Pedro António Pereira Prudêncio

# Pre-mRNA splicing and regulation of gene expression after *Drosophila* egg fertilization.

Universidade do Algarve
Departamento de Ciências Biomédicas e Medicina
2020

Pedro António Pereira Prudêncio

# Pre-mRNA splicing and regulation of gene expression after *Drosophila* egg fertilization.

Doutoramento em Ciências Biomédicas

Trabalho efetuado sob a orientação de:

Orientador: Prof. Doutor Rui Gonçalo Viegas Russo da Conceição Martinho

Co-orientador: Professora Doutora Maria do Carmo Salazar Velez Roque da Fonseca

Universidade do Algarve

Departamento de Ciências Biomédicas e Medicina

2020

Declaração de autoria de trabalho

Titulo:

Pre-mRNA splicing and regulation of gene expression after *Drosophila* egg fertilization.

Declaro ser o autor deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

_____

Copyright © 2020 Pedro António Pereira Prudêncio

A Universidade do Algarve reserva para si o direito, em conformidade com o disposto no Código do Direito de Autor e dos Direitos Conexos, de arquivar, reproduzir e publicar a obra, independentemente do meio utilizado, bem como de a divulgar através de repositórios científicos e de admitir a sua cópia e distribuição para fins meramente educacionais ou de investigação e não comerciais, conquanto seja dado o devido crédito ao autor e editor respetivos.

# Acknowledgments

A todos os colegas, amigos, conhecidos e desconhecidos que deram aquela “mãozinha” na altura certa, e especialmente às seguintes pessoas que foram tão importantes para mim durante esta jornada:

Rui Martinho

Maria Carmo-Fonseca

Sérgio de Almeida

Mónica Dias

Kenny Rebelo

Rosina Savisaar

Rui Luís

Teresa Silva

Rita Drago

Pedro Barbosa

Gaston Guilgur

Paulo Navarro

Tânia Ferreira

Ana Rita Marques

Dinora Levi

Delfina Pereira

Júlia Rocha

Manuel Prudêncio

Fátima Pereira

Juan López

Liliana Prudêncio

Raquel Mendes

Maria Prudêncio

Tiago Prudêncio

Francisco Prudêncio,

o meu Obrigado.

# Resumo

A expressão génica depende de uma coordenação complexa entre diferentes processos, os quais dependem de muitos e diferentes fatores de regulação. Devido a essa complexidade, espera-se que isso cause restrições nos estádios de desenvolvimento mais exigentes dos organismos, e vice-versa. Neste trabalho, tentamos explorar essas restrições e como influenciam a expressão genica, a relevância desses processos sob essas restrições e quão eficientes esses processos são para se adaptar sob esses estádios de desenvolvimento.

Nesse sentido, decidimo-nos focar na expressão de genes zigóticos durante as rápidas divisões nucleares antes a transição da metade da blástula (MBT) do embrião de *Drosophila*, em que as curtas interfases limitam o tamanho dos genes aí expressos. Sabendo também que a mitose inibe no geral tanto a transcrição de genes como o seu splicing, e uma vez que a maioria dos genes expressos nesta fase não contem intrões, nós colocámos a hipótese de que provavelmente haveria também restrições no splicing nesse mesmo estádio de desenvolvimento. De acordo com essa hipótese, dois alelos mutantes para um fator de splicing nomeado *fandango* e homologo ao gene Xab2, uma das subunidades to complexo NTC/Prp19 em humano, foram isolados. Estes alelos apresentam defeitos na formação da blastoderme típicos de alelos mutantes de genes zigóticos. Ao analisar os transcritos expressos, reparamos que especificamente os genes zigoticamente expresses apresentavam defeitos de splicing, defeito esse, não observado nos genes transcritos maternalmente e depositados no embrião. Além disso, a expressão ectópica materna de um transcrito zigótico foi suficiente para suprimir seus defeitos de splicing no mutante de *fandango*. Por fim, um pequeno transcrito zigótico modificado de forma a conter múltiplos intrões apresentou mais defeitos de splicing quando expresso no embrião do tipo selvagem, comparativamente a sua expressão maternal. Mostrando assim, a existência de um pré-requisito de splicing altamente eficiente durante o estádio de desenvolvimento embrionário inicial em *Drosophila*.

Apesar de 70% dos genes zigóticos expressos inicialmente serem curtos e não conterem intrões, a restante percentagem preserva intrões na sua arquitetura, sendo expressos durante um estádio critico para o processo de splicing. Desta forma, e uma vez que alguns destes intrões são bastante conservados entre espécies relativamente a sua posição, questionámo-nos se estes poderiam desempenhar alguma função relevante, nomeadamente durante a expressão do

próprio gene. Para testar essa hipótese, fomos testar a possível funcionalidade de dois intrões conservados pertencentes a dois genes que são expressos antes da fase MBT e muito bem caracterizados: knirps (*kni*) e even-skipped (*eve*). Pare esse efeito, foram produzidas moscas transgênicas de *Drosophila* contendo inserções genómicas que possuíam o respetivo gene com ou sem intrão e respetivas regiões intergénicas reguladoras. Infelizmente, não foi possível obter moscas transgénicas do gene *kni*, não permitindo tirar uma conclusão sobre a função desse intrão nesse gene. Surpreendentemente, ambos o transgénico de *eve* com e sem intrão, suprimiram completamente o fenótipo de viabilidade do alelo mutante nulo do próprio gene. Além disso, não foram observadas alterações nos níveis transcritos gerados entre os dois casos, sugerindo assim não existir uma função para este intrão testado, pelo menos nas condições testadas.

Para quantificar a eficiência de splicing em embriões de *Drosophila*, e ver se esta varia ao longo do desenvolvimento, tirámos partido da sequenciação dos transcritos em elongação nativa (*d*NET-seq). Esta técnica permite identificar a posição da RNA polimerase II (Pol II) com resolução de 1 nucleótido enquanto o gene esta a ser transcrito, e estabelecer a ligação dessa posição com a informação de splicing ocorrido no transcrito. Para tal foram isolados núcleos de embriões nos estádios correspondentes ao estádio MBT (early) e pós-MBT (late). A cromatina foi consequentemente digerida com Micrococcal nuclease (MNase) de modo solubilizar os complexos contento o DNA, e o RNA nascente ligado à Pol II. Estes complexos foram imunoprecipitados usando anticorpos específicos para os resíduos fosforilados, Serina 2 e Serina 5, do Domínio carboxi-terminal (CTD) de uma das subunidades da Pol II. Por consequente o RNA contido nesses complexos foi isolado, fracionado, e ligados especificamente a adaptadores de modo a garantir que unicamente os RNAs com grupo hidroxilo 3′ terminal fossem sequenciados.

Desta forma foi possível observar sinal proveniente de genes zigóticos transcricionalmente ativos no embrião, enquanto que genes expressos maternalmente e depositados nos embriões, mas transcricionalmente inativos, não apresentaram sinal significativo. Mostrando assim, que o sinal detetado era especificamente proveniente de transcritos nascentes. Verificou-se também que a eficiência na terminação da transcrição estava associada à ausência de genes proximais a jusante ao gene analisado, e curiosamente, que genes pré-MBT apresentavam essas mesmas características. Enquanto que genes sem genes posicionados nos 500 pb a jusante, apresentavam pouca densidade de Pol II jusante ao sinal de

terminação, genes convergentes que se sobrepõem exibiam mais Pol II a seguir a esse mesmo sinal, transcrição essa não correlacionada com a transcrição do gene convergente.

Detetámos também reads abrangendo junções recursivas de splicing e entre exões em transcritos conectados ao local ativo das moléculas de Pol II posicionadas alguns nucleotídeos a jusante dos locais de splicing recursivos e canônicos 3′. Indicando que o splicing pode ocorrer logo após um local de splicing, ou seja, muito próximo do canal de saída da Pol II. Quantificando a eficiência de splicing com base no rácio entre reads spliced e reads totais detetados nos primeiros nos primeiros 100 nucleóticos a jusante do local de splicing 3′, foi possível correlacionar essa eficiência com algumas características. Entre essas características, tanto a força do sinal de splicing presente na sequência 3′, como o alto conteúdo em nucleotidos GC e o facto dos intrões não pertencerem nem à primeira nem às últimas posições, correlacionou-se com o facto de haver mais casos com evidencia de splicing imediato. Inesperadamente, enquanto bastantes intrões curtos apresentavam evidencias de splicing imediato, concordante com o mecanismo de splicing de definição de intrão, intrões muito longos, onde era esperado splicing por definição exonica também apresentaram muitos casos em que o splicing ocorrera antes da Pol II terminar a transcrição do exão. Esta observação, argumenta que a definição de exão não é obrigatória para splicing em metazoários. Por último, e além do aumento de densidade de Pol II observado na generalidade dos exões relativamente aos intrões, encontrámos ainda padrões especificos de pausa da Pol II associados aos diferentes níveis de eficiência de splicing, sugerindo um mecanismo acoplado de splicing que diminui o alongamento da transcrição.

Tal como mostrado anteriormente em células humanas, *d*NET-seq foi capaz de detetar snRNAs pertencentes ao spliceosoma e productos intermediárias resultantes da primeira reação de splicing, mostrando uma vez mais que o spliceosoma se associa a Pol II ativa. Esses intermediários de splicicing juntamente com casos em que reads mostram a junção dos exões, permitiram concluir que 95% dos intrões analisados podem ser removidos co-transcricionalmente.

