-seq profile of a specific gene, TARS where we have compared its ANET-seq profiles between the four different Pol II antibodies (Figure 3A). Surprisingly, ANET-seq/S5P in particular detected prominent peaks in gene exons. We have reasoned that ANET-seq will specifically identify the nascent transcript 3'OH in the Pol II active site. However as previously noted (Churchman and Weissman, 2011) co-precipitated spliceosomes will also contain 3'OH RNA derived from intermediates in the splicing reaction. These 3'OH will potentially yield ANET-seq signal. Remarkably, ANET-seq/S5P on the PAPD7 gene yields peaks that are exactly located at exon 3' ends (Figure 3B). These observations suggest that ANET-seq/S5P detects the 5'SS cleavage splicing intermediate. This indicates that spliceosome complex C is directly associated with Pol II CTD S5P. To extend these observations we performed metagene analyses on exons that are either included or excluded in the mature mRNA, looking 50 nt upstream or downstream of exons. Notably the ANET-seq data shows a strong S5P specific peak at the 5'SS of included but not excluded exons. (Figure 3C-F). This result confirms that the ANET-seq 5'SS

signals derive from active splicing. We also studied the alternatively spliced (mutually exclusive) exon 9 and 10 of PKM. RT-PCR and ChrRNA-seq analyses show that exon 10 is predominantly included in mature PKM transcripts in HeLa cells (Figure 3G) (David et al., 2010). ANET-seq/S5P signals are largely accumulated at 3' end of exon 10 of PKM (Figure 3H). Together with metagene analyses, these results strongly suggest that spliceosomes are tethered to Pol II to promote co-transcriptional splicing. Remarkably, this first splicing step is specific to Pol II CTD S5P.

4.4 Co-transcriptional pre-miRNA biogenesis

Most pre-microRNAs (miRNA) are present within the introns of protein coding genes, where they are excised co-transcriptionally by the microprocessor complex, containing Drosha and DGCR8 (Morlando et al., 2008; Pawlicki and Steitz, 2008). Drosha cleavage generates 3'OH ends which have the potential for detection by ANET-seq. Since RNA cleavage sites on pre-miRNA generated by the microprocessor complex are highly variable, we individually checked the ANET-seq profiles for each pri-miRNA that is highly expressed in HeLa cells. We detect two peaks defining the pre-miRNA 5' and 3' ends for intronic hsa-mir-27b where the 3' peak is dominant. In contrast for intronic hsa-let-7g, the 5' peak is dominant (Figure 4A and 4B). Additionally we often observe a single 5' peak of ANET-seq for pre-miRNA sequences such as hsa-miR26b (Figure 4C). Interestingly, 5' end and 3' end peaks correspond to the 3' ends of the cleaved intron and the pre-miRNA which reaffirms the co-transcriptionality of pre-miRNA processing. As with spliceosomes we suggest that that microprocessor is co-precipitated with Pol II so that 3'OH intermediates of Drosha cleavage are detected by ANET-seq.

Two pre-miRNAs (hsa-mir181a-1 and hsa-mir181b-1) are located in the MIR181A1HG intron (Figure 4D). Although ENCODE Project (Consortium et al., 2012) shows both mature miRNAs are expressed in HeLa cells, only hsa-mir181a-1 yields significant ANET-seq peaks.

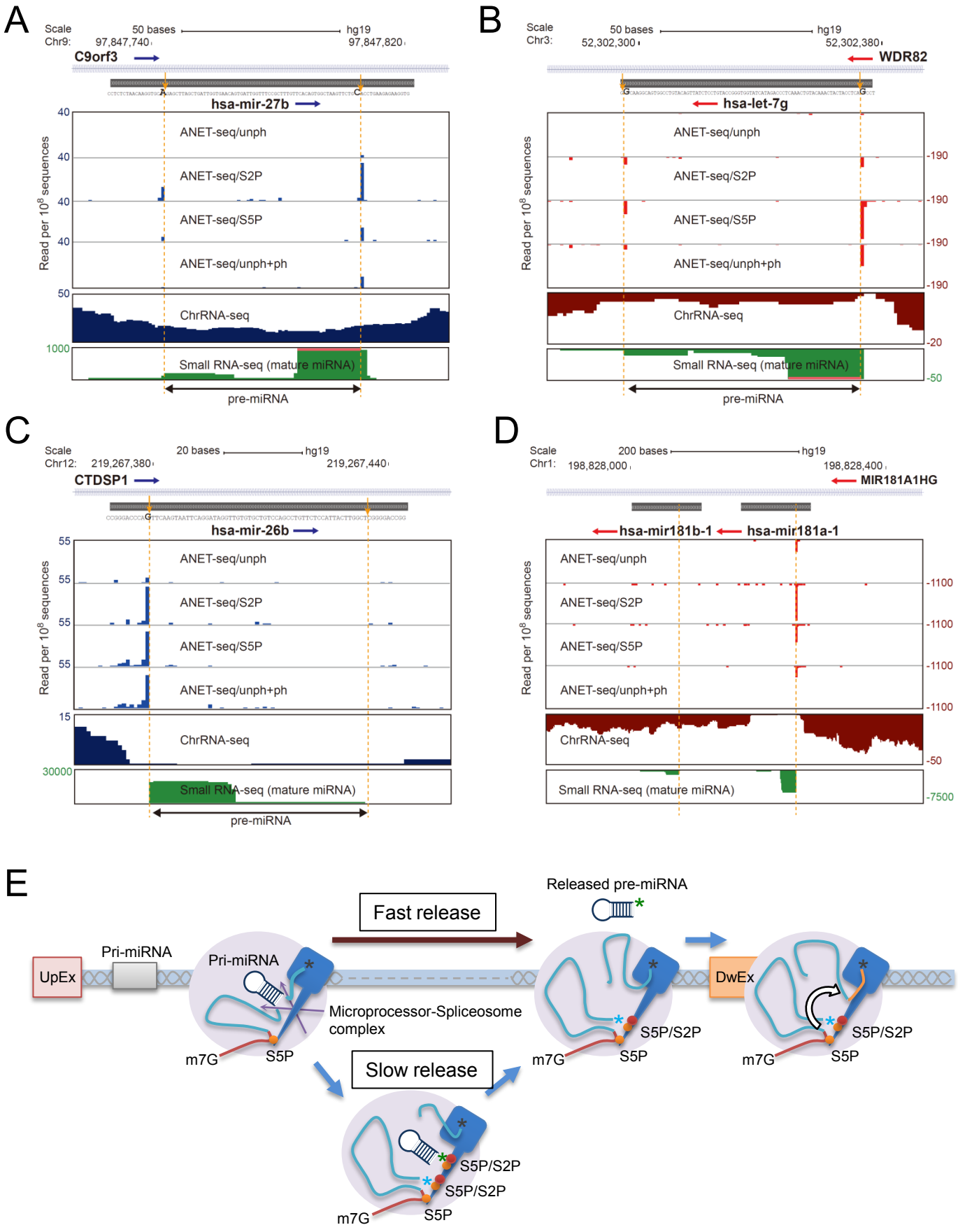

A
Scale
Chr9:
50 bases
hg19
97,847,740
97,847,820
C9orf3
hsa-mir-27b
ANET-seq/unph
ANET-seq/S2P
ANET-seq/S5P
ANET-seq/unph+ph
ChrRNA-seq
Small RNA-seq (mature miRNA)
pre-miRNA
Read per 10^8 sequences
B
Scale
Chr3:
50 bases
hg19
52,302,300
52,302,380
WDR82
hsa-let-7g
ANET-seq/unph
ANET-seq/S2P
ANET-seq/S5P
ANET-seq/unph+ph
ChrRNA-seq
Small RNA-seq (mature miRNA)
pre-miRNA
Read per 10^8 sequences
C
Scale
Chr12:
20 bases
hg19
219,267,380
219,267,440
CTDSP1
hsa-mir-26b
ANET-seq/unph
ANET-seq/S2P
ANET-seq/S5P
ANET-seq/unph+ph
ChrRNA-seq
Small RNA-seq (mature miRNA)
pre-miRNA
Read per 10^8 sequences
D
Scale
Chr1:
200 bases
hg19
198,828,000
198,828,400
MIR181A1HG
hsa-mir181b-1
hsa-mir181a-1
ANET-seq/unph
ANET-seq/S2P
ANET-seq/S5P
ANET-seq/unph+ph
ChrRNA-seq
Small RNA-seq (mature miRNA)
Read per 10^8 sequences
E
Released pre-miRNA
Fast release
UpEx
Pri-miRNA
Pri-miRNA
Microprocessor-Spliceosome complex
DwEx
m7G
S5P
S5P/S2P
Slow release

Figure 4. Pre-miRNA biogenesis from protein coding gene introns

(A-D) ANET-seq with different Pol II antibodies versus ChrRNA-seq over intronic pre-miRNAs, hsa-mir-27b (A), hsa-let-7g (B), hsa-mir-26b (C) and hsa-mir181a/b-1 (D) denoted by black rectangles. Frequent RNA cleavage sites identified by orange arrows and dashed lines. Small RNA-seq data are shown at bottom (green).

(E) Model of co-transcriptional pre-miRNA biogenesis. UpEx, DwEx and pre-miRNA DNA sequences in red, orange and grey. Co-transcriptional RNA cleavage by microprocessor and spliceosome (light blue) shown with 3' end of cleaved RNA and pre-miRNA tethered to phosphorylated CTD. Pre-miRNA release may occur from the transcription complex, fast (dark red arrow) or slow (blue arrows). Finally 5'SS is spliced to 3'SS in spliceosome. Curved arrow denotes exon splicing.

This correlates with ChrRNA-seq analysis showing a signal window over has-mir181a-1 but not b-1. We infer that only a-1 is co-transcriptionally processed. Evidently ANET-seq can be used to distinguish co-transcriptional and post-transcriptional pre-miRNA processing. We also note that the variable ANET-seq double peaks (i.e. hsa-mir-27b) and single peaks (i.e. hsa-mir-26b) suggest kinetic differences in pre-miRNA biogenesis. Some pre-miRNAs (such as pre-miRNA-26b and 181a-1) may be released immediately from the Pol II elongation complex after microprocessor cleavage (see model, Figure 4E). Other pre-miRNAs (such as pre-miRNA-27b and let-7g) may be more slowly released with the 3' ends of the pre-miRNA still tethered to the Pol II elongation complex. Significantly ANET-seq/S2P and S5P show larger peaks than ANET-seq/unph for pre-miRNA processing suggesting that CTD phosphorylation is important for co-transcriptional pre-miRNA biogenesis.

Four additional examples of pre-miRNA containing loci (Figure S5) emphasize the generality of our ANET-seq data. For MIR17HG locus containing six tandem pre-miRNA (Figure S5B) Drosha co-transcriptionally cleaves the outer pre-miRNA. However more inner pre-miR18a and pre-miR19a appear to be processed post-transcriptionally as judged by a lack of ANET-seq peaks and the absence of a hole in the ChrRNA-seq profile over these sequences.

4.5 Pol II pausing regulated by CPA factors at TES

Depletion of CPA factors (CPSF73 and CstF-64+CstF-64 tau) and Xrn2 proteins was performed by siRNA transfection and the protein level reductions were monitored by western blot using the indicated antibodies (Figure 5A-C, left panels). ChrRNA-seq analyses (both for specific genes and by metagene analysis) demonstrated clear Pol II termination defects following depletion of CPA factors (Figure 5A and 5B) We note that double-knockdown of CstF-64 and CstF-64 tau proteins was necessary to detect termination defects due to the functional redundancy in HeLa cells (Yao et al., 2012). Xrn2 knockdown showed no termination defect at protein-coding gene TES (Figure 5C) as suggested previously (Brannan

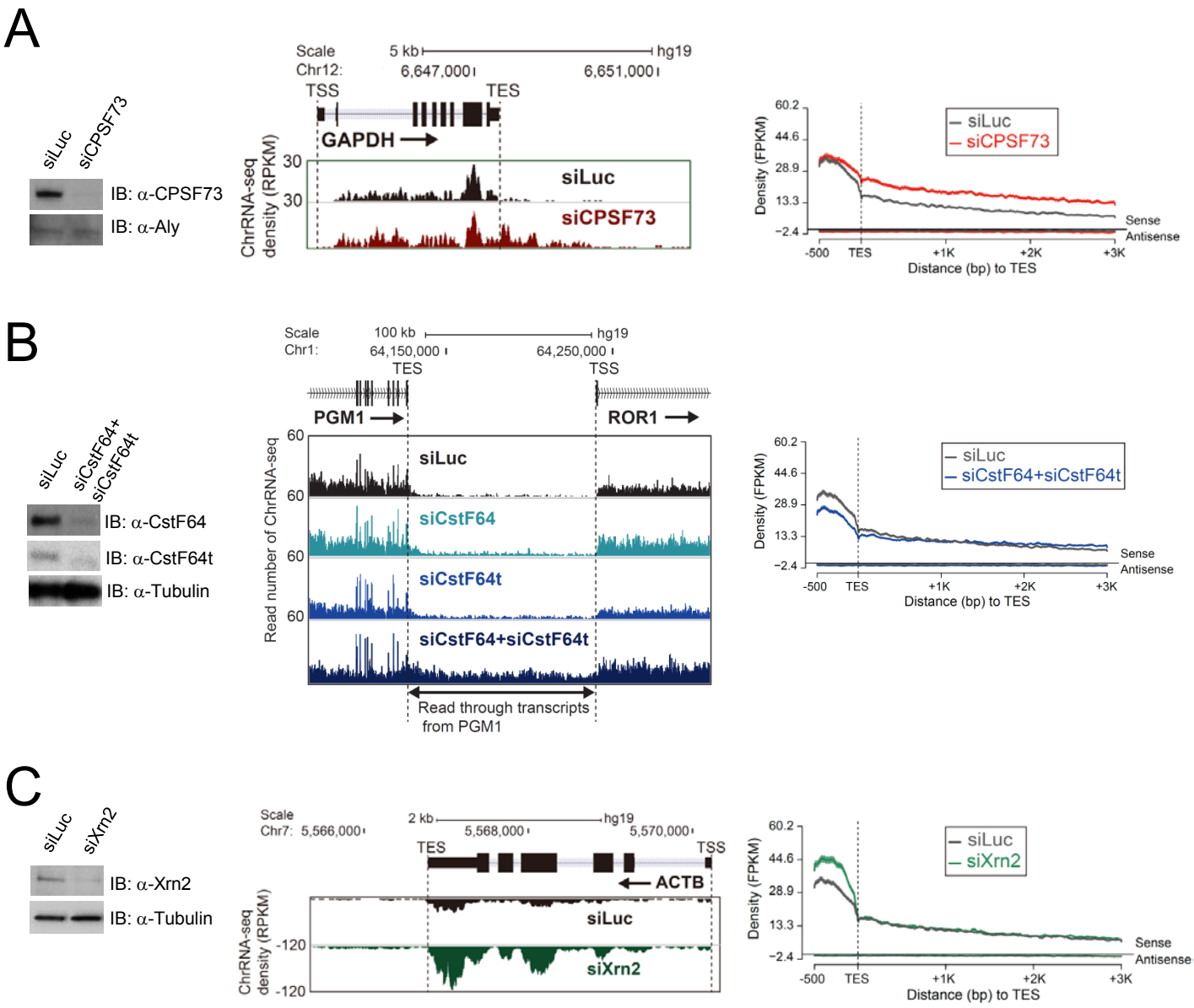


Figure 5. ChrRNA-seq reveals CPA knockdown causes TES termination defect
(A-C) Western blots showing knockdown efficiencies of siRNA treatments for CPSF73, CstF64/Tau and Xrn2. Aly and Tubulin proteins are loading controls (A) Termination defect detected following depletion of CPSF73 protein (red) on GAPDH gene and metagene profile over TES. n=1,647 (B) Additive termination defect seen following double knockdown of CstF64 and CstFTau (turquoise and blue and dark blue double) on PGM1 and metagene profile over TES. n=1,647. (C) No termination defect detected following Xrn2 depletion (green) on ACTB and metagene profile over TES. n=1,647.

et al., 2012). Possibly like CstF64 this factor acts redundantly with other termination factors. Interestingly Xrn2 depletion substantially increased transcript levels within the gene body suggesting a major role for Xrn2 in nuclear turnover (Davidson et al., 2012).