# Palavras-chave

Embriogénese em *Drosophila*;

Splicing;

Transcrição,

*d*NET-seq

# Abstract

Our aim is to explore the way developmental processes influence and are influenced by gene expression kinetics. We focused our work on *Drosophila* pre-midblastula transition (pre-MBT), during which the extremely fast syncytial nuclear divisions are known to impose significant constraints to transcription. Since most pre-MBT genes are intronless, we hypothesized significant constraints to splicing are also likely to occur. Accordingly, mutants for the spliceosome NineTeen complex (NTC) subunit Fandango impaired efficient splicing of pre-MBT but not maternally encoded transcripts. Furthermore, a small early zygotic transcript with multiple introns was poorly spliced in wild-type embryos, which suggests a developmental pre-requisite for highly efficient splicing during *Drosophila* early embryonic development.

Although most pre-MBT genes are intronless, many patterning genes have evolutionary conserved introns. Given this conservation, we hypothesized that some of these introns are functionally relevant for the correct expression of these genes. Nevertheless, a large genomic construct containing an intron-deleted even-skipped (eve) transgene fully complemented the patterning defects of a *eve* null allele, suggesting no obvious functional requirement for this intron.

In order to directly study splicing kinetics in a developing *Drosophila* embryo, we took advantage of the native elongating transcript sequencing (*d*NET-seq) to identify the position of RNA polymerase II when introns become spliced. Besides finding that most introns are co-transcriptionally spliced, we showed that in most cases this process occurs when Pol II is found paused few nucleotides past the 3′ss, which in your turn is associated to specific sequence and gene architecture features. Moreover, transcription termination efficiency was found to be associated to absence of proximal downstream genes, while genes with convergent genes show transcriptional-readthrough. Interestingly, pre-MBT genes were typically isolated and/or within large introns of other genes and splicing was typically immediate, which confirms a significant optimization of gene expression beyond small transcriptional unit size.

# Keywords

*Drosophila* embryogenesis;

Splicing;

Transcription;

*d*NET-seq.

# Table of contents

# Abbreviations

| | |
|---|---|
| 3′ss | 3′ splice site |
| 3′UTRs | 3′ untranslated regions |
| 5′ss | 5′ splice site |
| CBC | Capping Binding Complex |
| ChIP | Chromatin Immunoprecipitation |
| CPA | Cleavage and Polyadenylation |
| CTD | Carboxyl Terminal Domain |
| dPCR | Digital PCR |
| DNA | Deoxyribonucleic acid |
| DSE | Downstream Sequence Element |
| ESEs | Exonic Splicing Enhancers |
| ESSs | Exonic Splicing Silencers |
| FISH | Fluorescent in situ hybridization |
| GRO-seq | Global Run-on sequencing |
| IBP | Intron-Binding Complex |
| ISEs | Intronic Splicing Enhancers |
| ISSs | Intronic Splicing Silencers |
| LECA | Last Eukaryotic Common Ancestor |
| MBT | Mid-blastula transition |
| mRNA | Messenger RNA |
| MZT | Maternal to Zygotic Transition |
| ncRNAs | Non-Coding RNAs |
| NET-seq | Native Elongating Transcript sequencing |
| NMD | Nonsense-Mediated Decay |
| NTC | Nineteen Complex |
| pA | Polyadenylation |
| PCR | Polymerase Chain Reaction |
| PIC | Pre-Initiation Complex |
| Pol II | RNA Polymerase II |
| PRO-seq | Precision nuclear Run On sequencing |
| PTCs | Premature Termination Codons |

| | |
|---|---|
| qPCR | Quantitative Real time PCR |
| RNA | Ribonucleic acid |
| rRNAs | Ribosomal RNAs |
| RS | Recursive sites |
| SMRT | Single Molecule, Real-Time |
| snRNPs | Small Nuclear Ribonucleoproteins |
| SREs | Splicing Regulatory Elements |
| TBP | TATA-Binding Protein |
| TCR | Transcription Couple Repair |
| tRNAs | Transfer RNAs |
| TSS | Transcription Star Site |
| TT-seq | Transient Transcriptome sequencing |
| USE | U-rich upstream sequence element |

# 1. Introduction

## 1.1. The Code of Life

Genetic information is stored and decoded into three different biopolymers: Deoxyribonucleic acid (DNA), Ribonucleic acid (RNA) and Proteins, which are transversal to all known forms of life.

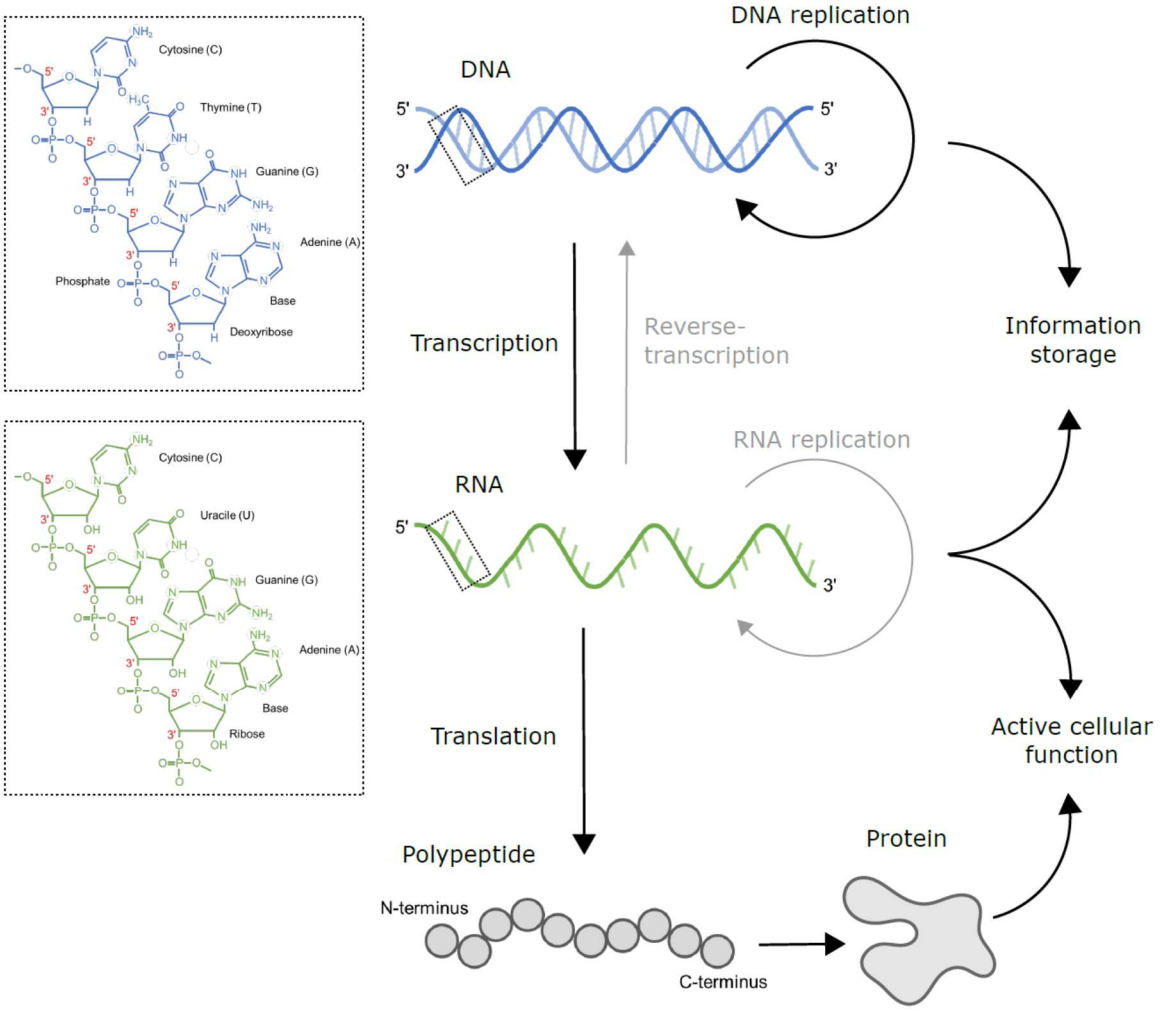



**Figure 1.1. The flow of genetic information.** Scheme depicting the central dogma in biology and how information can be transmitted between the different polymers, as well their general function (right). General information transfer (black) and special information transfer (grey). Chemical structure depicting the differences between DNA and RNA polymers (left).

DNA is composed by two chains, also known as DNA strands, containing monomeric units called nucleotides. Each nucleotide is composed of one of four nitrogen-containing nucleobases (cytosine [C], guanine [G], adenine [A] or thymine [T]), a sugar called deoxyribose, and a phosphate group. Along each strand the nucleotides are covalently bound between the sugar of one nucleotide and the phosphate of the next, resulting in an alternating sugar-phosphate backbone. Being their variable nucleobases distributed along the strand that will encode the genetic information. The orientation of the 3′ and 5′ carbons along the sugar-phosphate backbone confers directionality to each DNA strand. In a nucleic acid double helix, the strands are antiparallel, meaning that the direction of the nucleotides in one strand is opposite to their direction in the other strand. The two chains interact each other by complementarity of nucleobases where adenine pairs only to thymine in two hydrogen bonds, and cytosine to guanine in three hydrogen bonds, forming the base pairs (Figure 1.1). The structure of DNA helps to confer the necessary chemical stability to preserve genetic information, as well ensuring replication fidelity during cell division. At the same time, the DNA molecule can easily accommodate small alterations of the encoded genetic information, allowing cells and organisms to evolve.

Within the cell, genomic DNA is packed in structures called chromosomes. While in prokaryotes the DNA is stored in the cytoplasm, in supercoiled circular chromosomes. In eukaryotic cells their genomic DNA is stored inside the cell nucleus as nuclear DNA and organized into a polymeric complex called chromatin. This DNA-protein complex is composed by repeated structures called nucleosomes along the DNA structure, which are composed by an octameric complex of the core histone proteins wrapping 145 to 147 bp of DNA (McGinty and Tan, 2015; Richmond and Davey, 2003).