To extend our termination studies to ANET-seq we employed the CMA302 (S2P) Pol II antibody, as S2P CTD strongly correlates with 3' end processing (Ahn et al., 2004; Hirose and Manley, 1998). Metagene analyses of ANET-seq/S2P using control siRNA treatment (siLuc), illustrated significant Pol II pausing at the TES (Figure 6A) as with untreated cells (Figure S6). Interestingly, depletion of CPSF73 and CstF-64+CstF-64 tau proteins substantially reduced S2P Pol II pausing over the TES (Figure 6A, top and middle). In contrast Xrn2 knockdown showed no significant difference to the siLuc control (Figure 6A, bottom). We also observe that ANET-seq/S2P profiles upon knockdown of CPA factors crossed over the siLuc control profile approximately 2.5 kb downstream of the TES (Figure 6A, top and middle). ANET-seq on the specific genes GABARAPL1 and SMOC1 revealed that S2P Pol II pausing was suppressed by depletion of CPA factors and both ANET-seq and ChrRNA-seq show clear termination defects. Again Xrn2 depletion showed no significant effects in these assays (Figure 6B). A further S2P Pol II pausing effect was detected 10kbp downstream of the SMOC1 TES suggesting that S2P Pol II is paused in the 3' flanking region in a CPA factor independent manner. This effect may relate to nucleosome barriers, as previously described (Grosso et al., 2012).

We examined the Read-Through Index (see Experimental Procedure) following termination factor knockdown (Figure 6C). RTI demonstrates that depletion of CPA factors, but not Xrn2 decreases S2P Pol II occupancy within 2 kb downstream of the TES. This indicates that Xrn2 does not have a unique role in Pol II termination at TES. In contrast CPA factors promote Pol II pausing to enhance PAS recognition and PAS-dependent termination (Figure 6D).

We also analyzed the ANET-seq and ChrRNA-seq profiles for replication dependent histone genes (Schumperli, 1988). These small Pol II transcripts are intronless, not polyadenylated and associated with different Pol II CTD modifications; S7P (Egloff et al., 2007) and T4P (Hsin et al., 2011). We show that ANET-seq profiles appear quite different for these genes as compared to other protein coding genes. No TSS associated pausing or antisense transcription is evident and highest signals were observed for unphosphorylated CTD (Figure S7A). Also depletion of termination factors had different affects. CPSF73 knockdown gave a clear termination defect based on ChrRNA-seq (Figure S7B) consistent with the known association of CPSF with the histone 3’processing machinery (Kolev and

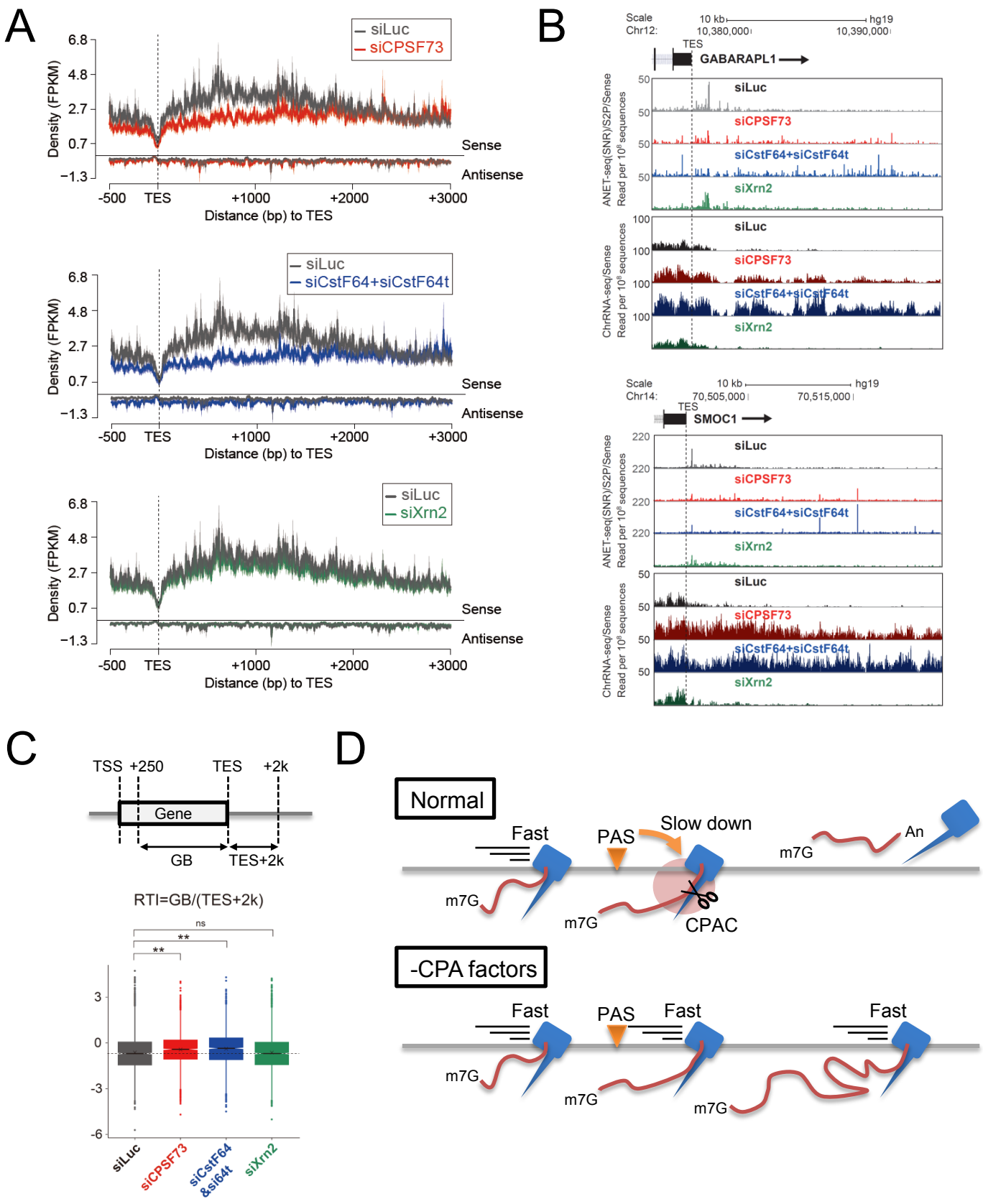

A
siLuc
siCPSF73
Density (FPKM)
6.8
4.8
2.7
0.7
−1.3
Sense
Antisense
-500
TES
+1000
+2000
+3000
Distance (bp) to TES
siLuc
siCstF64+siCstF64t
siLuc
siXrn2
B
Scale
10 kb
hg19
Chr12:
10,380,000
10,390,000
TES
GABARAPL1
ANET-seq(SNR)/S2P/Sense
Read per 10⁸ sequences
siLuc
siCPSF73
siCstF64+siCstF64t
siXrn2
ChrRNA-seq/Sense
Chr14:
70,505,000
70,515,000
SMOC1
C
TSS
+250
TES
+2k
Gene
GB
TES+2k
RTI=GB/(TES+2k)
ns
**
**
3
0
-3
-6
siLuc
siCPSF73
siCstF64 &si64t
siXrn2
D
Normal
Fast
PAS
Slow down
An
m7G
CPAC
-CPA factors

Figure 6. Nascent RNA within S2P Pol II complex at TES.

(A) Metagene analysis of ANET-seq/S2P over TES regions following termination factor depletions (Fig. 5). Data are mean +/- SE from 1,647 genes.

(B) ANET-seq/S2P (top) and ChrRNA-seq (bottom) of GABARAPL1 and SMOC1 gene TES from indicated siRNA treated HeLa cells.

(C) Read-Through Index (RTI) of ANET-seq/S2P following indicated knockdowns. RTI scheme is shown. Gene body (GB) signals were divided by signals in 2kb region from TES (TES+2k) for RTI (see Experimental Procedures). X in each boxplot marks the mean, and the dashed line is the median of siLuc. n=2,624. (**) P-value < 2 x 10^{-15} by two-sided Mann-Whitney test; (ns) indicates no difference between samples (p-value = 0.9894 by two-sided Mann-Whitney test).

(D) Model correlating Pol II pausing and PAS-dependent transcription termination at TES. RNA cleavage (scissors) by CPA complex (red circle) at PAS (orange triangle). Pol II elongation speed over 3' flank region is regulated by PAS recognition.

Steitz, 2005). Neither CstF64 with CstF64t nor Xrn2 depletion caused termination defects. Notably Xrn2 depletion significantly increased ChrRNA-seq levels across histone genes implying a major role in RNA stability (Davidson et al., 2012). Finally these termination factor knockdowns had no effect on TES pausing, in contrast to other protein coding genes (Figure S7C). Overall ANET-seq on histone gene nascent transcription reveals the potential for major differences between different gene classes.

4.6 3' end termination machinery regulates metabolism of promoter-associated RNA

Although RNA cleavage sites have been previously identified near TSS (Almada et al., 2013), which factors are involved in this process has not been determined. We therefore performed metagene analyses across TSS using ANET-seq/S2P following knockdown of CPA factors and Xrn2. CPSF73 contains the CPA cleavage activity and so could potentially cleave nascent RNA near the TSS by recognition of cryptic PAS. Interestingly we observe an equivalent increase in TSS-associated S2P Pol II pausing on both mRNA and PROMPT strands after depletion of CPA factors and Xrn2 (Figure 7A). Metagene analysis of CstF-64+CstF-64 tau double-knockdown shows an average 3.6 fold increase as compared to siLuc (Figure 7A middle). Also CPSF73 and Xrn2 knockdowns both show an average 2.3 fold increase in Pol II pausing. These effects extend from TSS+30 to TSS+100 on both mRNA and PROMPT strands (Figure 7A, top and bottom). Similar effects (average 3.1 (max 9.7), 6.0 (max 19), 5.7 (max 26.4) fold increase with siCPSF73, siCstF-64+siCstF-64 tau and siXrn2, respectively) were observed for the SLC30A6 gene (Figure 7B). The Escaping Index (EI) on over 2624 genes show that depletion of all three factors increases promoter-associated S2P Pol II pausing (Figure 7C). Additionally, EI also demonstrates that all three factors when knocked down have no effect on S2P Pol II distribution across the gene body (Figure 7C).

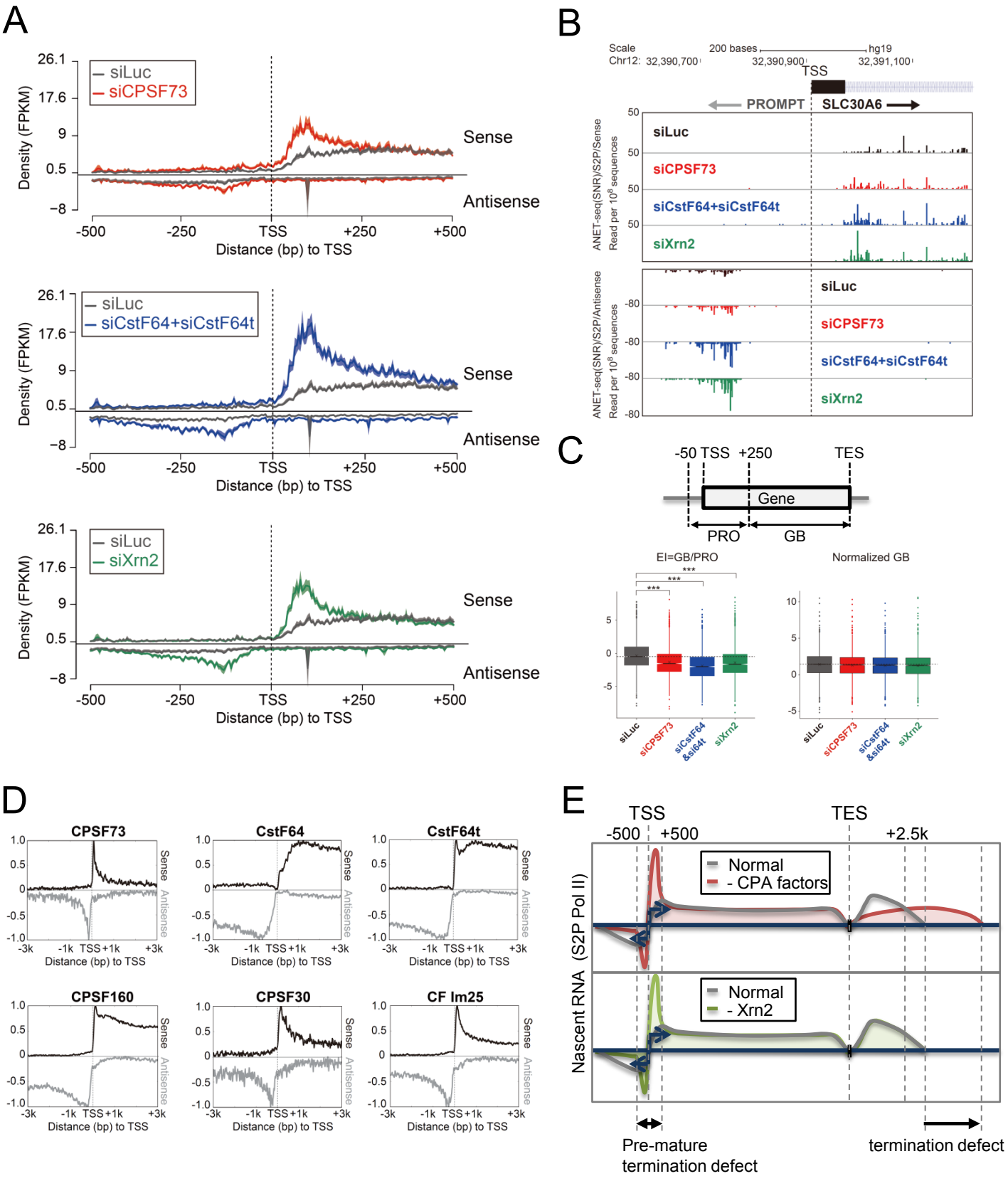

A
siLuc
siCPSF73
siCstF64+siCstF64t
siXrn2
Density (FPKM)
Sense
Antisense
Distance (bp) to TSS
B
Scale
Chr12: 32,390,700
200 bases
32,390,900
32,391,100
hg19
TSS
PROMPT
SLC30A6
ANET-seq(SNR)/S2P/Sense
ANET-seq(SNR)/S2P/Antisense
Read per 10^8 sequences
C
-50 TSS
+250
TES
Gene
PRO
GB
EI=GB/PRO
Normalized GB
D
CPSF73
CstF64
CstF64t
CPSF160
CPSF30
CF Im25
E
TSS
TES
-500
+500
+2.5k
Normal
- CPA factors
- Xrn2
Nascent RNA (S2P Pol II)
Pre-mature
termination defect
termination defect

Figure 7. Promoter-associated RNA metabolism regulated by termination factors.

(A) Metagene analyses of ANET-seq/S2P following knockdown of 3' end termination factors (Figure 5) at TSS. ANET-seq/S2P from siLuc, siCPSF73, siCstF64+siCstF64t and siXrn2 treated cells. Data are mean +/- SE from 1,647 genes.

(B) ANET-seq/S2P maps with indicated knockdowns around TSS of SLC30A6 gene on both mRNA and PROMPT strands.

(C) Escaping Index (EI) and normalized gene body (GB) profiles of ANET-seq/S2P. Representation of EI is shown above. GB signals were divided by signals in promoter region (PRO, -50 to +250 bp over TSS) for EI. The EI (left) and normalized GB (right) with indicated siRNA treatments are shown below. (***) P-value < 2.2×10^{-16} by two-sided Mann-Whitney test.

(D) CLIP analysis of CPA factors (Martin et al., 2012). Relative normalized counts and distance from TSS are shown at Y- and X-axis.

(E) Model showing effects of CPA factors and Xrn2 throughout a gene highlighting TSS and TES differences.

These results indicate that CPA factors and Xrn2 are involved in the metabolism of promoter-associated non-productive transcripts.

In order to examine whether CPA factors could directly bind to nascent RNA near TSS, we analyzed in vivo cross-linking and immuno-precipitation (CLIP) data which has been published for a genome wide alternative polyadenylation (APA) study at TES (Martin et al., 2012). Surprisingly all CPA factors, including CPSF73, CstF-64, CstF64 tau, CPSF160, CPSF30 and CF Im25 proteins, are significantly detected on both strands within 500 nt of the TSS. Especially CPSF73 shows a substantial peak 160 nt upstream and 80 nt downstream of TSS (Figure 7D and Supplementary Table S1). Together with our ANET-seq/S2P results, we conclude the CPA complex cleaves not only pre-mRNA at the PAS to promote 3' end termination, but also promotes promoter-associated premature termination (Figure 7E). Notably Xrn2 plays a unique role in TSS but not in TES termination.

5. Discussion

We present a powerful high-throughput sequencing strategy for mapping nascent RNA within the elongating Pol II complex across the human genome referred to as ANET-seq. This approach reveals precise maps of not only nascent RNA, but also the associated Pol II "CTD code". It is widely known that Pol II CTD heptad repeats are dynamically modified during the transcription cycle in eukaryotes (Brookes and Pombo, 2009; Egloff et al., 2012a; Heidemann et al., 2013). Thus we employed a S2P Pol II specific antibody to monitor transcription termination events (Figure 6), since CTD Ser^2 is highly phosphorylated at the TES where it acts to recruit the cleavage and polyadenylation (CPA) complex (Ahn et al., 2004). S2P CTD has also been shown to enhance CPA activity in vitro (Hirose and Manley, 1998). ANET-seq/S2P illustrates more coverage over the TES compared to other Pol II antibodies (Figure 2E). It also shows clear termination defects upon knockdown of CPA factors (Figure 6A). These data were coupled with Chromatin associated RNA sequencing (ChrRNA-seq).

We also used the S5P CTD specific antibody in ANET-seq analysis. Surprisingly, this demonstrates peaks at the 3' ends of actively spliced exons (Figure 3C-F) indicating that the upstream exon within the spliceosome is tethered to the Pol II elongation complex in a S5P dependent manner. Unspliced exons show much less peak compared to actively spliced exons. Furthermore the mutually exclusive exons of PKM show a selective peak of ANET-seq/S5P on exon 10, which is predominantly selected in HeLa cells (Figure 3H). Our ANET-seq technology will provide a novel way to unravel the complexity of the co-transcriptional splicing mechanism since it is possible to isolate a native splicing intermediate (C complex) in vivo (Figure 3B). Additionally, this technology may be useful to characterize recursively spliced introns as reported in *Drosophilla* (Burnette et al., 2005; Hatton et al., 1998). Thus ANET-seq/S5P peaks across introns may signify recursive 5'SS.

It has been reported that other CTD amino acids are highly phosphorylated during active transcription. For instance, phosphorylation of CTD Ser^7 (S7P) is important to recruit Integrator complex. This regulates 3' end processing of snRNA genes and so facilitates transcription termination (Egloff et al., 2007; Egloff et al., 2012b). Additionally, ChIP analysis in yeast has shown that S7P Pol II is intron enriched suggesting a link to pre-mRNA splicing (Kim et al., 2010). Mutation of CTD Thr^4 specifically represses histone gene expression suggesting that T4P is required for histone mRNA 3' end processing (Hsin et al., 2011). Another CTD phosphorylation Tyr^1 (Y1P) stimulates the binding of elongation factor Spt6

and blocks recruitment of termination factors in yeast (Mayer et al., 2012). Use of these different phosphorylation-specific Pol II antibodies may provide comprehensive maps of nascent RNA with all mammalian CTD codes.

The mechanistic and kinetic link between Pol II transcription and pre-mRNA splicing is well established (David and Manley, 2011; Moore and Proudfoot, 2009; Shukla and Oberdoerffer, 2012). In yeast, high resolution nascent RNA mapping and ChIP experiments have demonstrated that splicing-dependent Pol II pausing occurs in intron containing genes (Alexander et al., 2010; Carrillo Oesterreich et al., 2010; Chathoth et al., 2014). Similarly in *Drosophila*, Pol II pausing at 3' SS was detected by PRO-seq analysis (Kwak et al., 2013). However, the connection between co-transcriptional splicing and Pol II pausing in mammals has not been described. It has however been reported that phosphorylated Pol II CTD is important to recruit splicing factors onto spliced exons and so facilitate splicing efficiency (Ahn et al., 2004; Hirose et al., 1999). Furthermore alternative splicing of the multi exonic CD44 gene is associated with accumulation of S5P over variant exons (Batsche et al., 2006). Our nascent RNA profiles from ANET-seq (Figure 3) suggest new interpretations. Thus ANET-seq signals are enriched at 5'SS rather than at intron 3' ends as seen in yeast (Alexander et al., 2010). This may relate to differences between the intron definition model proposed for yeast and exon definition models for human splicing. It is thought that introns are recognized for splicing in lower species, since their lengths are generally much shorter than in mammals. On the other hand, an exon may need to be preferentially recognized by Pol II in mammals since here exons represent a very small part of the mainly intronic pre-mRNA. Our results are also consistent with exon-tethering models where upstream exons are retained on the elongating Pol II complex to facilitate splicing with downstream exons (Dye et al., 2006). Importantly, CTD S5P is involved in this exon-tethering model. It remains a possibility that other CTD modifications are also required for intron definition.

A substantial fraction of pre-miRNA are found in the introns of protein coding genes (Rodriguez et al., 2004). We show ANET-seq peaks that precisely delineate these intronic pre-miRNA sequences and are enriched for S2P and S5P (Figure 4A-D). Previous reports indicate co-transcriptional pre-miRNA processing can occur on chromatin by recruiting the microprocessor complex (Drosha and DGCR8 proteins) to these pri-miRNA sequences (Morlando et al., 2008). We show here that the kinetics of pre-miRNA biogenesis varies involving both co-transcriptional and post-transcriptional pre-miRNA processing events.

Pol II accumulation at TES also has been revealed by ChIP experiments and GRO-seq analysis (Core et al., 2008; Davidson et al., 2014; Proudfoot, 2011). It was thought that Pol II

pausing at TES regulates transcription termination, based on NRO analysis (Gromak et al., 2006). In addition, PAS comprising both an AAUAAA core sequence and downstream GU rich sequence element (DSE) are required for cleavage and polyadenylation (CPA) at the TES. Biochemical experiments isolated and characterized the cleavage and polyadenylation specificity factor (CPSF) complex and cleavage stimulating factor (CstF) complex from HeLa nuclear extracts. These protein complexes recognize the AAUAAA and GU-rich DSE, respectively. Importantly, CPA is functionally linked to Pol II transcription termination in vivo (Proudfoot, 2011). Here, we depleted components of the CPA complex (CPSF73 and CstF-64+CstF-64 tau) in HeLa cells using siRNA technology to examine the effect on Pol II pausing at TES. Consistent with previous reports, ChrRNA-seq reveals that siRNA-mediated CPSF73 and CstF-64 depletion causes transcriptional termination defects on protein-coding genes (Figure 5). Interestingly, our ANET-seq data shows that depletion of CPA factors causes significantly less pausing immediately downstream of TES (<2 kb from TES) and then more Pol II occupancy at further downstream region (> 2kb from TES) compared to siLuc transfected cells (Figure 6A). This result indicates that Pol II elongation speed is regulated by the CPA complex which may be important to mediate transcription termination at protein coding gene TES (Figure 6C). Moreover, depletion of CPA factors in some cases caused additional pausing further downstream of PAS-dependent Pol II pause site (Figure 6B, for examples GABARAPL1 and SMOC1 genes). This suggests other Pol II pausing mechanisms exist such as nucleosome barriers (Grosso et al., 2012; Mavrich et al., 2008), road blocks caused by DNA-binding protein (Shukla et al., 2011) or co-transcriptional RNA cleavage (CoTC) (Dye and Proudfoot, 2001; Nojima et al., 2013) in the termination region, possibly acting as fail-safe termination mechanisms.

We also demonstrate that no significant termination defect occurs following the TES upon knockdown of Xrn2 (Figure 6A, bottom). This observation is inconsistent with our previous reports which employed plasmid-based transfection studies (West et al., 2004). Additionally, it has been shown recently that Xrn2 has a required partner protein TTF2 for transcription termination (Brannan et al., 2012). It seems likely that Xrn2 associated termination is redundant with other termination factors.

Unexpectedly, ANET-seq analysis showed a drastic increase in Pol II pausing at the TSS (<100 base) for both mRNA and PROMPT transcription upon knockdown of CPA factors. Additionally, depletion of 5'-3' exonuclease Xrn2 also showed a similar increase in Pol II pausing at the TSS. This result suggests that Xrn2 is involves in premature termination at the TSS even though it may not play such a critical role at the TES (Brannan et al., 2012).

Although CPA factors and Xrn2 affects Pol II occupancy at TSS, all three protein knockdowns show no difference in Pol II distribution across the gene body.

Recent studies have pointed towards differences between promoter proximal termination for mRNA sense or antisense RNA (Almada et al., 2013; Grzechnik et al., 2014; Ntini et al., 2013). Antisense TSS transcripts (PROMPTs) are thought to utilize cryptic PAS close to the TSS while sense TSS transcripts may have reduced occurrence of cryptic PAS. Those that are present are thought to be blocked by nearby 5'SS U1snRNP recruitment (Kaida et al., 2010). These apparent differences in cryptic PAS usage between PROMPTs and sense TSS associated transcripts have been proposed to favor productive sense over non-productive antisense transcription. In contrast our ANET-seq data argue that CPA factors and Xrn2 play equivalent roles in restricting sense and antisense TSS transcription. Thus their depletion by siRNA treatment causes an equivalent increase in Pol II pausing in both transcriptional directions. We also show that CPA factors are directly and equally associated with these two transcript classes by CLIP analysis (Martin et al., 2012). Our data suggest that transcriptional directionality at TSS is unlikely to be regulated by CPA mediated termination. Rather both sense and antisense TSS associated transcripts are restricted by normally TES associated termination factors. Indeed we observe a redistribution of S2P Pol II from the TES to the TSS following CPA factor and Xrn2 knockdown. This argues for close interconnections between both ends of the Pol II transcription unit, as previously demonstrated by 3C analysis (Ansari and Hampsey, 2005; O'Sullivan et al., 2004; Tan-Wong et al., 2012).

Overall, the ANET-seq method shows Pol II pausing and RNA cleavage resulting in 3'OH at RNA 3' ends at single-nucleotide resolution. Critically the ANET-seq method can be applied to genome-wide analyses to check the occupancy of modified polymerase (even Pol I and Pol III) by selecting a range of different antibodies to pull down the associated nascent RNA. Furthermore ANET-seq can be applied to search for novel non-coding RNA that are rapidly degraded. We anticipate that ANET-seq will expand our knowledge of how different nascent RNA are associated with specific "CTD codes". This will illuminate the complexities of co-transcriptional RNA processing and regulated gene expression.

6. Experimental Procedures

Antibodies

Pol II antibodies CMA301, CMA302 and CMA303 were generated by Dr. H. Kimura (Stasevich et al., 2014). 8WG16 and Aly antibodies were purchased from Abcam. CPSF73, CstF-64 and CstF-64 tau antibodies were purchased from Bethyl laboratories. α-Tubulin antibody was purchased from Sigma. Xrn2 antibody was provided by Dr. N. Gromak.

Cell culture, NRO assay and RT-PCR

Cell culture and NRO assay were as previously described (Nojima et al., 2013). siRNA transfection, RT-PCR and primers are described in Extended Experimental Procedures.

ANET-seq, ChrRNA-seq and bioinfomatical analysis

ANET-seq and ChrRNA-seq were conducted according to Figure 1A and Supplemental Figure S2. For further details, data processing and bioinfomatical analysis see Supplemental Experimental Procedures.

ACCESSION NUMBERS

The accession number for sequence data will be submitted shortly in NCBI's gene Expression Omnibus.

SUPPLEMENTAL INFORMATION

Supplemental information includes extended Experimental Procedures, 7 figures, 1 table and can be found with the Article online.

ACKNOWLEDGMENTS

We thank the NJP lab and Dr. M. Dienstbier for critical discussion. T.N. was supported by the KANAE foundation. This work was supported by funding to NJP (Wellcome Trust Programme and ERC Advanced Grants) and to MCF (Fundação Ciência e Tecnologia, Portugal).

7. References

Ahn, S.H., Kim, M., and Buratowski, S. (2004). Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3' end processing. Mol Cell *13*, 67-76.

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.

Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B., and Sharp, P.A. (2013). Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature *499*, 360-363.

Ansari, A., and Hampsey, M. (2005). A role for the CPF 3'-end processing machinery in RNAP II-dependent gene looping. Genes Dev *19*, 2969-2978.

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

Brannan, K., Kim, H., Erickson, B., Glover-Cutter, K., Kim, S., Fong, N., Kiemele, L., Hansen, K., Davis, R., Lykke-Andersen, J., *et al.* (2012). mRNA decapping factors and the exonuclease Xrn2 function in widespread premature termination of RNA polymerase II transcription. Mol Cell *46*, 311-324.

Brookes, E., and Pombo, A. (2009). Modifications of RNA polymerase II are pivotal in regulating gene expression states. EMBO Rep *10*, 1213-1219.

Burnette, J.M., Miyamoto-Sato, E., Schaub, M.A., Conklin, J., and Lopez, A.J. (2005). Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. Genetics *170*, 661-674.

Carrillo Oesterreich, F., Preibisch, S., and Neugebauer, K.M. (2010). Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. Mol Cell *40*, 571-581.

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

Churchman, L.S., and Weissman, J.S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature *469*, 368-373.

Consortium, E.P., Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845-1848.