Unlike DNA, RNA is often found in nature as a single-strand folded onto itself, rather than a paired double-strand. Composed by one chain of nucleotides, RNA differs in two characteristics from DNA:  instead of a deoxyribose has a ribose and while DNA has thymine nucleobase, RNA in its place has an uracil. One possible explanation for this, is the fact that thymine has grater resistance to photochemical mutations, those contributing for a more stable DNA molecule (Lesk, 1969), at the same time preventing the spontaneous deamination of cytosines into uracil to alter the content of genomes during evolution (Lehninger et al., 2013). All this evidences and others, support the theory that DNA evolved from RNA molecules. Due their versatile structure, the RNA molecule not only stores DNA genetic information for protein synthesis, but also performs many other regulatory functions within the cell, which are likely

to represent the majority of the cases, since whereas only 2% of the human genome code for proteins, 90% is known to be transcribed (The ENCODE Project Consortium, 2007; Wilusz et al., 2009). Although non-coding RNAs represents a large fraction of functional molecules in the cell, their precise number and functions are still poorly understood (Uszczynska-Ratajczak et al., 2018).

Proteins are composed by a chain of amino acids that are linked by peptide bonds stablished by the polymerization between amino and carboxylic acid groups attached to the alpha carbon of each of the twenty amino acids. The number and biochemical variability of amino acids, allowed proteins to acquire more elaborated conformations and therefore, more diverse and specialized functions in the cell.

### 1.1.1. Transfer of Information

These three polymers (DNA, RNA and proteins), enable cells to store genetic information, transfer it whenever needed, and use it to synthesize new functional molecules. Despite, some especial cases, like in virus (grey arrows in Figure 1.1), in most types of living cells (from prokaryotes to eukaryotes), the transfer of information occurs in one way: the DNA is copied into DNA by the DNA polymerase and all replisome machinery associated in a processed called DNA replication. In addition, DNA can be also copied into RNA by the RNA polymerase and in many cases requires post processing (RNA transcription and processing) to form a messenger RNA (mRNA), that in your turn, will be copied into proteins by the ribosomes (translation), or instead, be used as regulatory non-coding RNAs (ncRNAs) (Figure 1.1). In eukaryotes, each of these processes occurs in a specific compartment of the cell. Transcription - takes place in the nucleus, where synthesized mRNAs are exported to the cytoplasm and where translation into proteins happens. In prokaryotes, and since there is no nucleus, these processes are not compartmentalized and can occur at the same time in structures called polysomes. In this way, while prokaryotes control their gene expression at the level of transcription, eukaryotic cells do it at different levels in a highly regulated and coordinated process that controls the timing, amount and localization of each molecule synthetized in each cell (review in (Madhani, 2013)).

## 1.2. The Complex Gene Expression regulation during Transcription Cycle

To maintains the cell homeostasis, differentiate or respond to external stimulus, cells rely on an elaborated and complex process that ensures the expression of the correct genes at the correct amount and time. Requirements of more complex systems like eukaryotes, increased also the complexity in the way how each gene expression is regulated. In a process that starts with the definition of gene transcription start site, followed by initiation of transcription, elongation that normally is coupled to transcript processing and finally termination, elaborated mechanisms of regulation can be found in each of these steps which do in many cases influence each other.

Understand the mechanisms beyond transcription regulation has been an important key to perceive the causes of many diseases (Lee and Young, 2013), which has been opening the opportunity for development of new targeted therapies in the past years (reviewed in (Desterro et al., 2020)). However, and although the first human genome was sequenced almost 20 years ago (Venter et al., 2001) among other species, the way this code is interpreted by the cell, still has many questions to be answered. During the last decades many molecular mechanisms has been proposed to explained how each of those processes are regulated. Although, "single gene" experiments performed with reporter genes were crucial to better understand how each of this process works, latest genome-wide studies came to show that some of these cannot be the rule, meaning that transcription and mRNA processing is much more complex than previously though.

### 1.2.1. RNA Polymerase

From prokaryotes to eukaryotes, RNA Polymerase is the enzyme responsible to synthesize the new RNA molecule from the DNA template, and therefore, the centre of transcription regulation. This is due to its ability to interact with many regulatory factors. Depending on species or the type of gene transcribed, different mechanisms of regulation, as well as different polymerases can be found. In bacteria the RNA polymerase together with a sigma factor, can both recognize two short sequences located -10 and -35 nucleotides from the transcription star site (TSS). Additional sequences that recruit other associated factors or

variation on those sequence elements, will influence RNA polymerase binding, thus affecting the expression of downstream genes (Barne et al., 1997; Browning and Busby, 2004; Haugen et al., 2006; Helmann and deHaseth, 1999; Ross et al., 1993). On the other side, RNA polymerase recruitment can be blocked by repressors like Lac or λ that bind to promoter, or instead, by anti-sigma factors that sequestrates sigma factors preventing their binding to RNA polymerase or to the promoter (Browning and Busby, 2004; Helmann, 1999).

In eukaryotes, three different RNA polymerases (Roeder and Rutter, 1969), and an additionally two in the case of plants (Zhou and Law, 2015), are responsible for the synthesis of different sets of genes. RNA polymerase I is responsible to transcribe genes encoding the large ribosomal RNAs (rRNAs), whereas RNA polymerase III makes transfer RNAs (tRNAs) and other small ncRNAs that includes: the small rRNA and U6 small nuclear RNA. In its turn, RNA polymerase II (Pol II) is responsible to transcribe all protein-coding genes and also many ncRNAs. Pol II is composed by 12 subunits, which are quite conserved from yeast to human (Myer and Young, 1998) and share 10 subunits with Pol I and Pol III. (Wild and Cramer, 2012). However, the fundamental difference that distinguishes Pol II from all other polymerases is the carboxyl terminal domain (CTD) of the largest subunit RPB1. This domain is involved in many processes of RNA biogenesis typical of mRNAs and non-coding RNAs transcribed by this polymerase (see below). In plants, additional Pol IV and V have dedicated roles to gene silencing (Zhou and Law, 2015).

### 1.2.2. Promoter recognition

Before transcription initiation, Pol II is recruited to sequence regions located upstream to TSS of genes, the so-called promoters. These regions are prone for gene transcription regulation through the control of three different characteristics, that could either, block or facilitate the binding of Pol II: chromatin DNA accessibility, sequence motifs which are recognized by transcription factors and the ability of DNA to be epigenetically modified (reviewed in (Müller and Tora, 2014))

Since nucleosomes can inhibit transcription initiation by blocking the access of Pol II to the DNA, active promoters are thus characterized by nucleosome-free regions which are flanked by specialized -1 and +1 nucleosomes (Schones et al., 2008; Talbert et al., 2019). These open chromatin regions can result for example, from: an inhibition to form nucleosome due to

the nucleotide composition or through the recruitment of transcription factors, also known as pioneers, that can recruit chromatin remodelers to alter the position of nucleosomes (Müller and Tora, 2014). For example, during *Drosophila* embryogenesis the transcription factor Zelda binds to nucleosome DNA and establishes chromatin accessible regions, which in turn will regulate the zygotic genome activation (Liang et al., 2008; McDaniel et al., 2019).

Promoter nucleotide sequences play an essential role defining the location where transcription is initiated. For instance, canonical promoters containing a TATA-box, are recognized by a TATA-binding protein (TBP), have a very focus transcription initiation, and are typically associate to cell-type-specific genes, tightly regulated during cell differentiation (Müller and Tora, 2014). However, the large majority of promoters in vertebrates lack the presence of this motif, instead they are characterized by the high density in CG dinucleotides, which are inhibitors of nucleosome formation, composing the so called CpG islands (Saxonov et al., 2006). These promoters have more dispersed transcription initiation, a characteristic commonly found in housekeeping genes (Larsen et al., 1992; Zhu et al., 2008) and are susceptible to inhibition through DNA methylation. By lacking the canonical TATA-box promoter sequence, many of these promoters rely on the recognition of their CG-rich sequences by Sp1 transcription factor, which recruits other transcription factors and promote initiation in a TATA-box independent manner (Butler and Kadonaga, 2002). Additionally, it was shown that the +1 nucleosome can also play a role on Pol II recruitment, by interacting with basal transcription factor complex (TFIID), which contains a TBP subunit (Vermeulen et al., 2007). In *Drosophila* and since there is no evidences of methylation patterns (Raddatz et al., 2013), dispersed promoters like CpG islands, such as the DNA replication-related element (DRE), Ohler1, Ohler6 and Ohler7 promoter motifs can be found (Ohler et al., 2002).

Besides promoters, transcription factors can also regulate transcription by binding on distant located regulatory regions, known as enhancers. Looping of DNA between these two distant structures, allows the interaction between these transcription factors at enhancers and the promoters (Kulaeva et al., 2012). This way, is possible to imagine the versatility in gene expression regulation of a gene under the control of several enhancers, that can bind to different transcription factors either activators or repressors. One of the most well documented cis-regulatory regions of this type, is the enhancer that controls the local expression of stripe 2 of *even skipped (eve)* gene in *Drosophila* embryo. This element, with just 480 bp, comprises 12 different binding sites for 4 different transcription factors, the repressors: Kruppel (Kr) and Giant (gt) and the activators: Bicoid (bcd) and Hunchback (hb). Therefore, the enhancer

becomes active only in cells containing activators and devoid of repressors, thus ensuring the correct expression of *eve* in stripe 2 region of the embryo (Borok et al., 2010).