David, C.J., Chen, M., Assanah, M., Canoll, P., and Manley, J.L. (2010). HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. Nature *463*, 364-368.

David, C.J., and Manley, J.L. (2011). The RNA polymerase C-terminal domain: a new role in spliceosome assembly. Transcription *2*, 221-225.

Davidson, L., Kerr, A., and West, S. (2012). Co-transcriptional degradation of aberrant pre-mRNA by Xrn2. EMBO J *31*, 2566-2578.

Davidson, L., Muniz, L., and West, S. (2014). 3' end formation of pre-mRNA and phosphorylation of Ser2 on the RNA polymerase II CTD are reciprocally coupled in human cells. Genes Dev *28*, 342-356.

Dye, M.J., Gromak, N., and Proudfoot, N.J. (2006). Exon tethering in transcription by RNA polymerase II. Mol Cell *21*, 849-859.
Dye, M.J., and Proudfoot, N.J. (2001). Multiple transcript cleavage precedes polymerase release in termination by RNA polymerase II. Cell *105*, 669-681.
Egloff, S., Dienstbier, M., and Murphy, S. (2012a). Updating the RNA polymerase CTD code: adding gene-specific layers. Trends Genet *28*, 333-341.
Egloff, S., O'Reilly, D., Chapman, R.D., Taylor, A., Tanzhaus, K., Pitts, L., Eick, D., and Murphy, S. (2007). Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression. Science *318*, 1777-1779.
Egloff, S., Zaborowska, J., Laitem, C., Kiss, T., and Murphy, S. (2012b). Ser7 phosphorylation of the CTD recruits the RPAP2 Ser5 phosphatase to snRNA genes. Mol Cell *45*, 111-122.
Glover-Cutter, K., Kim, S., Espinosa, J., and Bentley, D.L. (2008). RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. Nat Struct Mol Biol *15*, 71-78.
Gromak, N., West, S., and Proudfoot, N.J. (2006). Pause sites promote transcriptional termination of mammalian RNA polymerase II. Mol Cell Biol *26*, 3986-3996.
Grosso, A.R., de Almeida, S.F., Braga, J., and Carmo-Fonseca, M. (2012). Dynamic transitions in RNA polymerase II density profiles during transcription termination. Genome Res *22*, 1447-1456.
Grzechnik, P., Tan-Wong, S.M., and Proudfoot, N.J. (2014). Terminate and make a loop: regulation of transcriptional directionality. Trends Biochem Sci *39*, 319-327.
Hah, N., Danko, C.G., Core, L., Waterfall, J.J., Siepel, A., Lis, J.T., and Kraus, W.L. (2011). A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. Cell *145*, 622-634.
Hatton, A.R., Subramaniam, V., and Lopez, A.J. (1998). Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions. Mol Cell *2*, 787-796.
Heidemann, M., Hintermair, C., Voss, K., and Eick, D. (2013). Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription. Biochim Biophys Acta *1829*, 55-62.
Hirose, Y., and Manley, J.L. (1998). RNA polymerase II is an essential mRNA polyadenylation factor. Nature *395*, 93-96.
Hirose, Y., Tacke, R., and Manley, J.L. (1999). Phosphorylated RNA polymerase II stimulates pre-mRNA splicing. Genes Dev *13*, 1234-1239.
Hsin, J.P., Sheth, A., and Manley, J.L. (2011). RNAP II CTD phosphorylated on threonine-4 is required for histone mRNA 3' end processing. Science *334*, 683-686.
Ip, J.Y., Schmidt, D., Pan, Q., Ramani, A.K., Fraser, A.G., Odom, D.T., and Blencowe, B.J. (2011). Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. Genome Res *21*, 390-401.
Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L., and Dreyfuss, G. (2010). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature *468*, 664-668.

Kim, H., Erickson, B., Luo, W., Seward, D., Graber, J.H., Pollock, D.D., Megee, P.C., and Bentley, D.L. (2010). Gene-specific RNA polymerase II phosphorylation and the CTD code. Nat Struct Mol Biol *17*, 1279-1286.
Kolev, N.G., and Steitz, J.A. (2005). Symplekin and multiple other polyadenylation factors participate in 3'-end maturation of histone mRNAs. Genes Dev *19*, 2583-2592.
Kornblihtt, A.R., de la Mata, M., Fededa, J.P., Munoz, M.J., and Nogues, G. (2004). Multiple links between transcription and splicing. RNA *10*, 1489-1498.
Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science *339*, 950-953.
Lacoste, N., Woolfe, A., Tachiwana, H., Garea, A.V., Barth, T., Cantaloube, S., Kurumizaka, H., Imhof, A., and Almouzni, G. (2014). Mislocalization of the centromeric histone variant CenH3/CENP-A in human cells depends on the chaperone DAXX. Molecular cell *53*, 631-644.
Martin, G., Gruber, A.R., Keller, W., and Zavolan, M. (2012). Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. Cell Rep *1*, 753-763.
Mavrich, T.N., Jiang, C., Ioshikhes, I.P., Li, X., Venters, B.J., Zanton, S.J., Tomsho, L.P., Qi, J., Glaser, R.L., Schuster, S.C., *et al.* (2008). Nucleosome organization in the Drosophila genome. Nature *453*, 358-362.
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.
Min, I.M., Waterfall, J.J., Core, L.J., Munroe, R.J., Schimenti, J., and Lis, J.T. (2011). Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. Genes Dev *25*, 742-754.
Moore, M.J., and Proudfoot, N.J. (2009). Pre-mRNA processing reaches back to transcription and ahead to translation. Cell *136*, 688-700.
Morlando, M., Ballarino, M., Gromak, N., Pagano, F., Bozzoni, I., and Proudfoot, N.J. (2008). Primary microRNA transcripts are processed co-transcriptionally. Nat Struct Mol Biol *15*, 902-909.
Munoz, M.J., Perez Santangelo, M.S., Paronetto, M.P., de la Mata, M., Pelisch, F., Boireau, S., Glover-Cutter, K., Ben-Dov, C., Blaustein, M., Lozano, J.J., *et al.* (2009). DNA damage regulates alternative splicing through inhibition of RNA polymerase II elongation. Cell *137*, 708-720.
Nojima, T., Dienstbier, M., Murphy, S., Proudfoot, N.J., and Dye, M.J. (2013). Definition of RNA polymerase II CoTC terminator elements in the human genome. Cell Rep *3*, 1080-1092.
Ntini, E., Jarvelin, A.I., Bornholdt, J., Chen, Y., Boyd, M., Jorgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P.R., *et al.* (2013). Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nat Struct Mol Biol *20*, 923-928.
O'Sullivan, J.M., Tan-Wong, S.M., Morillon, A., Lee, B., Coles, J., Mellor, J., and Proudfoot, N.J. (2004). Gene loops juxtapose promoters and terminators in yeast. Nat Genet *36*, 1014-1018.
Pawlicki, J.M., and Steitz, J.A. (2008). Primary microRNA transcript retention at sites of transcription leads to enhanced microRNA production. J Cell Biol *182*, 61-76.

Perez-Lluch, S., Blanco, E., Carbonell, A., Raha, D., Snyder, M., Serras, F., and Corominas, M. (2011). Genome-wide chromatin occupancy analysis reveals a role for ASH2 in transcriptional pausing. Nucleic Acids Res *39*, 4628-4639.
Proudfoot, N.J. (2011). Ending the message: poly(A) signals then and now. Genes Dev *25*, 1770-1782.
Rahl, P.B., Lin, C.Y., Seila, A.C., Flynn, R.A., McCuine, S., Burge, C.B., Sharp, P.A., and Young, R.A. (2010). c-Myc regulates transcriptional pause release. Cell *141*, 432-445.
Rodriguez, A., Griffiths-Jones, S., Ashurst, J.L., and Bradley, A. (2004). Identification of mammalian microRNA host genes and transcription units. Genome Res *14*, 1902-1910.
Saunders, A., Core, L.J., and Lis, J.T. (2006). Breaking barriers to transcription elongation. Nat Rev Mol Cell Biol *7*, 557-567.
Schumperli, D. (1988). Multilevel regulation of replication-dependent histone genes. Trends Genet *4*, 187-191.
Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature *479*, 74-79.
Shukla, S., and Oberdoerffer, S. (2012). Co-transcriptional regulation of alternative pre-mRNA splicing. Biochim Biophys Acta *1819*, 673-683.
Skourti-Stathaki, K., Proudfoot, N.J., and Gromak, N. (2011). Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. Mol Cell *42*, 794-805.
Stasevich T., Hayashi-Takanaka Y., Sato Y., Maehara K., Ohkawa Y., Sakata-Sogawa K., Tokunaga M., Nagase T., Nozaki N., McNally J. G., and Kimura H. (2014) Regulation of RNA polymerase II activationby histone acetylation in single living cells, Nature, in press
Tan-Wong, S.M., Zaugg, J.B., Camblong, J., Xu, Z., Zhang, D.W., Mischo, H.E., Ansari, A.Z., Luscombe, N.M., Steinmetz, L.M., and Proudfoot, N.J. (2012). Gene loops enhance transcriptional directionality. Science *338*, 671-675.
West, S., Gromak, N., and Proudfoot, N.J. (2004). Human 5' --> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. Nature *432*, 522-525.
West, S., Proudfoot, N.J., and Dye, M.J. (2008). Molecular dissection of mammalian RNA polymerase II transcriptional termination. Mol Cell *29*, 600-610.
Yao, C., Biesinger, J., Wan, J., Weng, L., Xing, Y., Xie, X., and Shi, Y. (2012). Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation. Proc Natl Acad Sci U S A *109*, 18773-18778.

8. Supplementary Material

8.1 Extended experimental procedures

siRNA transfection

SMARTpool siRNA against human CPSF73 (CPSF3) and CstF64 (CSTF2) were purchased from Thermo scientific. ON-TARGET plus siRNA against Xrn2 was made by Thermo Scientific as following sequences. Sense: AAGAGUACAGAUGAUCAUGUU, Antisense: 5'-P CAUGAUCAUCUGUACUCUUUU. Silencer select siRNA against CstF64 tau (CSTF2T) was designed by Life technologies as following sequence, Sense: CCAUUAUUGACUCACCCUAtt, Antisense: UAGGGUGAGUCAAUAAUGGgc. These siRNA (final conc. 30nM) were transfected into HeLa cell using Lipofectamine RNAiMAX reagent (Life technologies) according to the manual and incubated for 72 hours.

RT-PCR analysis

RNA was isolated from HeLa cells and cells were transfected with Trizol. For reverse transcription, 500 ng of total RNA was incubated with oligo $(dT)_{20}$ and Superscript II reverse transcriptase (Life Technologies). PCR was performed using GO taq polymerase (Promega) and following primer set.

PKMex8_Fw: 5'- GATGGAGCCGACTGCATCATG -3',

PKMex11_Rv: 5'- ATTCCGGGTCACAGCAATGAT -3'

PCR products were digested by either NcoI (NEB) or PstI (NEB) for six hours. The PCR products were analyzed by 2% agarose gel electrophoresis, followed by ethidium bromide staining.

Chromatin-bound RNA (ChrRNA)-seq method and RNA library preparation

Chromatin RNA fraction was prepared from ~80% confluent HeLa cells in 100mm Dishes. Approximately $7x10^6$ cells were washed with ice-cold PBS twice. The cells were lysed with ice-cold 4 ml of HLB/NP40 buffer (10 mM Tris-Hcl pH 7.5, 10 mM NaCl, 0.5% NP40 and 2.5 mM $MgCl_2$) and incubated on ice for 5 min. After the incubation, 1 ml of ice-cold HLB/NP40/Sucrose buffer (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 0.5% NP40, 2.5 mM $MgCl_2$ and 10 % Sucrose) was under-laid and then the nuclei were collected under 1,400 rpm centrifuge at 4^0C for 5 min. Isolated nuclei were resuspended in 1.25 μl of NUN1 solution (20 mM Tri-HCl pH 8.0, 75mM NaCl, 0.5 mM EDTA, 50% Glycerol and proteinase inhibitor

1xComplete (Roche)) and added 1.2 ml NUN2 buffer (20 mM HEPES-KOH pH 7.6, 7.5 mM $MgCl_2$, 0.2 mM EDTA, 300 mM NaCl, 1 M Urea, 1% NP40, proteinase inhibitor 1xComplete and phosphatase inhibitor 1xPhosStop (Roche)). 15 min incubation was carried out on ice with mixing by max speed vortex for 5 sec every ~4 min and then chromatin pellets were precipitated under 13,000 rpm centrifuge at 4^{o}C for 10min. Chromatin pellet was resuspended in 200 μl HSB (10mM Tris-HCl pH 7.5, 500 mM NaCl and 10 mM $MgCl_2$) with 0.25 U/μl TURBO DNase (Life technologies) at 37^{o}C for 10 min and then treated with Proteinase K for 10 min. RNA was extracted by Trizol reagent (Life technologies). This extraction steps were repeated three times.

In prior to RNA library preparations, rRNAs were depleted using Ribo-Zero rRNA removal kits (Epicentre) from 5 μg of Chromatin RNA. RNA was also fragmented 150-200 nt by heat treatment (94 ^{o}C) for 15 min in 1xNEB first strand synthesis buffer. 100 ng or chromatin RNA was used for RNA library preparations. These were carried out according to NEBNext Ultra Directional RNA Library Prep kit for Illumina (NEB). Deep sequencing using Hiseq2000 and Hiseq2500 were performed by the Wellcome Trust Centre for Human Genetics (WTCHG) Oxford UK.

ANET-seq method and RNA library preparation

Approximately 1.6x10^{8} cells were used to generate nuclear and chromatin fractions. Isolated chromatin was washed in 1 ml of 1x Micrococcal Nuclease (MNase) buffer (NEB) and then incubated with MNase (40 u/μL) on Thermomixer (eppendorf, 1,400 rpm) at 37^{o}C for 90 sec. In order to inactivate MNase, EGTA (25 mM) was added immediately after the reaction and soluble digested chromatin was collected by 13,000 rpm centrifuge for 5 min. The supernatant was diluted with 9 ml of NET-2 buffer and add Pol II antibody-conjugated beads. 40 μg of Pol II antibody was used for each ANET-seq experiment. Immunoprecipitation was performed at 4^{o}C for one hour. The beads were washed with 1 ml of NET-2 buffer six times and with 500 μl of 1xPNKT (1xPNK buffer and 0.05 % Triton X-100) buffer once in the cold room. The washed beads were incubated in 100 μl of PNK reaction mix (1xPNKT, 1 mM ATP and 0.05 U/ml T4 PNK 3'phosphatase minus (NEB)) on Thermomixer (1,400 rpm) at 37^{o}C for 6 min. After the reaction the beads were washed with 1 ml of NET-2 buffer once and RNA was extracted with Trizol reagent.

RNA was resolved on 8 % denaturing acrylamide 7 M urea gels for size purification. 35-100 nt fragments were eluted from the gel using RNA elution buffer (1 M NaOAc and 1 mM EDTA) and RNA was precipitated in 75 % Ethanol. RNA libraries were prepared according to the manual of Truseq small RNA library prep kit (Illumina). Deep sequencing was conducted by WTCHG in Oxford.

Analyses of in vivo Cross-linking and Immuno-precipitation (CLIP) assay for TSS

CLIP-sequencing datasets (Martin et al., 2012) were downloaded for the following transcription factors, CPSF-73, CstF-64, CstF-64tau, CPSF-160, CPSF-30 and CF-Im25. Normalized read counts were calculated for sense and antisense strands relative to the direction of gene transcription for a region of 3 kb upstream and downstream of annotated Refseq TSS and plotted for 10 bp bin (Supplementary Table S1).