### 1.2.3. Pre-initiation complex assembly

Even sufficient to perform RNA transcription from a DNA template, Pol II alone is not capable to precisely initiate transcription from a specific promoter (Young, 1991). Canonical Pol II recruitment to the DNA promoter, implies action of the well characterized general transcription factors (GTFs). TFIID the multiprotein complex composed by TBP, interacts with the TATA-box sequence in the promoter, binds and localizes Pol II to a downstream position in the DNA. TFIIB and TFIIF multiprotein complex helps to stabilize the pre-initiation complex and stimulates initial RNA synthesis forming the preinitiation complex (PIC) that extends over ~40 bp on either side of the TSS (Nogales et al., 2017). Then, TFIIE and TFIIH factors, which contains the DNA translocase XPB, binds downstream to Pol II and stimulates the DNA opening forming the transcription bubble that will allow Pol II access the DNA template strand (Schilbach et al., 2017). In the end of this process, transcription initiation by Pol II is stimulated by a complex known as Mediator. This complex contact both Pol II and TFIIH, stimulating the phosphorylation of carboxyl terminal domain (CTD) by TFIIH subunit CDK7 (Kornberg, 2005) (Figure 1.2 A). Although, this is the canonical view from PIC, recent evidences show that different complex conformation can be achieved to activate the transcription initiation (Andersen et al., 2017).

### 1.2.4. Promoter proximal Pol II Pausing

After recruitment to the promoter by PIC, Pol II becomes apt to initiate transcription elongation. However, a very early pause in this elongation is verified in many genes including both, active as well as less transcribed, but preferentially genes associated to developmental control (Zeitlinger et al., 2007). In fact, this phenomenon that normally occurs 50 bp downstream of TSS (promoter proximal pausing), appears to be highly regulated. This pause is controlled by the DRB sensitivity-inducing factor (DSIF) that binds to Pol II around the exiting of DNA and RNA, and the negative elongation factor (NELF), that binds near the Pol

II funnel, preventing TFIIS to release Pol II from pause. Pause release is performed by the positive transcription elongation factor b (P-TEFb) that contains the CDK9 kinase, which phosphorylates both DSIF and NELF and Pol II CTD at serine 2 residues, causing NELF to dissociate and converting DSIF to a positive elongation factor (Figure 1.2 B)(Yamaguchi et al., 2013; Zhou et al., 2012). Moreover, promoter sequence elements emerge also to play an important role controlling Pol II pausing. For instance, highly paused Pol II is associated to focused initiation promoters containing the following sequence elements: an initiator (Inr), a downstream promoter element (DPE), a motif ten element (MTE), the pause button (PB) and GAGA, while more broad initiation promoters are associated to moderated pausing. On the other hand, TATA promoters lacking any other pausing element, are associated with low degree of pausing (Gaertner and Zeitlinger, 2014). Evidences suggest that this mechanism of pausing, prepares the gene for rapid activation, like the case of *heat shock protein 70 (hsp70)* in *Drosophila* (Wang et al., 2005). Absence of pausing is associated to transcription bursts, allowing the production of many transcripts in a short period of time, as it happens in the case of early zygotic expressed genes in *Drosophila* embryo (see below) (Chen et al., 2013; Zenklusen et al., 2008).

### 1.2.5. Pol II CTD

One of the essential characteristics of Pol II is the ability to couple transcription with downstream mRNA processes, allowing the efficient and regulated expression of a variety of different types of coding genes and many non-coding RNAs. One of the main players in this process is the carboxyl terminal domain (CTD) of RPB1 subunit of Pol II. CTD is composed by repeated heptapeptides with the following consensus sequence: Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7. CTD is involved in the regulation of different processes during transcription cycle, through its ability to interact with a variety of factors involved from transcription initiation to 5′ capping, splicing, 3′end processing and also chromatin remodelling (Hsin and Manley, 2012).

A link between the number of repeats and the genomic complexity of species has been observed. Thus, while human CTD contains 52 repeats, yeast only contains 26. Work performed with reporter mammal genes, showed that, a minimum of 22 repeats are required and sufficient for the efficient transcription, splicing of constitutive but not alternative exons

and 3′ end cleavage functions (Rosonina, 2004). However, distinct CTDs mutants showed that specific regions of CTD are important for other specific functions (Custódio et al., 2007). On the other hand, yeast can grow under demanding conditions, with only half (13) of their 26 repeats (Nonet and Young, 1989). CTDs across species are composed by consensus and consensus divergent repeats. A recent work in *Drosophila*, which contains only 2 consensus repeats out of 44, showed that by replacing the divergent by many consensus, results in Pol II aggregations and impairment of *Drosophila* development, which gives an extra importance for the divergent CTD repeats (Lu et al., 2019). In fact, this observation can be explained by the recent findings related with the ability of low-complexity properties that CTD repeats have to establish non-covalent interactions. Thereafter, contributing to liquid-liquid phase separation and consequent formation of membraneless compartments or the so called condensates (Boehning et al., 2018; Mitrea and Kriwacki, 2016). Moreover, it has being suggested the existence of distinct types of condensates associated to transcription initiation and splicing, which may be important catalysts by compartmentalizing these processes (Cho et al., 2018; Guo et al., 2019).

CTD phosphorylation plays an important role during different stages of transcription cycle. These repeats can be modified at residues Tyr1, Ser2, Thr4, Ser5 and Ser7. CTD is unphosphorylated during PIC assembly and becomes phosphorylated at Ser5 and S7 by the THIIF-subunit Cyclin-dependent Kinase 7 (CDK7). Transcription elongation is promoted by the phosphorylation of Ser2 by CDK9 subunit of P-TEFb, which fits with the idea that phosphorylation of CTD help to dissolve the condensates referred before (Boehning et al., 2018). During elongation, and although many studies report a general reduction of Ser 5 and 7 phosphorylation, others correlate specifically the phosphorylation of Ser 5 with co-transcriptional splicing (Nojima et al., 2015a, 2018). During transcription termination, Ser2 and Thr4 phosphorylation are involved in the recruitment of cleavage and polyadenylation factors, as well as termination factors that release Pol II from the DNA (review in (Harlen and Churchman, 2017)). Since CTD must be hypophosphorylated for the polymerase to bind the promoter of transcribed genes, several phosphatases like Fcp1 are known to dephosphorylates Ser 2, while Scps, Sssu72 and Rtr1, were proposed to be involved in Ser 5 and other CTD residues dephosphorylation (Hsin and Manley, 2012).

### 1.2.6. Productive elongation

Transcriptional initiation and promoter pausing regulation steps are associated with a rapid turnover of Pol II near the promoter, being approximately only 1% of total recruited polymerases engaged into elongation. Since Pol II engaged in elongation is almost 100% stable (Price, 2018; Steurer et al., 2018), this clearly shows, how robust and regulated is the transition transcription elongation.

Although transcriptional elongation rate can vary between genes and even within the same gene, in general, slow rates of approximately 0.5 kb per minute are observed in the first kilobases, increasing to 2-5 kb per minute after ~15 kb (Jonkers and Lis, 2015). In agreement, similar rates has been observed recently in *Drosophila* for genes devoid of promoter proximal pausing (2.4-3.0 kb per minute) (Fukaya et al., 2017). However, Pol II have to face many obstacles during elongation that normally results in stalling or premature termination. These obstacles can be diverse such as: DNA sequences (consecutive thymidine residues in the non-transcribed DNA strand (T-runs) that stimulate termination) (Dedrick et al., 1987), DNA secondary structures, epigenetic DNA modification (5-formyl and 5-carboxyl-cytosine) (Kellinger et al., 2012), DNA lesions (Brueckner et al., 2007), small-molecules DNA-binders (distamycin or mithramycin) (Hardenbol and Van Dyke, 1992) or even other Pol II that can be either convergent (Hobson et al., 2012) or in tandem (Saeki and Svejstrup, 2009). But beyond all, the nucleosomes seem to be the major roadblocks during elongation of Pol II, thus causing significant pause few bp upstream of their location (Churchman and Weissman, 2011a). Although, many of pauses during elongation has been linked to different causes, the high frequency of Pol II pauses observed every 20-30 bp in yeast clearly shows that Pol II behaviour during transcription is much more complex than previously thought. After pausing, Pol II have the tendency to backtrack between 8-15 bp (Churchman and Weissman, 2011a; Nudler et al., 1997) leaving a 3′-end of nascent RNA exposed, which turn out to be cleaved by TFIIS, thus releasing again Pol II into elongation (Kulish and Struhl, 2001) (Figure 1.2 C).

Several factors are known to facilitate Pol II transcriptional elongation. After proximal promoter pausing release by P-TEFb, new elongation factors are recruited to Pol II. PAF1 complex (PAF) for instance, is known to recruit other elongation factors, and to play an important role in binding the Pol II and to compete with NELF. On the other hand, SPT6 binds to Pol II CTD, completing the elongation complex (Figure 1.2 D), and promoting efficient

transcription elongation by displacing and rearranging the nucleosomes, through their chaperon activity (reviewed in (Cramer, 2019; Kwak and Lis, 2013)). Moreover, histones modifications that compose the nucleosomes, also contribute to an increase in transcriptional elongation rates. For instance, acetylation of H3 at Lys56 (H3K56; mediated by histone acetyltransferase RTT109) and trimethylation of H3 at Lys36 (H3K36me3; mediated by the methyltransferase SET2). These marks are known to influence Pol II elongation rate directly, or indirectly, by forming a platform for histone chaperone binding, and subsequent histone or nucleosome removal and restoration in the wake of elongating Pol II (Jonkers and Lis, 2015).