Data pre-processing

ANET-seq data adaptors were trimmed using Cutadapt (v1.1) (Martin, 2011), discarding reads with less than 10 bases. Then a Perl script was used to remove the reads left unpaired. The remaining reads were then aligned to the reference human genome (hg19) using TopHat (v2.0.9) (Kim et al., 2013) with a minimum anchor length of 5 bases, and only allowing for one alignment to the reference. It was necessary to determine the last nucleotide incorporated by the polymerase and its directionality. This nucleotide was defined as the 5' end of read two of the pair, with the directionality indicated by read one. Knowing this, the properly aligned pairs of reads were trimmed to solely keep the 5' nucleotide of read two. This was done using SAMtools (Li et al., 2009) and a python script. SAMtools was also used to separate the reads by strand for further analysis.

ChrRNA-seq data was aligned using the same version of TopHat, but allowing for the read pairs to be separated by 3kb. For the metagene representation, SAMtools was used to separate the reads by strands.

ChIP-seq data for unphosphorylated Pol II, H3K4m3 and H3K36m3 (GEO accession numbers GSM935395, GSM945201 and GSM733711, respectively) were generated as part of the ENCODE Project (Consortium et al., 2012).

Determination of expressed genes

To determine the genes expressed in HeLa S3 cells, strand-specific RNA-seq data from a previously published study (Lacoste et al., 2014) was used (GEO accession number

GSM1155630). The data was aligned with TopHat and then Cufflinks (v2.1.1) (Trapnell et al., 2010) was used to acquire a FPKM value for each gene. These values were then converted to log2 and their distribution was plotted. The cut off value chosen to determine the expressed genes was the local minimum of the log2 (FPKM) distribution between the primary peak of high expression genes and the long left shoulder of low-expression transcripts as previously reported (Hart et al., 2013). This defined 11560 expressed genes, of which 10473 were protein coding. From these genes a further selection of ones where the gene body and the adjacent regions (TSS-1000bp and TES+3kbp) do not intersect other genes was made. This resulted in 1647 genes used to generate the metagene profiles.

Metagene profiles

The metagene profiles represent average profiles across expressed genes for Pol II or RNA abundance. To generate these, genes were aligned by their annotated TSS and TES. The 5' end, showing a span of 1kb up and downstream of the TSS, and the 3' end, showing the interval from TES-500bp to TES+3kb, were unscaled and averaged in a 5bp window. The remaining gene body was scaled to 100 equally sized bins, so that all the genes appear the same length.

Metagene profiles were generated using this same method, but the window around the TSS extended from TSS-250 bp to TSS+250 bp, and around the TES from TES-250 bp to TES+1 kb.

The individual profiles were plotted in single base windows and using a scale of reads per 10^8 sequences.

Determination of included and excluded exons

To determine if alternative exons were included or excluded in the transcripts produced, previously described RNA-seq data used for determining expressed genes was analysed with MISO (Katz et al., 2010). These results were compared to RefSeq exon reference data. Exons were then divided according to the Ψ-value calculated by MISO, which indicates the fraction of inclusion of an exon predicted for a dataset. Only exons with more than 0.9 or less than 0.1 were considered included or excluded, respectively.

Escaping and Read-Through Index

Escaping Index (EI) is defined as the proportion of Pol II from the TSS that proceeds to the elongation phase of transcription. It was calculated as follows:

$$EI = \log_2\left(\frac{GB}{TSS} + c\right),\ c = \frac{\min\left(\frac{GB}{TSS}>0\right)}{2}$$

where GB is the Reads per Kilobase per Milion reads (RPKM) of sense reads in the interval [TSS+500, TES], TSS is the RPKM of sense reads in [TSS-50, TSS+250]. The constant c was used to log the zeros in the data. The first 500 bases of each gene are excluded from the definition of the gene body to prevent TSS polymerase accumulation from interfering with the counts for the gene body.

The Read-Through Index was calculated using the same approach, but instead of considering the TSS interval, the RPKM of sense reads for [TES, TES+2000] was used.

Normalized Gene Body counts use the same formula but without dividing the RPKM from the gene body region by any of the others.

Significance of the differences between control and knock-down for each index was calculated using a two-sided Mann-Whitney test. The p-values were then adjusted using the Holm method.

PCR Primers sequences for ChIP assay

GAPDH_TSS_F	5'-cggctactagcggttttacg-3'
GAPDH_TSS_R	5'-gctgcgggctcaatttatag-3'
GAPDH_int1_F	5'-CCCCTTCATACCCTCACGTA-3'
GAPDH_int1_R	5'-GACAAGCTTCCCGTTCTCAG-3'
GAPDH_I6E7_F	5'-acccagaagactgtggatgg-3'
GAPDH_I6E7_R	5'-ttcagctcagggatgacctt-3'
GAPDH_PAS_F	5'-CTGAATCTCCCCTCCTCACA-3'
GAPDH_PAS_R	5'-TGCCCCAGACCCTAGAATAA-3'
GAPDH_PAS+1.1k_F	5'-TCCAGCCTAGGCAACAGAGT-3'
GAPDH_PAS+1.1k_R	5'-TGTGCACTTTGGTGTCACTG-3'
IST1_-2k_F	5'-TGTTAGCCAGGGTGGTCTTC-3'
IST1_-2k_R	5'-GGTCAGGAGTTGGAGAGCAG-3'
IST1_TSS_F	5'-aaccctgaagtcggtgtctg-3'
IST1_TSS_R	5'-ctccgaagtcgtttgaatcc-3'
IST1_B_F	5'-caccatgcccagctaatttt-3'
IST1_B_R	5'-accctcaggtggttctgatg-3'
IST1_LE_F	5'-tgaaggcctcgcttagttgt-3'
IST1_LE_R	5'-gcaccttgtcctttctctgc-3'
IST1_+4k_F	5'-TCCGCTGTCACTGCATAAAC-3'
IST1_+4k_R	5'-TTCCCATGGAGAGGAACATC-3'
MYC_TSS_F	5'-gggatcgcgctgagtataaa-3'
MYC_TSS_R	5'-cctattcgctccggatctc-3'
MYC_I2_F	5'-tggcagggagtgtatgaatg-3'
MYC_I2_R	5'-cacccactcttgaggcagtt-3'
MYC_+0.8K_F	5'-ACATCAACCCCATGAAGGAG-3'
MYC_+0.8K_R	5'-GTGGCTTGGACAGGTTAGGA-3'
MYC_+2.5k_F	5'-GATGGAGACCATCCTGGCTA-3'
MYC_+2.5k_R	5'-ATGCAGTGGCACAATCTCAG-3'

Supplementary Table - p-values of every two-sided Mann-Whitney test for every index, before and after adjusting using the Holm method

sample (vs. siLuc)	EI		RTI		GB	
	before adjusting	after adjusting	before adjusting	after adjusting	before adjusting	after adjusting
siCPSF73	3.53E-63	3.53E-63	9.72E-16	1.94E-15	0.0087335989	0.0087335989
siCstF64si64t	1.36E-138	4.08E-138	3.34E-20	1.00E-19	0.0008596245	0.0017192490
siXrn2	6.23E-72	1.25E-71	0.9894088	0.9894088	0.0000645689	0.0001937067

8.2 Supplementary Figures

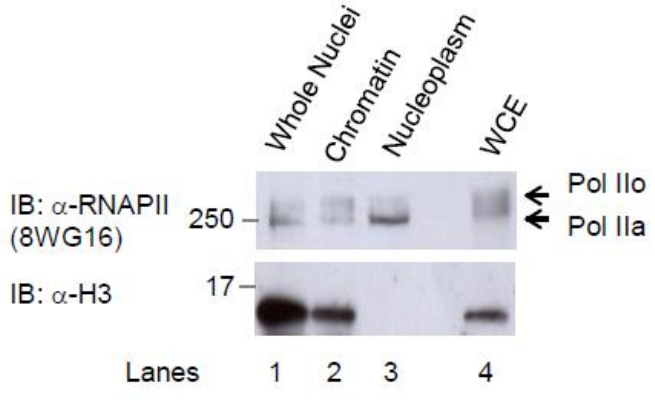


Figure S1. **Pol II phosphorylation in different fractions**

HeLa cell extracts were prepared from whole cell (WCE), whole nuclei, chromatin and nucleoplasm fractions. Two major phosphorylated forms of Pol II, hypophosphorylated (Pol IIa) and hyperphosphorylated Pol II (Pol IIo) were detected by western blot using 8WG16 Pol II antibody. H3 was detected as a chromatin marker.

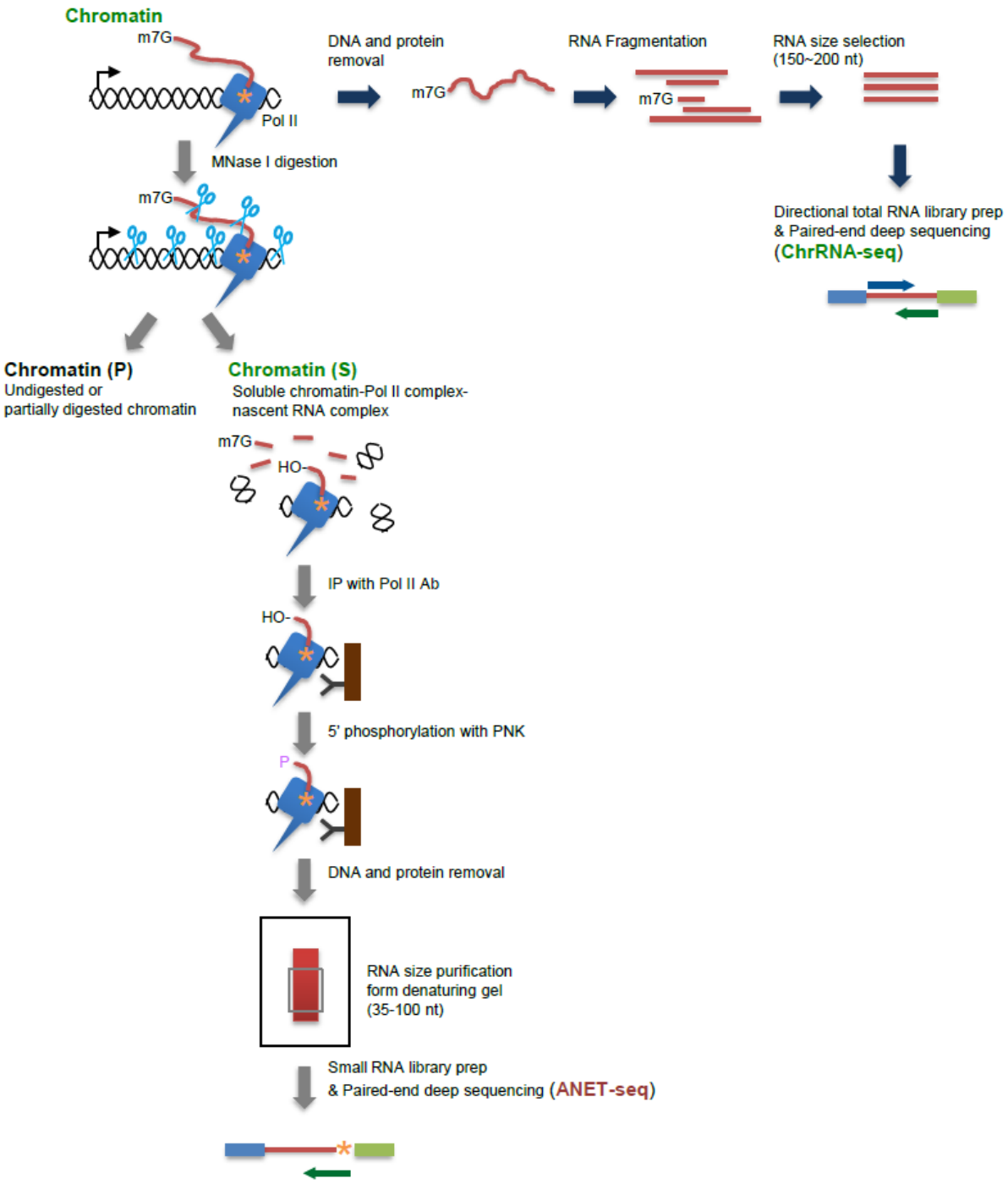


Figure S2. Detailed ChrRNA-seq and ANET-seq methods, Related to Figure 1

(Right) ChrRNA-seq method. Chromatin-bound RNA (red line) is purified from isolated chromatin fraction by micrococcal nuclease (MNase1) and proteinase K treatments. Pol II and RNA synthesizing site are shown as tailed blue box and orange asterisk, respectively. RNA is fragmented to 150-200 nt by heat and adapters ligated on both ends for paired-end 51bases directional deep sequencing (blue and green arrows).

(Below) ANET-seq method. Chromatin DNA and chromatin-bound RNA are digested with MNase I (light blue scissors). To separate insoluble pellet (P) and soluble chromatin supernatant (S), digested chromatin is centrifuged. Soluble Pol II-nascent RNA complex is immuno-precipitated (IPed) with Pol II antibody. 5' hydroxyl (OH) is then phosphorylated with PNK on beads and phenol extraction performed to remove DNA and proteins. IPed RNA is purified from denaturing gel (size range 35-100 nt). RNA adapters are added to both ends strand-specifically and deep sequencing is conducted from reverse sequence primer (green arrow) to read 3' end of insert (orange asterisk).