Other phenomenon that causes pausing is the hybrid formation between DNA and nascent RNA, so-called R-loops, that normally occur in G-rich DNA regions. Accumulation of R-loops triggers histone modifications, which result in transcriptional termination and avoidance of genomic instability caused by those structures (Skourti-Stathaki et al., 2014). Accordingly, low CG content, normally found in introns is associated with higher elongation rates, whereas exons show slow transcriptional elongation rate, probably due to their higher CG content (Amit et al., 2012a) and/or higher nucleosome occupancy (Tilgner et al., 2009). Indeed, it has being hypothesised, whether this delay over exons could be associated also with co-transcription splicing, since the time that has been proposed for this reaction (20-30sec) matches with the delay caused during exon transcription (Jonkers and Lis, 2015; Jonkers et al., 2014a; Martin et al., 2013). Still, until now no direct evidences were shown.

Moreover, and similarly to exons, G-rich sequences located downstream to the poly (A) site, were shown to favour pausing of Pol II, which allows the RNA exonuclease XRN1 to contribute for an efficient transcription termination after RNA cleavage (Gromak et al., 2006) (see section 1.3.3 and Figure 1.5).

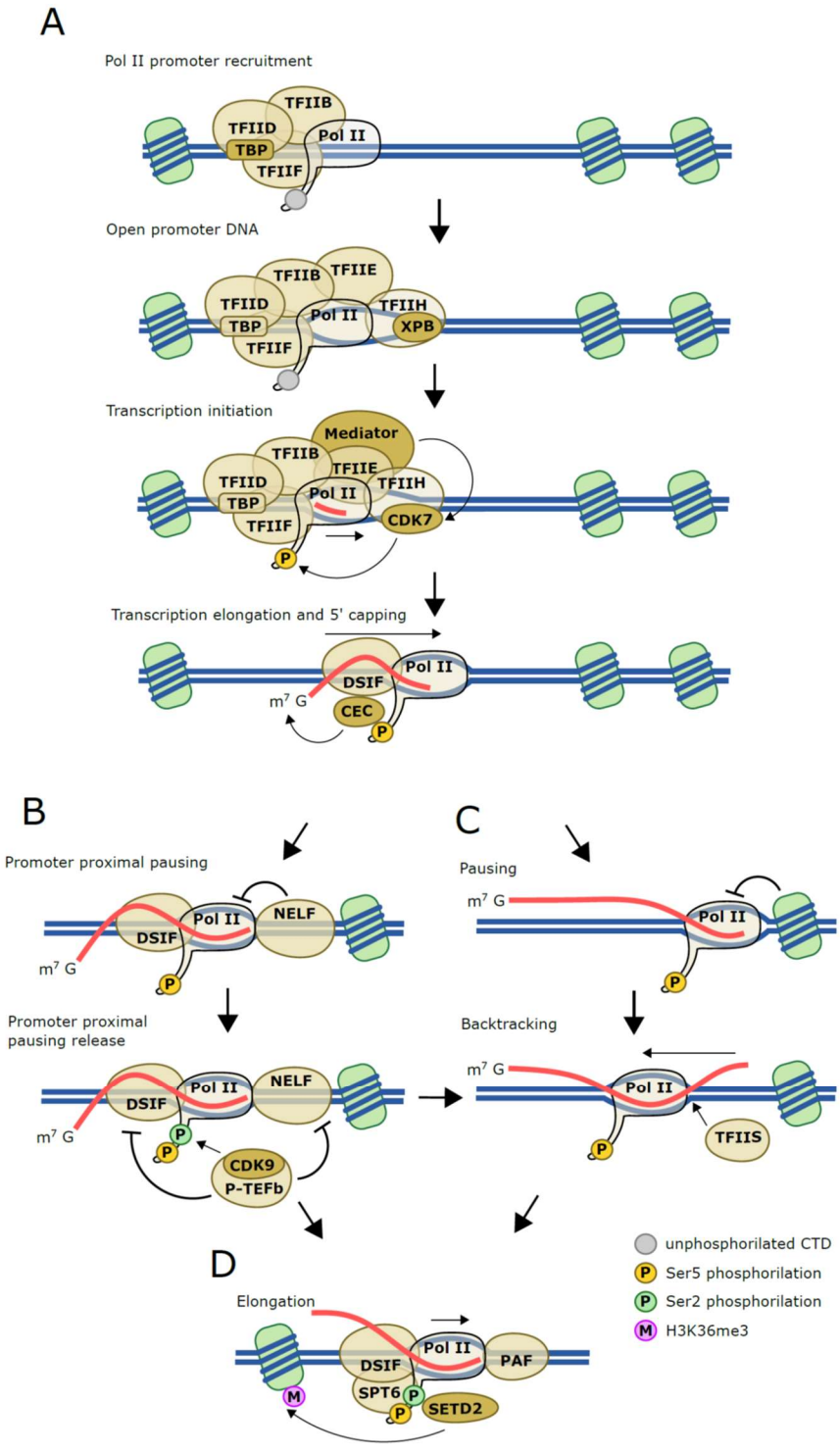



**Figure 1.2. Representation of processes associated to transcription initiation and elongation. (A)** Pre-initiation complex assembly and opening of DNA bubble by the DNA translocase XPB. Mediator complex stimulates the CTD Ser5 phosphorylation (P yellow) by TFIIH subunit CDK7. 5′ capping by the Capping enzyme complex (CEC). **(B)** Promoter proximal pausing is stimulated by NELF and DSIF. Phosphorilation by P-TEFb kinase CDK9 of CTD Ser2 (P green), NELF and DSIF promotes release from pausing. **(C)** Backtracking after Pol II pausing in solved by TFIIS 3′end RNA clevage. (D) Productive elongation is stimulated by several different elongation factors. DNA (blue), RNA (red) and nucleosomes (green).

## 1.3 mRNA processing

During and after transcription, pre-mRNA molecules goes under several modifications, that will be essential to regulate their stability and function in the cell. These processes include: the formation of secondary structures, chemical modifications like 5′ capping and 3′-end polyadenylations, interaction with regulatory proteins and editing by the removal of internal sequences (splicing).

### 1.3.1. 5′ capping

Maturation of pre-mRNA begins shortly after 20 to 30 of 5′-end nucleotides are transcribed by Pol II (Rasmussen and Lis, 1993). Capping factors (including the mRNA-capping enzyme and guanine-7-methyltransferase) are recruited to phosphorylated Ser5 CTD when Pol II is still paused in the promoter proximal region (Rasmussen and Lis, 1993). They catalyse 5′ capping by linking a 7-methylguanosine (m7G) to the last 5′ nucleoside of the mRNA through a 5′–5′ triphosphate bridge. While m7G protects the RNA from exonucleases degradation activity, it also facilitates the interaction of the mRNA 5′-end to other protein complexes. One of these complexes, the capping binding complex (CBC) was shown to be involved in many processes related with mRNA biogenesis like: transcription, splicing, 3′ processing and translation (review in (Gonatopoulos-Pournatzis and Cowling, 2014)). For example, in HeLa cells depletion of CBC resulted in the inhibition of pre-mRNA splicing and reduced the recruitment of U1 snRNP (a *small nuclear ribonucleoprotein*, composed by and RNA and several auxiliary proteins, which is part of the spliceosome) to the 5′ splice site of the 5′ proximal intron (Izaurralde et al., 1994; Lewis et al., 1996). Moreover, CBC was also shown to interact with U4/U6·U5 tri-snRNP and promote splicing of internal introns (Pabis et al., 2013).

Capping is a reversible modification that can occur in the cytoplasm and even co-transcriptionally in the nucleus, thus contributing to regulate the number of transcripts or even to provoke premature termination through degradation in a 5′→3′ direction (Davidson et al., 2012; Jiao et al., 2010).

### 1.3.2. Splicing

Splicing is the process by which the intronic sequence if an RNA is removed from the Pre-mRNA followed by ligation of the two flaking sequences, the exons. This process is catalysed by the Spliceosome, macromolecule complex, composed by 5 small nuclear ribonucleoproteins (snRNPs), each one containing a snRNA, as well more than 150 associated proteins (Will and Luhrmann, 2011). The Canonical spliceosome is composed by the U1, U2, U4, U6 and U5 snRNPs, which gradually acts together with other proteins to bind, rearrange, cleave, and ligate RNA sequences that define the exon-intron boundaries (splice sites) with single-nucleotide precision. This is very important, since it is necessary to keep the open reading frame intact after splicing. U1 snRNP, which contains the U1 snRNA, will recognize the 5′ splice site (5´ss) forming the Complex E. Definition of 3′ss region depends on splicing factor 1 (SF1) which binds the branch point site (BPS) that is located upstream of 3′ss, and U2 snRNP auxiliary factors (U2AF65 and U2AF35), which bind to the polypyrimidine tract (located between BPS and 3′ss) and to the 3′ss, respectively. This preassembly step is essential for the recruitment of U2 snRNP to the BPS which establishes the A Complex (Figure 1.3). Then, U1 and U2 snRNPs which are separated by the intron length and are brought together (by a mechanism discussed below), establishing a bridge between their component proteins and Prp5 (Shao et al., 2012). Next, preassembled U4/U6.U5 tri-snRNP, join the spliceosome forming the B complex, where U5 snRNP helps to join both exons. Through a series of rearrangements between RNA-RNA interactions, U6 interaction with U4 is lost, resulting in a new contact between the 5′ss U6 and U2, and the release of U1 and U4 from the spliceosome. During this stage the recruitment of the nineteen complex (NTC) and NTC-related in yeast or Prp19/CDC5L and the recently denominated intron-binding complex (IBP) () in humans, helps to remodel and maturate the spliceosome into the $B^{act}$ Complex (De et al., 2015). However, the spliceosome is still not ready for the first transesterification reaction. The ATPase Prp2, was shown to be essential to destabilize U2 snRNP subunits, SF3a and SF3b from the complex, exposing this way the BPS adenosine for the first transesterification reaction and thus converting the $B^{act}$ into B* Complex. Interestingly, this step does not require the 3′ss, suggesting that, during co-transcriptional splicing the first reaction can potentially occur as soon as the BPS is transcribed by the Pol II (Herzel et al., 2017; Rymond and Rosbash, 1985). C Complex results from the nucleophilic attack of 2'-hydroxyl group of the BPS adenosine on

the phosphate group at the 5′ss phosphodiester bond. This will generate a splicing intermediate 5′ exon with a 3′-hydroxyl group (3′-OH) and an intron lariat-3′ exon intermediate. In the second transesterification reaction the 3´OH of the 5´ exon attacks the first residue of the 3´ exon, resulting in the ligated exons and the release of the intron lariat. In the end, spliceosome components are release from the ligated exons and recycled for further rounds of splicing (Fourmann et al., 2013) (Figure 1.3).