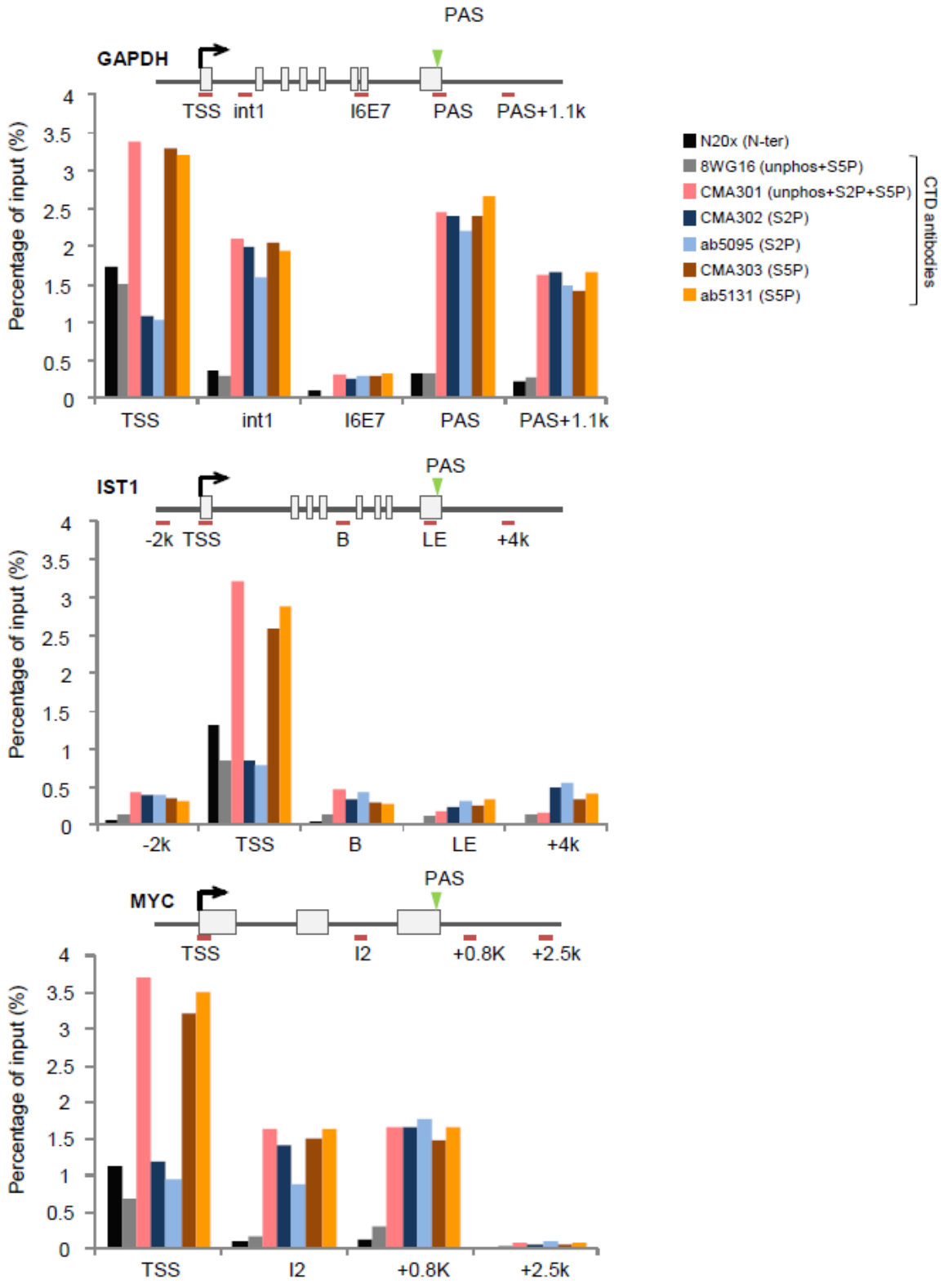


Figure S3. Gene specific ChIP analysis using indicated Pol II antibodies, Related to Figure 2

Pol II ChIP was conducted with indicated Pol II antibodies on GAPDH, IST1 and MYC genes. Positions of primer sets and PAS are shown by red bars and green triangles, respectively. TSS denoted by black arrow.

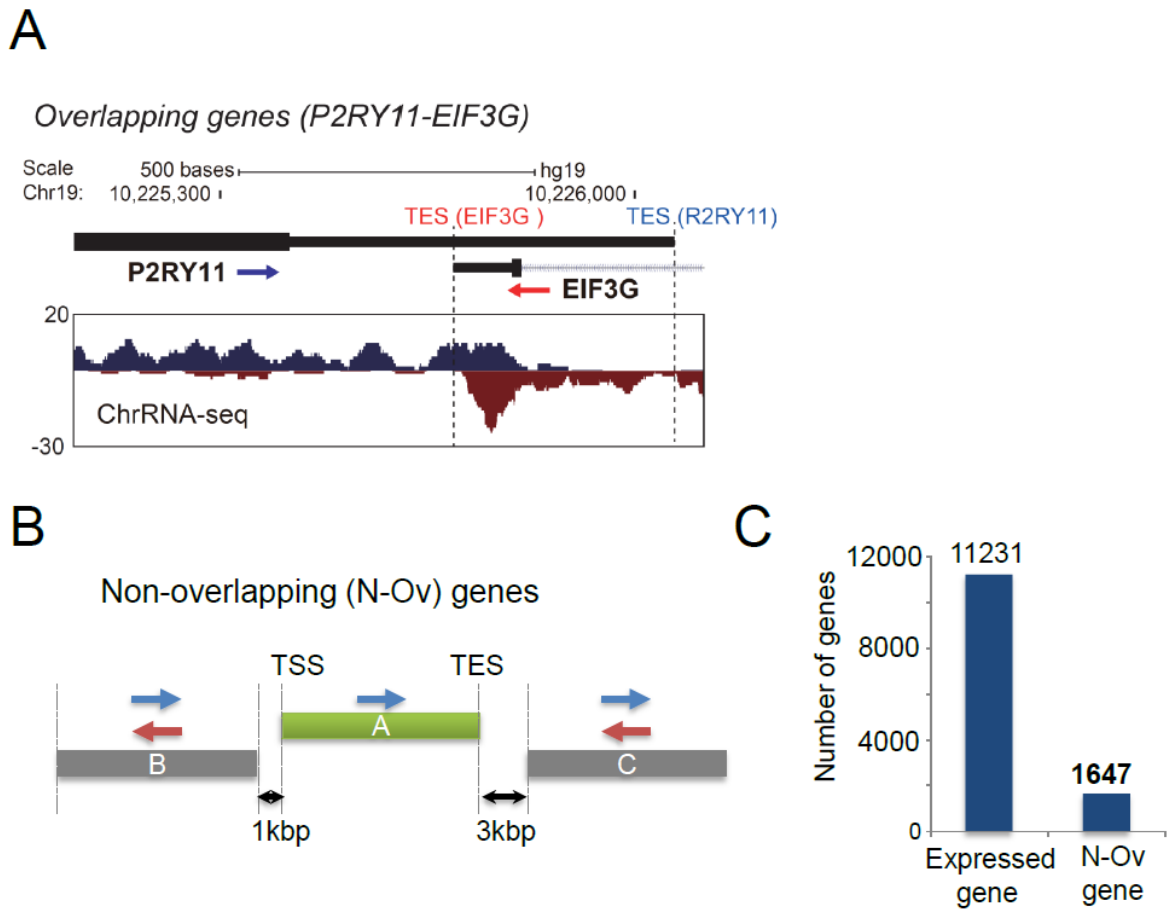


Figure S4. Overlapping gene units

(A) Example of overlapping genes. ChrRNA-seq signals from P2RY11 (dark blue) and EIF3G (dark red) genes overlap at their TES.

(B) Selection criteria of non-overlapping (N-Ov) genes. Genes of interest, A (green) are isolated from neighboring genes. Gene B is 1 kb upstream of gene A TSS. Gene C is 3 kb downstream of gene C TES. Blue and red arrows show transcription direction.

(C) 11231 genes are selected as expressed genes in HeLa cell. 1647 genes are not overlapping based on criteria set in (B).

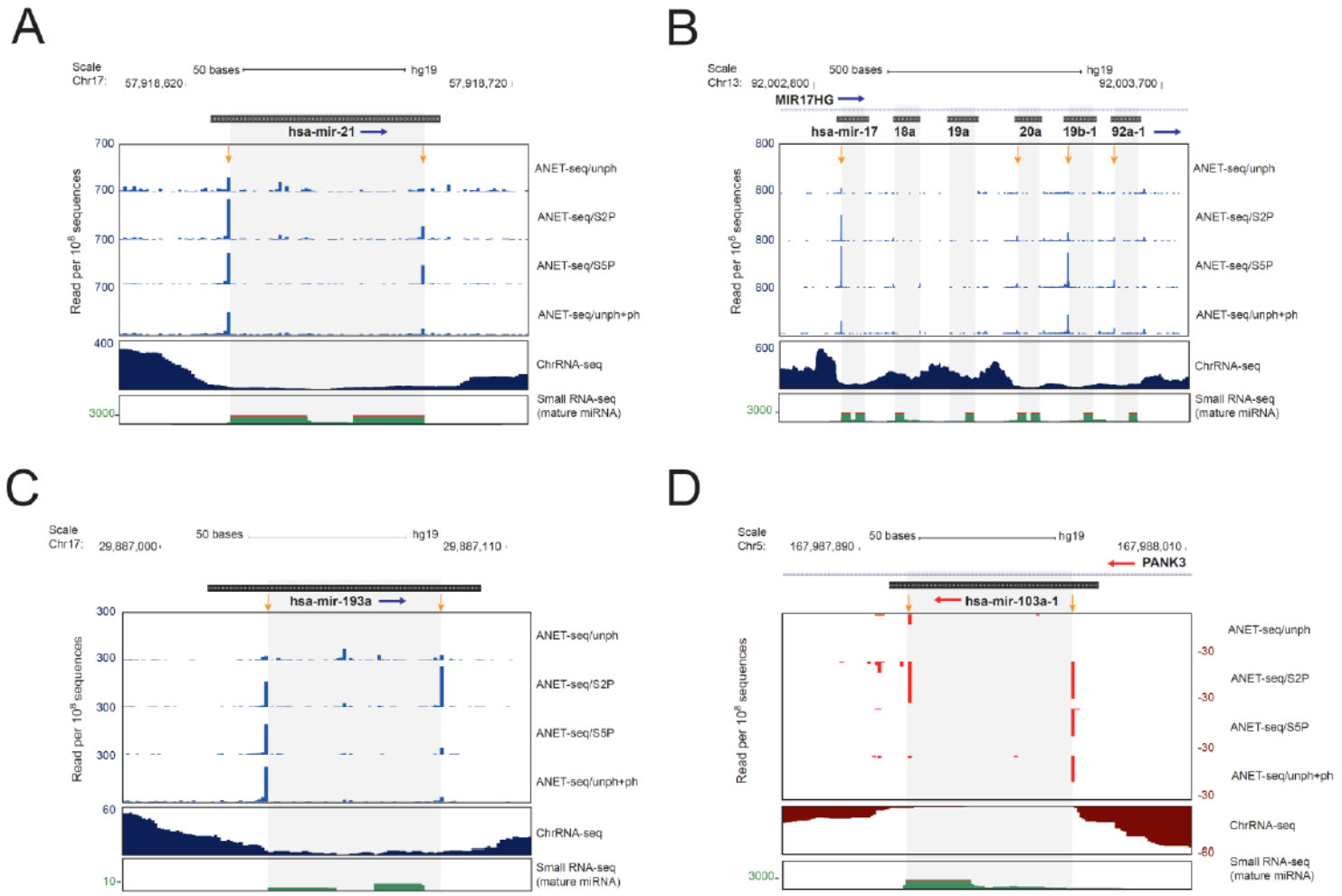


Figure S5. Further examples of pre-miRNA ANET-seq and ChrRNA-seq profiles, Related to Figure 3

ANET-seq analysis with unph, S2P, S5P and unph+ph antibodies compared with ChrRNA-seq and small RNA-seq profiles for hsa-mir-21 (A), MIR17HG (B), hsa-mir-193a (C) and has-mir-103a-1 (D). Details as for Figure 4 legend. Note MIR17HG harbors polycistronic pre-miRNAs.

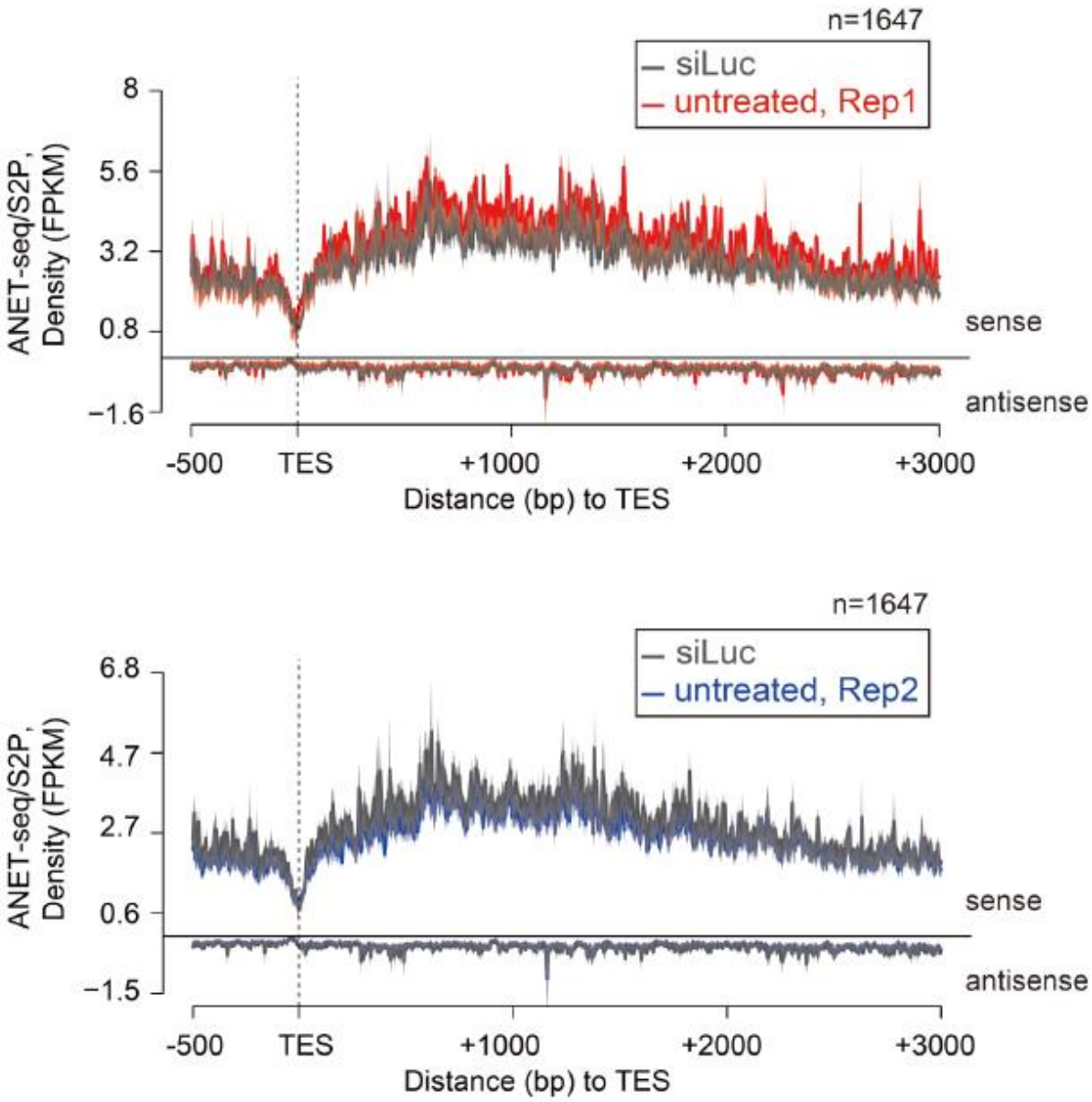


Figure S6. Comparison of siLuc control treated ANET-seq with untreated profiles, Related to Figure 6

siLuc treated HeLa cell ANET-seq/S2P metagene profile over TES compared with untreated cell replicates. See Figure 6 legend for further details.