Although the majority of introns (U2 type) are spliced by the canonical spliceosome as previously described, a small number of introns (U12 type) are spliced by a coexisting, but less abundant spliceosome that differs in some snRNPs. Thus, the minor spliceosome is composed by U11, U12, U4atac, and U6atac rather than U1, U2, U4 and U6 respectively, but shares the U5snRNP (Patel and Steitz, 2003).

**The Prp19 complex and related**

Although lacking the snRNA that characterizes the main components of spliceosome, NineTeen Complex (NTC) (also known as Prp19 complex) also plays important functions during the splicing reaction, behaving as a scaffold for recruiting other proteins to the spliceosome, and by facilitating conformational rearrangements during (Chan, 2003; McGrail et al., 2009) and after spliceosome activation (Chang et al., 2009). The core components of this complex are well conserved from yeast to human. However, the way these components are recruited to the spliceosome varies slightly between yeast and human. In yeast, NTC is composed by the core components: Prp19, Cef1, Syf1, Syf2, Syf3, Snt309, Isy1 and Ntc20, which are recruited as a pre-assembled complex to the spliceosome (Fabrizio et al., 2009; Hogg et al., 2010). In contrast, human Prp19/CDC5L only shares the homologous of Prp19, Cef1 and Snt309 (hPrp19, CDC5L and SPF27, respectively) (Makarova et al., 2004). Whereas the homologous of Syf1 and Isy1 (hSyf1/Xab2 and hIsy1, respectively) are found associated to the intron binding complex (IBC) composed by CypE, CCDC16 and Aquarius. Aquarius is an RNA helicase that is absent in yeast and binds to U2 subunits during $B^{act}$ spliceosome, increasing the splicing efficiency (De et al., 2015). Thus, suggesting that high eukaryotes have more sophisticated mechanisms during the splicing reaction.

Interestingly, other functions have also been attributed to NTC. For instance, this complex was found to be associated with transcribing Pol II, through the interaction with U2AF65 that binds to the CTD (David et al., 2011). Moreover, a different study also in yeast,

showed that Prp19 complex through his component Syf1, has a function in the transcription processivity by recruiting TREX complex to elongating Pol II (Chanarat et al., 2011). TREX complex onsists of the heterotetrameric THO complex (comprised of the proteins Tho2, Hpr1, Mft1, and Thp2), which turns out to be involved in different processes, like: transcriptional processivity, mRNA export, Transcription Couple Repair (TCR), and prevention of R-loops formation during transcription elongation (Gaillard et al., 2007; Jimeno, 2002; Katahira and Yoneda, 2009; Rondón et al., 2003). Supporting this idea, Xab2 in humans was also found to be involved in pre-mRNA splicing, transcriptional elongation, and TCR (Kuraoka et al., 2008; Nakatsu et al., 2000).

**Co-transcriptional splicing**

Originally considered a post-transcriptional process, splicing is currently considered to be mostly co-transcriptional. The first time that a transcript was found to be spliced, while still bound to the chromatin before termination, was by "Miller spread" under electron microscope (Beyer and Osheim, 1988). Since then, many studies have shown that the frequency of co-transcriptional is highly conserved among different species (Ameur et al., 2011; Brugiolo et al., 2013; Carrillo Oesterreich et al., 2010; Khodor et al., 2011a; Windhager et al., 2012) and typically range between 75-85% of splicing. Although the majority of introns are removed co-transcriptionally, there is however a class of introns that are retained during transcription, being their post-transcriptional splicing regulated by responses to changes in the cellular environment (Boutz et al., 2015).

Splicing coupled to transcription requires a crosstalk between the spliceosome and transcriptional machinery. Work with truncated carboxy terminal domain (CTD) forms of Pol II show that although genes were normally transcribed, splicing was impaired. Besides its role in 5′ Capping and transcriptional termination, Pol II CTD behaves as platform crucial for the crosstalk between Pol II and the spliceosome (Hirose and Manley, 2000). In agreement, several spliceosomal components were reported to interact with CTD. *In vitro* assays showed that polypyrimidine tract-binding protein-associated splicing factor (PSF), as well a closely related protein, p54nrb/NonO, bind to both hypophosphorylated and hyperphosphorylated forms of Pol II CTD when associated with RNA (Emili et al., 2002). U2AF65 was found to bound to a phosphorylated form of Pol II CTD and recruiting Prp19 complex, which is essential for co-transcriptional splicing (David et al., 2011). Prp40, a subunit of U1 snRNP, was also found to

associate to phosphorylated CTD, which facilitated U5 snRNP and Prp19 recruitment to the spliceosome during transcription (Gornemann et al., 2011; Morris and Greenleaf, 2000; Yuryev et al., 1996). More recently, immunoprecipitation of MNase-digested chromatin with antibodies against Pol II showed that both snRNA and proteins from active spliceosomes are complexed to phosphorylated Ser 5 CTD during co-transcriptional splicing (Nojima et al., 2018). Altogether, these evidences support a model capable of explaining how exons separated by long introns are recognized and correctly spliced by the spliceosome. This model suggests that the 5′ss is kept associated to Pol II CTD by U1 snRNP and SR proteins, immediately after being transcribed, and remain tethered there until the Pol II transcribes the 3′ss, which then interacts with U2AF65 and the Prp19 complex. Therefore, the proximity of these complexes facilitates the interaction of the following splicing components (Dye et al., 2006; Hollander et al., 2016; Zeng and Berget, 2000) ( Figure 1.3).

Co-transcriptional splicing can occur very efficiently after transcription of splice sites. By using live imaging to monitor the lifetime of the intron, from intron transcription to intron release using the RNA loops MS2 system in human cells, splicing rates could be as fast as 20-30s (Martin et al., 2013). However, a possible delay in loop formation, as well the release from spliceosome, can cause an overestimation of the real splicing rates. Supporting these results, similar times of 15-30s were observed for U2 and U5 snRNPs, associated with pre-mRNA (Huranová et al., 2010). From a different perspective, recent work taking advantage of long read sequencing from nascent chromatin RNA in yeast, showed that splicing could occur as soon as the 3′ss is exposed by the Pol II (Carrillo Oesterreich et al., 2016).

#### 1.3.2.1. Splicing regulation

Splicing is a powerful editor of the cell transcriptome. If on the one hand can originate more than 38.000 different transcript isoforms from the same original transcript (Neves et al., 2004) on the other hand, it can simply send it for rapid degradation (Lejeune and Maquat, 2005). This is typically a consequence of the way spliced sites are defined during splicing, which can result in exon skipping, intron retention, alternative 3′ and 5′ splice sites (Figure 1.4). The mechanisms that regulate splice site choice are complex and they can depend on many different factors.

### *cis* regulatory sequences

The first layer on splice site recognition are sequence elements that surround each splice site. While the 5′ splice site (5′ss) consensus is relatively short: AG/GUAUGU in yeast and AG/GURAGU in mammals (where R=purine, and / indicates a splice site), the 3′ splice site (3′ss) can extend up to 100 nucleotides into the intron. The 3′ss is defined by three separate sequence elements: the BPS composed UACUAAC in yeast and YNYURAC, in mammals, (where Y=pyrimidine, N=any nucleotide and A is the site of branch formation), the polypyrimidine tract, and the 3′ss consensus composed by CAG/G in yeast and YAG/G, in mammals (Moore et al., 1993). Although, the underlined sequences, that are ubiquitous to almost all splice site consensus, are required for splicing to occur, in many cases they are not sufficient to define the exons (Lim and Burge, 2001). Thus, auxiliary splicing regulatory elements (SREs) play important roles in splice site choice by recruiting specific trans-acting factors to the pre-mRNA that function by either activating or repressing splicing. This is possible due to their interaction with other regulatory factors or core components of the spliceosome. These elements can be found in introns, the so called intronic splicing enhancers (ISEs) or silencers (ISSs) or they can be also found in exons, also known as exonic splicing enhancers (ESEs) or silencers (ESSs). The last ones, can promote either, the exclusion or inclusion of exons or other type of alternative splicing (reviewed in (Matlin et al., 2005)).

### GC content

Another type of nucleotide composition that has also been observed to influence the way exons are defined, especially exons flanked by long introns, is the CG content. High CG content in exon compared with flanking introns, favours exon inclusion (Amit et al., 2012a). Although, not directly involved in the recruitment of splicing factors like SREs, CG-rich sequences are associated to nucleosome occupancy (Tillo and Hughes, 2009) which is a major determinant for exon recognition during splicing (Kornblihtt et al., 2009; Schwartz et al., 2009; Tilgner et al., 2009).

## Histone marks

Several epigenetic marks have also been found to influence splice site definition. For instance, DNA methylation (Gelfman et al., 2013), high levels of histone marks H3K36me3 (de Almeida et al., 2011) or hyper acetylation (Zhou et al., 2011) were found to be positively correlated with inclusion levels of alternative exons. These observations, the fact that Pol II has been found to accumulate at constitutive exons (Mayer et al., 2015), and that slow Pol II elongation rates promote the inclusion of alternative exons (de la Mata et al., 2003). This suggested a kinetic model, by which nucleosome and epigenetic marks regulate Pol II elongation rates, which can influence the way splicing occurs. A distinct non-mutually exclusive hypothesis suggests the recruitment of splicing factors by Pol II during transcription. For instance, phosphorylation of Pol II CTD is associated to splicing both, *in vitro* (Bird et al., 2004) and *in vivo*, especially phosphorylation of CTD Ser 5 (Nojima et al., 2015a), as it is known to recruit several splicing factors like U1 snRNP and U2Af65 (as described before). Moreover, chromatin markers can also facilitate recruitment of splicing factors. For example, U2 snRNP was shown to associate with chromatin via chromatin remodeler CHD1, which binds H3K4me3 in HeLa cells (Sims et al., 2007), and to heterochromatin protein 1 (HP1) that binds to H3K9me3, which in turn is associated with methylated DNA (Yearim et al., 2015). Additionally, BS69, a U5 snRNP component, recognizes the H3K36me3 modification and promotes intron retention (Guo et al., 2014). In summary, chromatin structure and Pol II elongation rates can influence splicing factors recruitment to the nascent RNA, being crucial for the regulation of alternative splicing (de Almeida and Carmo-Fonseca, 2012; Hollander et al., 2016; Iannone and Valcárcel, 2013).

## Spliceosome dynamics

The spliceosome has recently been characterized as a highly dynamic and flexible molecular complex, where the splicing reaction is reversible for each reaction (Hoskins and Moore, 2012), it can follow alternative splicing pathways (Shcherbakova et al., 2013) and shows a highly disordered composition (Korneta and Bujnicki, 2012). Since knockdown of many core subunits of the spliceosome did not impair splicing in general, but only had an effect on alternative splicing (Papasaikas et al., 2015), this suggests that the factors involved in

catalytic activation of the spliceosome are also likely to display important regulatory properties on splicing (reviewed in (Papasaikas and Valcárcel, 2016)).

### Gene architecture

Another important factor is the size of intron and exons, that have also been shown to influence splice site recognition (Sterner et al., 1996). For instance, in human cells length expansion of short exons to over 300 nucleotides flanked by long introns, strongly inhibit the formation of the spliceosome (Robberson et al., 1990) at least in some cases (Chen and Chasin, 1994). Moreover, it was also shown the importance of 5′ss recognition by U1 snRNP for the splicing of the upstream intron (Robberson et al., 1990). These observations lead to the exon definition model, in which both splice sites of exons are recognized by all five spliceosomal snRNPs before splicing of upstream intron (Schneider et al., 2010). Exon definition is characteristic of high eukaryotic species containing short exons flanked by long introns, contrasts with the situation in low eukaryote species, that are characterized by having short introns. In these cases, the 5′ss was observed to be dispensable for splicing of upstream short introns, thus suggesting that the intron is defined between their splice sites (intron definition). Interestingly, increasing the length of that intron in absence of a stronger 3′ss (no pyrimidine track) promotes aberrant splicing, showing that there is an optimal intron length for this to occur ((Talerico and Berget, 1994) and reviewed in (Berget, 1995)).

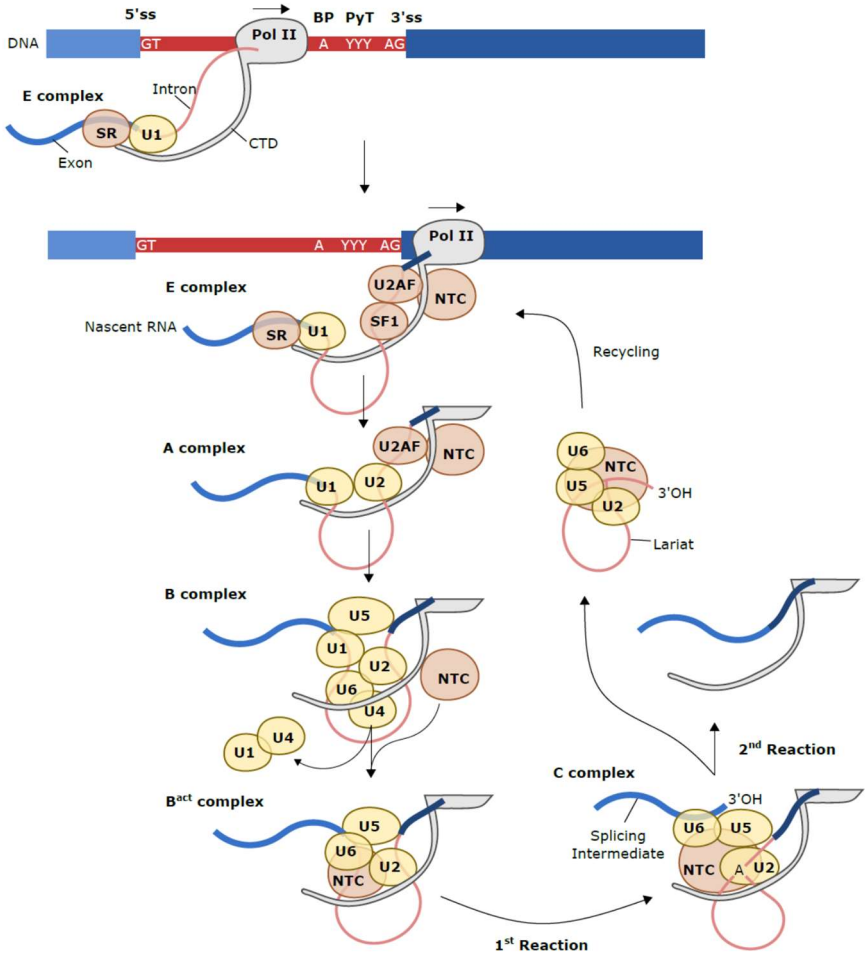



**Figure 1.3. Representation of co-transcriptional spliceosome assembly and splicing reactions.** Protein complexes (pink) and snRNPs (yellow) are recruited during transcription to the Pol II CTD (grey). DNA (thick lines), RNA (thin lines), introns (red) and exons (red).

Rather than length, intron position along the gene body has also been implicated in splicing efficiency. There are evidences that Cap-Binding Complex facilitates the association of U1sRNP, thus increasing splicing efficiency of 5′ proximal introns (Lewis et al., 1996). Accordingly, both mouse and human first introns are more efficiently spliced compared with downstream introns (Khodor et al., 2012a; Pandya-Jones and Black, 2009). However, these results are in contrast to what has been observed in *Drosophila*, in which the first introns are less efficiently spliced compared with other introns. One possible explanation, is the fact that U1 snRNP, which is recruited during transcription initiation, will accumulate several functions, such as: recruitment of initiation factors and preventing premature cleavage and polyadenylation, thus interfering with efficient co-transcriptional splicing of the first intron (Khodor et al., 2011b, 2012a).

Moreover, cleavage and polyadenylation factors are also known to play a role in exon-definition of last exons. Meaning that splicing of the last intron occurs preferentially during the 3′end processing (Li et al., 2001; Niwa and Berget, 1991; Rigo and Martinson, 2008). Accordingly, splicing of terminal introns in both mammals and flies are less efficient compared with other introns (Khodor et al., 2012a).

#### 1.3.2.2. Non canonical types of splicing

Besides the canonical splicing performed, non-canonical splice sites or other sequences can also alter the way how introns and exons are recognized, resulting in an atypical inclusion of cryptic exons, ligation of distant located exons, partial spliced introns, formation of circular RNAs, or even splicing in absence of the spliceosome.

##### Self splicing

Thought to be the ancestral mechanism of pre-mRNA splicing, self-splicing does not use spliceosome to catalyse the reaction. Instead, the intronic RNA, have the ability to fold and form secondary structures with ribozyme activity that will carry self-splicing. There are different classes of self-spliced introns, based on the type of reaction. In both classes two transterification reactions are performed, being the class II the most similar to the spliceosome pre-mRNA splicing. This type of splicing can be helped by other protein that is normally found only in bacteria, mitochondria, chloroplasts and the nuclei of some eukaryotic microorganisms (Nielsen and Johansen, 2009; Smathers and Robart, 2019).

##### Non canonical

Besides the consensus motifs of canonical splice sites, genes may contain many other putative cryptic splice sites that can be recognized by the spliceosome. Mechanisms like exon definition have been proposed to prevent uncontrolled splicing at these sites, thus maintaining the splicing fidelity (De Conti et al., 2013; Robberson et al., 1990). Although, some cases can

result in aberrant splicing such as inclusion of cryptic exons, in other cases these splice sites have been shown to be functional relevant for gene expression like recursive slicing (Kelly et al., 2015). In the case of cryptic exons, these often contains premature termination codons (PTCs) that when included in the transcript can target this transcript for nonsense-mediated decay (NMD) (Ni et al., 2007). In other cases, abnormal splicing can lead to NMD independent transcription-coupled surveillance mechanisms, that results in the reduction in expression of resulting transcripts (Vaz-Drago et al., 2015).