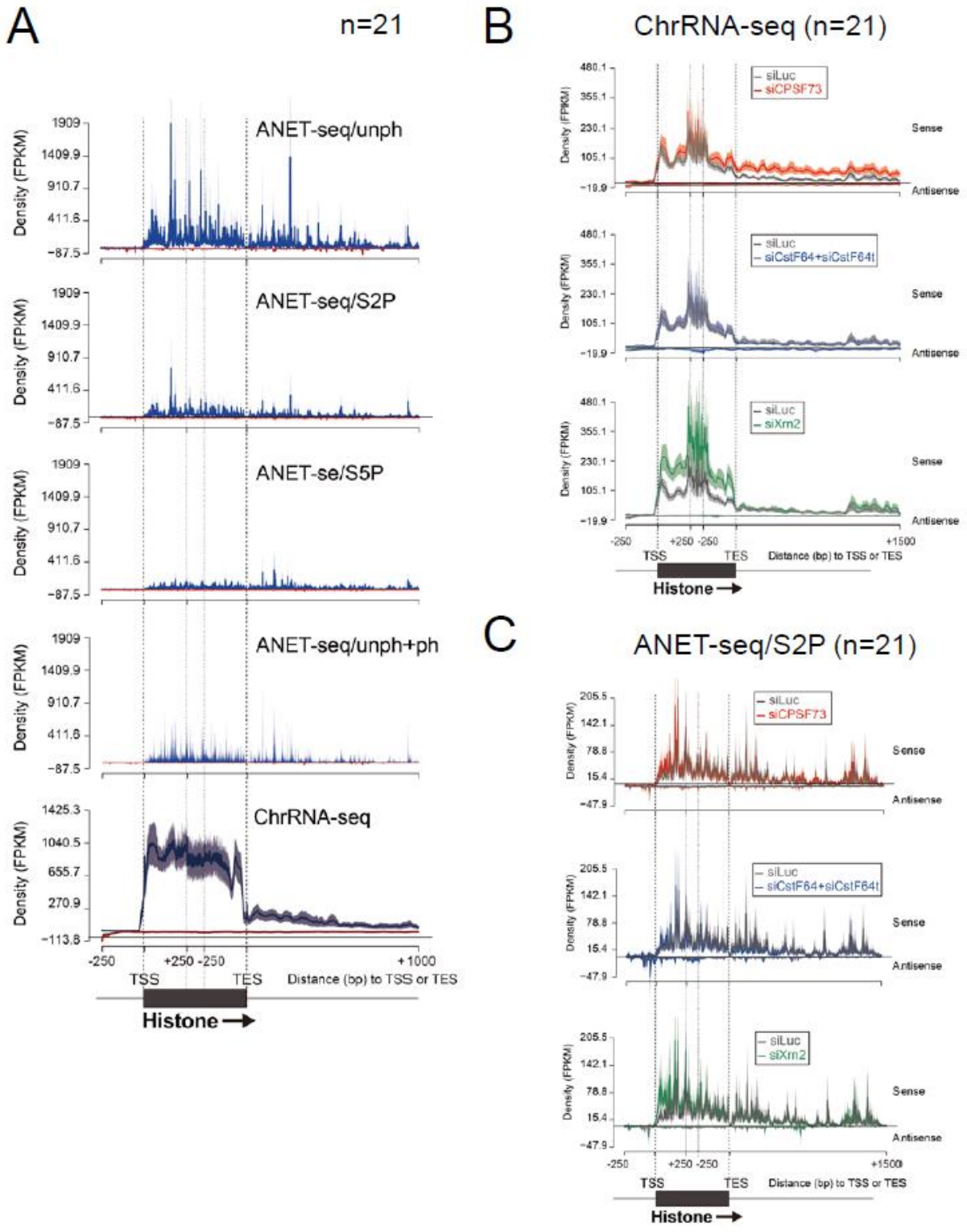


Figure S7. Histone gene ANET-seq and ChrRNA-seq profiles.

(A) Histone metagene analysis using ANET-seq with different Pol II CTD antibodies compared to ChrRNA-seq.

(B) ChrRNA-seq following termination factor knockdown by siRNA. See Figure 5 legend.

(C) ANET-seq/S2P following termination factor knockdown by siRNA. See Figure 5 legend.

8.3 Supplementary References

Consortium, E.P., Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

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.

Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature methods *7*, 1009-1015.

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 biology *14*, R36.

Lacoste, N., Woolfe, A., Tachiwana, H., Garea, A.V., Barth, T., Cantaloube, S., Kurumizaka, H., Imhof, A., and Almouzni, G. (2014). Mislocalization of the centromeric histone variant CenH3/CENP-A in human cells depends on the chaperone DAXX. Molecular cell *53*, 631-644.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Martin M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal

Martin, G., Gruber, A.R., Keller, W., and Zavolan, M. (2012). Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. Cell Rep *1*, 753-763.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.

Chapter 3

Conclusions and Perspectives

Evolution of research methodologies in life sciences has come a long way in the last fifty years. In the past, technical hindrances limited the understanding of biological systems, which invariably led to having a narrow scope when taking conclusions from data. This did not hold back the growth of knowledge, yet many questions were raised because of the high variability and noise in collected data, and few were answered at that time. Only with technology development and constant increase in throughput became clear that living systems are far more complex, and the web of interactions in a cell is far more intricate than first imagined, with some degree of functional redundancy that could justify the questions before raised. It is therefore crucial that high-throughput techniques are utilized in regulation studies, so as to accurately draw the big picture of biological functioning.

ANET-seq is an advantageous improvement in this type of approaches. While NET-seq (Churchman and Weissman, 2011) was useful for describing polymerase occupancy in detail, its advanced counterpart here described allows the distinction of CTD modifications in data interpretation. The Pol II C-terminal domain has been increasingly described as a key regulator in transcription and its associated events, and future experiments using this high-throughput, high-precision system focusing on its various modifications will certainly add valuable insights to all fields surrounding transcription. Additionally, extracting this information from different circumstances is also of great importance. Like in this study, where the use of siRNA opened new perspectives about CPA and termination factors function, so other conditions - such as different cell stages, types and organisms – can be compared to reveal novel aspects of transcription in concrete scenarios.

Besides revalidating some of the previous findings about CTD isoform distribution in genes, ANET-seq data showed its relationship with splicing and miRNA processing. Constant activity of processing machinery is increasingly obvious, but it still surprises the evident duality of co and post-transcriptional processes, and what exactly defines these processing timing differences. Further investigation is required in order to explain in real time the decision-making processes of the cell in regards to transcription. And while it seems clear that the signal captured at the end of exons belongs to co-transcriptionally spliced exons, it is not yet fully understood how these intermediates are captured and if they can accurately measure the degree of splicing that is occurring.

Some outstanding knowledge came out from siRNA experiments as well. ANET-seq preformed in CPA factors-depleted cells showed a clear decrease in effective transcription termination, but not when the removed factor was the termination-related protein Xrn2. However, perhaps even more surprising were the differences observed in TSS Pol II accumulation when any of the tested factors was removed. These differences pointed to novel roles of these factors in early termination, and hinted perhaps on one more motive for gene looping that is the sharing of transcription termination factors for early and late transcription.

This work also showed the great need of a constant cooperation between molecular and computational biologists. Increasing complexity in data generation creates a higher demand in robust and adaptable bioinformatical tools, so as to extract the significant elements from a dataset. It becomes clear, however, that an increase in the number of tools also makes it difficult to establish a standard analysis pipeline, thus increasing debate on whether the right methodology is being used or not. But this is a positive situation, as it is important to adequate the analysis to the data. This is why a good understanding of the workings of the referred tools, together with a robust pipeline, is so necessary in these methods, since it allows a clear justification of why a specific analysis was preformed. In this particular situation, the aspect which is most demanding of the data might be the single-nucleotide resolution of ANET-seq reads. Other methodologies have been created in order to make peak calling of ChIP-seq or related data, these tools are not suited for working with this kind of data. Hence, it can also be argued that this new generation of high-precision sequencing techniques creates a new niche for the development of new tools able to accommodate their properties.

Future research on transcription should make use of emerging high precision tools. Single-cell technologies are soon to be established, revealing the intercellular variability that is currently considered all together, and may consequently help filter some patterns that only exist as a result of a mixture of different cells' information. Integration of ever more clear-cut microscopy technologies might also be helpful in tracking individual molecules, which would help describe certain events in real time. Finally, a great challenge that is transversal to many biology fields is the ability to efficiently mine the continuously growing high-throughput data flow. New methods of extracting and presenting information are of great demand wherever these technologies apply. As for biological targets, transcription will also have to focus in subsets of genes and their transcription, highlighting the expanding topic of non-coding RNAs and how they are synthesized and their role in gene expression regulation. It is also of great importance to better understand the co-transcriptionally associated phenomena, namely splicing and small RNA processing, and how these may change by altering the cellular

environment. ANET-seq will certainly be a central point in future studies related to these fields, consequently making the capacity for analyzing this type of data a very valuable skill, just like it is for RNA-seq or ChIP-seq.

References

Acker, J., Graaff, M., Cheynel, I., Khazak, V., Kedinger, C., and Vigneron, M. (1997). Interactions between the Human RNA Polymerase II Subunits. Journal of Biological Chemistry*272*, 16815–16821.

Allfrey, V., Faulkner, R., and Mirsky, A. (1964). Acetylation and Methylation of Histones and their Possible Role in the Regulation of RNA Synthesis. Proceedings of the National Academy of Sciences *51*, 786–794.

Ambros, V., Bartel, B., Bartel, D., Burge, C., Carrington, J., Chen, X., Dreyfuss, G., Eddy, S., Griffiths-Jones, S., Marshall, M., et al. (2003). A uniform system for microRNA annotation. RNA *9*, 277–279

Barbosa-Morais, N., Irimia, M., Pan, Q., Xiong, H., Gueroussov, S., Lee, L., Slobodeniuc, V., Kutter, C., Watt, S., Çolak, R., et al. (2012). The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. Science *338*, 1587–1593.

Bartel, D. (2004). MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. Cell *116*.

Bartkowiak, B., Liu, P., Phatnani, H., Fuda, N., Cooper, J., Price, D., Adelman, K., Lis, J., and Greenleaf, A. (2010). CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. Genes & Development.

Baugh, L., Demodena, J., and Sternberg, P. (2009). RNA Pol II accumulates at promoters of growth genes during developmental arrest. Science (New York, N.Y.) *324*, 92–94.

Belotserkovskaya, R., Oh, S., Bondarenko, V., Orphanides, G., Studitsky, V., and Reinberg, D. (2003). FACT Facilitates Transcription-Dependent Nucleosome Alteration. Science*301*, 1090–1093.

Bernstein, E., Caudy, A.A., Hammond, S.M., and Hannon, G.J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature *409*, 363–366.

Beyer, A., and Osheim, Y. (1988). Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. Genes & Development *2*, 754–765.

Black, D.L. (2003). Mechanisms of alternative pre-messenger RNA splicing. Annual Review of Biochemistry *72*, 291–336.

Boehm, A.K., Saunders, A., Werner, J., and Lis, J.T. (2003). Transcription factor and polymerase recruitment, modification, and movement on dhsp70 in vivo in the minutes following heat shock. Molecular and Cellular Biology*23*, 7628–7637.

Boelens, M., Meerman, G., Gibcus, J., Blokzijl, T., Boezen, H., Timens, W., Postma, D., Groen, H., and Berg, A. (2007). Microarray amplification bias: loss of 30% differentially expressed genes due to long probe – poly(A)-tail distances. BMC Genomics *8*, 277.

Bonen, L. (1993). Trans-splicing of pre-mRNA in plants, animals, and protists. FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology *7*, 40–46.

Borchert, G., Lanier, W., and Davidson, B. (2006). RNA polymerase III transcribes human microRNAs. Nature Structural & Molecular Biology *13*, 1097–1101.

Brannan, K., Kim, H., Erickson, B., Glover-Cutter, K., Kim, S., Fong, N., Kiemele, L., Hansen, K., Davis, R., Lykke-Andersen, J., et al. (2012). mRNA decapping factors and the

exonuclease Xrn2 function in widespread premature termination of RNA polymerase II transcription. Molecular Cell *46*, 311–324.
Brengues, M., Teixeira, D., and Parker, R. (2005). Movement of Eukaryotic mRNAs Between Polysomes and Cytoplasmic Processing Bodies. Science *310*, 486–489.
Bushnell, D., and Kornberg, R. (2003). Complete, 12-subunit RNA polymerase II at 4.1-Å resolution: Implications for the initiation of transcription. Proceedings of the National Academy of Sciences *100*, 6969–6973
Chamberlin, M., and Berg, P. (1962). Deoxyribonucleic acid-directed synthesis of ribonucleic acid by an enzyme from escherichia coli. Proceedings of the National Academy of Sciences *48*, 81–94.
Chen, D. (1999). Regulation of Transcription by a Protein Methyltransferase. Science *284*, 21742177.
Churchman, L.S., and Weissman, J.S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature *469*, 368–373.
Churchman, L.S., and Weissman, J.S. (2012). Native elongating transcript sequencing (NET-seq). Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.] *Chapter 4*, Unit 4.14.1–17.
Cock, P., Fields, C., Goto, N., Heuer, M., and Rice, P. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research *38*, 1767–1771.
Connelly, S., and Manley, J. (1988). A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. Genes & Development *2*, 440–452.
Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science (New York, N.Y.)*322*, 1845–1848.
Cotten, M., Gick, O., Vasserot, A., Schaffner, G., and Birnstiel, M. (1988). Specific contacts between mammalian U7 snRNA and histone precursor RNA are indispensable for the in vitro 3' RNA processing reaction. The EMBO Journal*7*, 801–808
Cramer, P., Bushnell, D., and Kornberg, R. (2001). Structural Basis of Transcription: RNA Polymerase II at 2.8 Angstrom Resolution. Science *292*, 18631876.
Crick, F. (1970). Central dogma of molecular biology. Nature *227*, 561-563.
Dominski, Z., Yang, X., and Marzluff, W. (2005). The Polyadenylation Factor CPSF-73 Is Involved in Histone-Pre-mRNA Processing. Cell *123*.
Egloff, S., O'Reilly, D., Chapman, R., Taylor, A., Tanzhaus, K., Pitts, L., Eick, D., and Murphy, S. (2007). Serine-7 of the RNA Polymerase II CTD Is Specifically Required for snRNA Gene Expression. Science *318*, 1777–1779.
Ezkurdia, I., Juan, D., Rodriguez, J., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., and Tress, M. (2014). Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Human Molecular Genetics ddu309.
Fong, N., and Bentley, D. (2001). Capping, splicing, and 3′ processing are independently stimulated by RNA polymerase II: different functions for different segments of the CTD. Genes & Development *15*, 1783–1795.
Fuda, N.J., Ardehali, M.B., and Lis, J.T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature *461*, 186–192

Garber, M., Mayall, T., Suess, E., Meisenhelder, J., Thompson, N., and Jones, K. (2000). CDK9 Autophosphorylation Regulates High-Affinity Binding of the Human Immunodeficiency Virus Type 1 Tat-P-TEFb Complex to TAR RNA. Molecular and Cellular Biology *20*, 6958-6969.

Griffiths-Jones, S. (2004). The microRNA Registry. Nucleic Acids Research *32*, D109–D111.

Grosso, A.R., de Almeida, S.F., Braga, J., and Carmo-Fonseca, M. (2012). Dynamic transitions in RNA polymerase II density profiles during transcription termination. Genome Research *22*, 1447–1456.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77–88.

Guiner, C., Lejeune, F., Galiana, D., Kister, L., Breathnach, R., Stévenin, J., and Gatto-Konczak, F. (2001). TIA-1 and TIAR Activate Splicing of Alternative Exons with Weak 5′ Splice Sites followed by a U-rich Stretch on Their Own Pre-mRNAs. Journal of Biological Chemistry *276*, 40638–40646.

He, S., Su, H., Liu, C., Skogerbø, G., He, H., He, D., Zhu, X., Liu, T., Zhao, Y., and Chen, R. (2008). MicroRNA-encoding long non-coding RNAs. BMC Genomics *9*, 236.

Hebbes, T.R., Thorne, A.W., and Crane-Robinson, C. (1988). A direct link between core histone acetylation and transcriptionally active chromatin. The EMBO Journal *7*, 1395–1402.

Hsin, J., Sheth, A., and Manley, J. (2011). RNAP II CTD Phosphorylated on Threonine-4 Is Required for Histone mRNA 3' End Processing. Science 334, 683-686.

Hu, X., Ma, C., and Zhou, Y. (2013). A novel two-layer SVM model in miRNA Drosha processing site detection. BMC Systems Biology *7*, S4.

Hughson, F., and Schedl, P. (1999). Two domains and one RNA: a molecular threesome. Nature Structural Biology *6*, 499–502.

Hutvágner, G., and Zamore, P. (2002). A microRNA in a Multiple-Turnover RNAi Enzyme Complex. Science *297*, 2056-2060.

Isogai, Y., Keles, S., Prestel, M., Hochheimer, A., and Tjian, R. (2007). Transcription of histone gene cluster by differential core-promoter factors. Genes & Development *21*, 2936–2949

Jacob, S. (1995). Regulation of ribosomal gene transcription. The Biochemical Journal *306 (Pt 3)*, 617–626.

Johnson, D., Mortazavi, A., Myers, R., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science *316*, 1497–1502.

Johnson, J., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P., Armour, C., Santos, R., Schadt, E., Stoughton, R., and Shoemaker, D. (2003). Genome-Wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays. Science *302*, 2141–2144

Johnson, C., Primorac, D., McKinstry, M., McNeil, J., Rowe, D., and Lawrence, J.B. (2000). Tracking COL1A1 RNA in osteogenesis imperfecta. splice-defective transcripts initiate transport from the gene but are retained within the SC35 domain. The Journal of Cell Biology *150*, 417–432.

Jonkers, I., Kwak, H., and Lis, J.T. (2014). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons.

Jonsson, J.J., Foresman, M.D., Wilson, N., and McIvor, R.S. (1992). Intron requirement for expression of the human purine nucleoside phosphorylase gene. Nucleic Acids Research *20*, 3191–3198.
Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature Methods *7*, 1009–1015.
Kent, W., Sugnet, C., Furey, T., Roskin, K., Pringle, T., Zahler, A., and Haussler, and (2002). The Human Genome Browser at UCSC. Genome Research *12*, 9961006.
Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D., and Ren, B. (2005). A high-resolution map of active promoters in the human genome. Nature *436*, 876–880.
Kim, M., Suh, H., Cho, E.-J., and Buratowski, S. (2009). Phosphorylation of the yeast Rpb1 C-terminal domain at serines 2, 5, and 7. The Journal of Biological Chemistry *284*, 26421–26426.
Kim, T.H., and Ren, B. (2006). Genome-wide analysis of protein-DNA interactions. Annual Review of Genomics and Human Genetics *7*, 81–102.
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 Biology *14*, R36.
Kim, E., Magen, A., and Ast, G. (2007). Different levels of alternative splicing among eukaryotes. Nucleic Acids Research *35*, 125–131.
Komarnitsky, P., Cho, E., and Buratowski, S. (2000). Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. Genes & Development *14*, 24522460.
Kouzarides, T. (2007). Chromatin Modifications and Their Function. Cell *128*.
Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Research *42*, D68–73.
Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science (New York, N.Y.) *339*, 950–953.
Lacadie, S., Tardiff, D., Kadener, S., and Rosbash, M. (2006). In vivo commitment to yeast cotranscriptional splicing is sensitive to transcription elongation mutants. Genes & Development *20*, 2055–2066.
Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods *9*, 357–359.
Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S., and Kim, V. (2004). MicroRNA genes are transcribed by RNA polymerase II. The EMBO Journal *23*, 4051-4060.
Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S., et al. (2003). The nuclear RNase III Drosha initiates microRNA processing. Nature *425*, 415–419.
Lee, Y., Jeon, K., Lee, J., Kim, S., and Kim, V. (2002). MicroRNA maturation: stepwise processing and subcellular localization. The EMBO Journal.
Lee, R., Feinbaum, R., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell *75*, 843854.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) *25*, 2078–2079.

Li, H., Zhang, Z., Wang, B., Zhang, J., Zhao, Y., and Jin, Y. (2007). Wwp2-Mediated Ubiquitination of the RNA Polymerase II Large Subunit in Mouse Embryonic Pluripotent Stem Cells. Molecular and Cellular Biology *27*, 5296–5305.

Liu, D., Brockman, J., Dass, B., Hutchins, L., Singh, P., McCarrey, J., MacDonald, C., and Graber, J. (2007). Systematic variation in mRNA 3′-processing signals during mouse spermatogenesis. Nucleic Acids Research *35*, 234–246.

Logan, J., Falck-Pedersen, E., Darnell, J., and Shenk, T. (1987). A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene. Proceedings of the National Academy of Sciences *84*, 8306–8310.

Lund, E., Güttinger, S., Calado, A., Dahlberg, J., and Kutay, U. (2004). Nuclear Export of MicroRNA Precursors. Science*303*, 95–98.

Lutz, C. (2008). Alternative polyadenylation: a twist on mRNA 3' end formation. ACS Chemical Biology *3*, 609–617.

Lynch, M., and Conery, J. (2003). The Origins of Genome Complexity. Science *302*, 1401–1404.

Ma, H., Wu, Y., Choi, J., and Wu, H. (2013). Lower and upper stem-single-stranded RNA junctions together determine the Drosha cleavage site. Proceedings of the National Academy of Sciences.

Mandel, C., Kaneko, S., Zhang, H., Gebauer, D., Vethantham, V., Manley, J., and Tong, L. (2006). Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. Nature *444*, 953–956.

Martin M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal

Maxwell, E.K., Ryan, J.F., Schnitzler, C.E., Browne, W.E., and Baxevanis, A.D. (2012). MicroRNAs and essential components of the microRNA processing machinery are not encoded in the genome of the ctenophore Mnemiopsis leidyi. BMC Genomics *13*, 714.

Mendell, J. (2008). miRiad roles for the miR-17-92 cluster in development and disease. Cell *133*, 217–222.

Morin, R., O'Connor, M., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.-L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., et al. (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Research *18*, 610–621.

Myer, V., and Young, R. (1998). RNA Polymerase II Holoenzymes and Subcomplexes. Journal of Biological Chemistry *273*, 27757–27760.

Napolitano, G., Lania, L., and Majello, B. (2013). RNA polymerase IICTD modifications: How many tales from a single tail. Journal of Cellular Physiology 229, 538-544.

Nechaev, S., Fargo, D., Santos, G., Liu, L., Gao, Y., and Adelman, K. (2010). Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science (New York, N.Y.) *327*, 335–338.

Neugebauer, K. (2002). On the importance of being co-transcriptional. Journal of Cell Science *115*, 3865–3871.
Ni, Z., Saunders, A., Fuda, N.J., Yao, J., Suarez, J.-R.R., Webb, W.W., and Lis, J.T. (2008). P-TEFb is critical for the maturation of RNA polymerase II into productive elongation in vivo. Molecular and Cellular Biology *28*, 1161–1170.
Nojima, T., Dienstbier, M., Murphy, S., Proudfoot, N.J., and Dye, M.J. (2013). Definition of RNA polymerase II CoTC terminator elements in the human genome. Cell Reports *3*, 1080–1092.
Orphanides, G., LeRoy, G., Chang, C.H., Luse, D.S., and Reinberg, D. (1998). FACT, a factor that facilitates transcript elongation through nucleosomes. Cell *92*, 105–116.
Pandya-Jones, A., and Black, D. (2009). Co-transcriptional splicing of constitutive and alternative exons. RNA (New York, N.Y.) *15*, 1896–1908.
Pertea, M., and Salzberg, S. (2010). Between a chicken and a grape: estimating the number of human genes. Genome Biology *11*, 206.
Peterlin, B.M., and Price, D.H. (2006). Controlling the elongation phase of transcription with P-TEFb. Molecular Cell *23*, 297–305.
Phatnani, H., and Greenleaf, A. (2006). Phosphorylation and functions of the RNA polymerase II CTD. Genes & Development *20*, 2922–2936.
Proudfoot, N.J., Furger, A., and Dye, M.J. (2002). Integrating mRNA processing with transcription. Cell *108*, 501–512.
Quinlan, A., and Hall, I. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England) *26*, 841–842.
Rahl, P.B., Lin, C.Y., Seila, A.C., Flynn, R.A., McCuine, S., Burge, C.B., Sharp, P.A., and Young, R.A. (2010). c-Myc regulates transcriptional pause release. Cell *141*, 432–445.
Ramsey-Ewing, A., Van Wijnen, A.J., Stein, G.S., and Stein, J.L. (1994). Delineation of a human histone H4 cell cycle element in vivo: the master switch for H4 gene transcription. Proceedings of the National Academy of Sciences of the United States of America *91*, 4475–4479.
Rana, T. (2007). Illuminating the silence: understanding the structure and function of small RNAs. Nature Reviews Molecular Cell Biology *8*, 23–36.
Rearick, D., Prakash, A., McSweeny, A., Shepard, S., Fedorova, L., and Fedorov, A. (2011). Critical association of ncRNA with introns. Nucleic Acids Research *39*, 2357–2366.
Reed, R., and Maniatis, T. (1988). The role of the mammalian branchpoint sequence in pre-mRNA splicing. Genes & Development *2*, 1268–1276.
Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature Methods *4*, 651–657.
Roeder, R., and Rutter, W. (1969). Multiple Forms of DNA-dependent RNA Polymerase in Eukaryotic Organisms. Nature *224*, 234–237.
Sammeth, M., Foissac, S., and Guigó, R. (2008). A general definition and nomenclature for alternative splicing events. PLoS Computational Biology *4*, e1000147.
Saunders, A., Core, L., and Lis, J. (2006). Breaking barriers to transcription elongation. Nature Reviews Molecular Cell Biology *7*, 557-569.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science (New York, N.Y.) *270*, 467–470.
Shi, Y., Giammartino, D., Taylor, D., Sarkeshik, A., Rice, W., Yates, J., Frank, J., and Manley, J. (2009). Molecular architecture of the human pre-mRNA 3' processing complex. Molecular Cell *33*, 365–376.
Shpakovski, G.V., Acker, J., Wintzerith, M., Lacroix, J.F., Thuriaux, P., and Vigneron, M. (1995). Four subunits that are shared by the three classes of RNA polymerase are functionally interchangeable between Homo sapiens and Saccharomyces cerevisiae. Molecular and Cellular Biology *15*, 4702–4710.
Sims, R.J., Rojas, L.A., Beck, D., Bonasio, R., Schüller, R., Drury, W.J., Eick, D., and Reinberg, D. (2011). The C-terminal domain of RNA polymerase II is modified by site-specific methylation. Science (New York, N.Y.) *332*, 99–103.
Siomi, H., and Siomi, M. (2009). On the road to reading the RNA-interference code. Nature *457*, 396–404.
Stargell, L.A., and Struhl, K. (1996). Mechanisms of transcriptional activation in vivo: two steps forward. Trends in Genetics : TIG *12*, 311–315.
Stein, G.S., van Wijnen, A.J., Stein, J.L., Lian, J.B., Montecino, M., Zaidi, S.K., and Braastad, C. (2006). An architectural perspective of cell-cycle control at the G1/S phase cell-cycle transition. Journal of Cellular Physiology *209*, 706–710.
Steitz, J., Dreyfuss, G., Krainer, A., Lamond, A., Matera, A., and Padgett, R. (2008). Where in the cell is the minor spliceosome? Proceedings of the National Academy of Sciences of the United States of America *105*, 8485–8486.
Talerico, M., and Berget, S. (1990). Effect of 5' splice site mutations on splicing of the preceding intron. Molecular and Cellular Biology *10*, 6299–6305.
Tarn, W.Y., and Steitz, J.A. (1996). A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. Cell *84*, 801–811.
Teixeira, A., Tahiri-Alaoui, A., West, S., Thomas, B., Ramadass, A., Martianov, I., Dye, M., James, W., Proudfoot, N., and Akoulitchev, A. (2004). Autocatalytic RNA cleavage in the human β-globin pre-mRNA promotes transcription termination. Nature *432*, 526–530.
Terzi, N., Churchman, L., Vasiljeva, L., Weissman, J., and Buratowski, S. (2011). H3K4 trimethylation by Set1 promotes efficient termination by the Nrd1-Nab3-Sen1 pathway. Molecular and Cellular Biology *31*, 3569–3583.
Tilgner, H., Knowles, D., Johnson, R., Davis, C., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T., and Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Research *22*, 1616–1625.
Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.
Uzawa, T., Yamagishi, A., and Oshima, T. (2002). Polypeptide synthesis directed by DNA as a messenger in cell-free polypeptide synthesis by extreme thermophiles, Thermus thermophilus HB27 and Sulfolobus tokodaii strain 7. Journal of Biochemistry *131*, 849–853.

Veerla, S., and Höglund, M. (2006). Analysis of promoter regions of co-expressed genes identified by microarray analysis. BMC Bioinformatics *7*, 384.
Weinmann, R., and Roeder, R. (1974). Role of DNA-Dependent RNA Polymerase III in the Transcription of the tRNA and 5S RNA Genes. Proceedings of the National Academy of Sciences *71*, 1790–1794.
West, S., Proudfoot, N.J., and Dye, M.J. (2008). Molecular dissection of mammalian RNA polymerase II transcriptional termination. Molecular Cell *29*, 600–610.
West, S., Gromak, N., and Proudfoot, N.J. (2004). Human 5' --> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. Nature *432*, 522–525.
Will, C., and Lührmann, R. (2005). Splicing of a rare class of introns by the U12-dependent spliceosome. Biological Chemistry *386*.
Will, C., and Lührmann, R. (2011). Spliceosome structure and function. Cold Spring Harbor Perspectives in Biology *3*.
Willis, I. (1993). RNA polymerase III. Genes, factors and transcriptional specificity. European Journal of Biochemistry*212*, 111
Woychik, N., and Hampsey, M. (2002). The RNA Polymerase II MachineryStructure Illuminates Function. Cell *108*, 453–463.
Yamaguchi, Y., Takagi, T., Wada, T., Yano, K., Furuya, A., Sugimoto, S., Hasegawa, J., and Handa, H. (1999). NELF, a Multisubunit Complex Containing RD, Cooperates with DSIF to Repress RNA Polymerase II Elongation. Cell *97*.
Yao, C., Choi, E.-A., Weng, L., Xie, X., Wan, J., Xing, Y., Moresco, J., Tu, P., Yates, J., and Shi, Y. (2013). Overlapping and distinct functions of CstF64 and CstF64τ in mammalian mRNA 3' processing. RNA (New York, N.Y.) *19*, 1781–1790.
Yeom, K.-H., Lee, Y., Han, J., Suh, M., and Kim, V. (2006). Characterization of DGCR8/Pasha, the essential cofactor for Drosha in primary miRNA processing. Nucleic Acids Research *34*, 4622–4629.
Zarkower, D., Stephenson, P., Sheets, M., and Wickens, M. (1986). The AAUAAA sequence is required both for cleavage and for polyadenylation of simian virus 40 pre-mRNA in vitro. Molecular and Cellular Biology *6*, 2317–2323.
Zeng, Y., Yi, R., and Cullen, B. (2004). Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. The EMBO Journal *24*, 138148.