### Recursive Splicing

Recursive splicing is an evolutionarily conserved process of removing long introns via multiple steps of splicing. This is possible because recursive splice (RS) sites are defined by a sequence that combines the 3′ splice site immediately followed 5′ splice site consensus motifs, also known as zero length exons, within the long intron (Duff et al., 2015). The RS 3′ splice site is then used as an acceptor to splice the first part of the intron, which will expose the RS 5′ splice site to be used as donor to splice the next or last segment of the intron (Figure 1.4). RS sites have been reported in many different species (Hatton et al., 1998a), and are more frequently found in longer introns (Joseph et al., 2018a). Recursive splicing enhances efficiency of long introns splicing (Pai et al., 2018a) (Kelly et al., 2015). Interestingly, and although RS sites contains a 5′splice site, it has also been verified the existence of downstream splice sites that define cryptic exons. Thus, supporting the idea that recursive splicing follows the exon definition (Joseph et al., 2018a; Sibley et al., 2015).

### *cis* and *trans* splicing

It is also possible to observe splicing reactions that enable the ligation of exons from different, is some cases very distantly located, genes. Cis-splicing has been proposed to result from transcriptional readthrough, in which proximal genes are transcribed as a single unit. Thus, resulting in splicing of the penultimate exon of the upstream gene to the second exon of the downstream gene (Grosso et al., 2015)(Figure 1.4). By contrast, trans-splicing joins exons derived from distantly located genomic locations (Figure 1.4). The resulting chimeric

transcripts have been best documented in trypanosomes, *C. elegans* and insects (Allen et al., 2011; McManus et al., 2010; Sutton and Boothroyd, 1986), and to a lesser extent, in humans (Wu et al., 2014).

**Circular splicing**

Not all pre-mRNA splicing follows the canonical 5′ to 3′ order. Circular RNAs formation results from regulated mechanisms such as back splicing, also known as head-to-tail splicing, in which a branch point upstream of an exon attacks a downstream splice donor (Figure 1.4), or alternatively, from intron lariats that are resistant to de-branching owing to the presence of C-rich motifs near the branch point (Chen, 2016). In the first case, introns flanking an exon are brought together by hybridization of intronic Alu sequences (Jeck et al., 2013), or regulated by the binding of RNA binding proteins (Kramer et al., 2015). Depending on the splice site used, these RNA species can then result in single or multi-exons circularized RNAs, in which the multi exon can harbour or not the intron. Many circular RNAs have tissue-specific expression patterns, which suggests an important regulatory role (Chen, 2016; Memczak et al., 2013). In fact, different functions have already been suggested, like behaving as "sponges" for miRNA sponge (Hansen et al., 2013) and RNA binding protein (Ashwal-Fluss et al., 2014), among others (Barrett and Salzman, 2016).

**1.3.2.3. Intron functions**

Although introns are removed from the mRNAs during splicing, and therefore do not encode for proteins, they show however a high degree of conservation in many of their positions and sequences, indicating that these sequences play important functions within eukaryotic cells (Hare, 2003; Mattick and Gagen, 2001).

How introns emerged and evolved into the actual gene architectures of different species is still not consensual. However, evidences suggest that the last eukaryotic common ancestor (LECA) and early eukaryotes were relatively intron rich (Csuros et al., 2008, 2011). From this perspective, a progressive loss and/or gain of introns is thought for being responsible for shaping gene architecture of the present eukaryotic species. While in eukaryotic unicellular

organisms just few introns are typically found in the entire genome, complex multicellular organisms like mammals are classified as intron-rich (reviewed in (Rogozin et al., 2012)). Although some unicellular organisms, like Schizosaccharomyces pombe (fission yeast), have intron-dense genes, compared to budding yeast, and retained many features of complex splicing observed in mammals, it is consensual that introns were one of the major contributors for the evolution of complex multicellular eukaryotes, due to their ability to diversify both, the type of transcripts produced and their regulation.

**Alternative splicing**

One of the most significant advantages for genes to have many introns, is the ability to generate many different transcripts isoforms from the same DNA encoded gene by alternative splicing. Through the action of trans-acting factors directed to the precursor mRNA cis-acting regulatory elements, as well as through other factors referred before (e.g. Pol II elongation rates), alternative choice of splice sites can result is distinct patterns of alternative splicing. Alternative splicing includes inclusion or exclusion of exons, alternative 5′ and 3′ splice sites, intron retention and alternative splicing coupled with alternative first or last exons (Figure 1.4). Alternative splicing therefore greatly increases transcriptomic and proteomic diversity. One remarkable example is the *Down syndrome cell adhesion molecule 1* (Dscam1) gene of *Drosophila*, which potentially generates more than 38,000 transcript isoforms (Schmucker et al., 2000).

In human, near 95% of multi-exon genes undergo some degree of alternative splicing, mostly in a very tissue-specific way (Pan et al., 2008), and interestingly with high frequency in complex tissues, like the brain. Genetic variants that modulate alternative splicing influence phenotypic variability and disease susceptibility in human populations (Park et al., 2018).

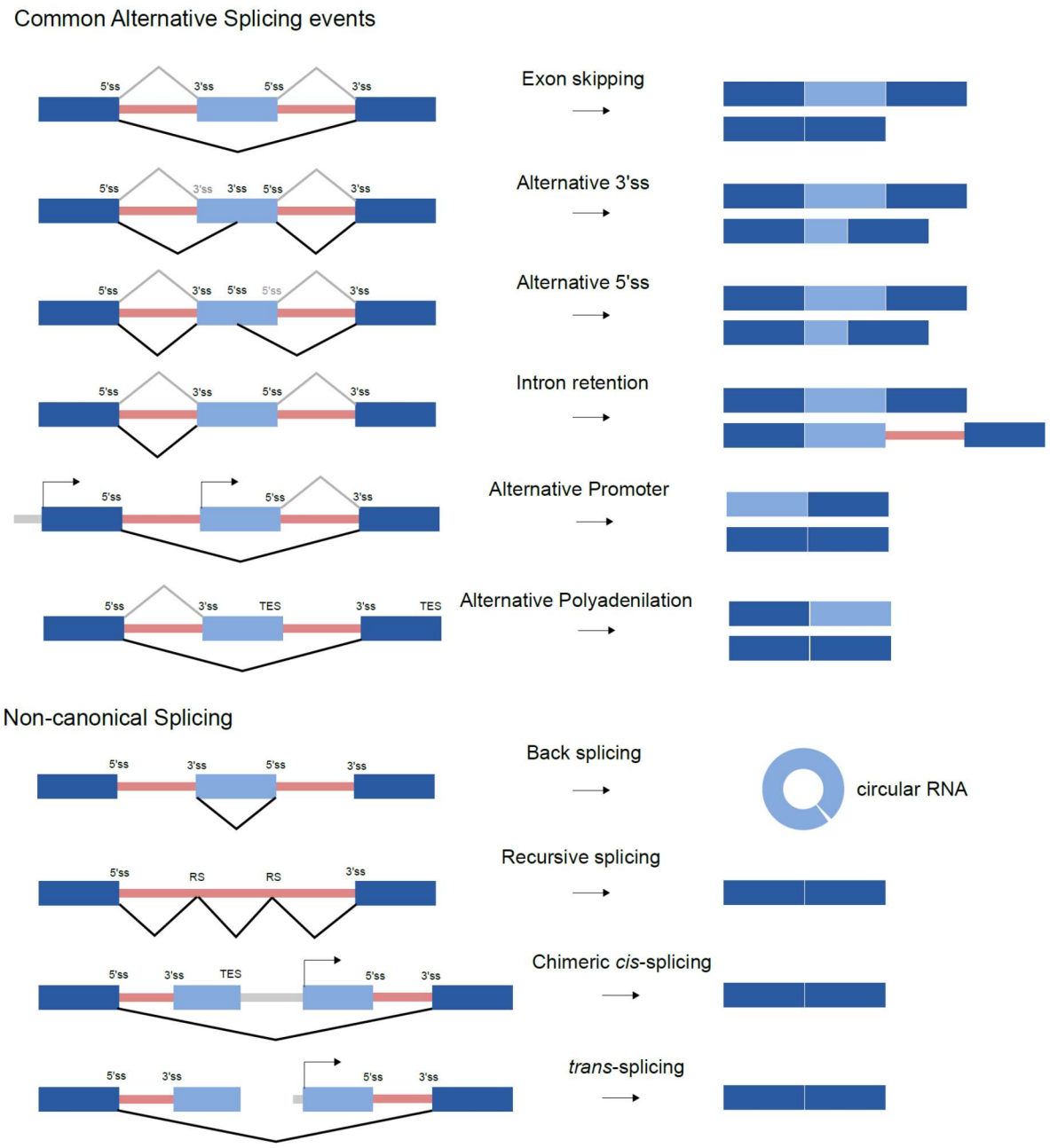



**Figure 1.4. Schematic representation of general Alternative and Non-Canonical Splicing events.** The black and grey lines indicate the alternative splicing events in each represented process. Genomic gene architecture (left), depicting the 5′ splice site (5′ss), the 3′ splice site (3′ss), the recursive sites (RS), the transcription start site (arrow), the transcription end site (TES), the introns (red lines), the exons (blue boxes), and the intergenic regions (grey line). Expected transcript configuration (right).

## Gene expression regulation

Total mRNA levels present in the cell are determined by the balance between transcription and mRNA degradation. Curiously, introns play fundamental role in both processes. For instance, the first intron of the shrunken-1 (Sh1) locus in maize, was shown to increase expression at least 10 times more efficiently than other maize introns when incorporated in another genes (Vasil et al., 1989). In agreement, removal of promoter proximal introns from endogenous genes markedly reduces transcription (Furger, 2002). Accordingly, posterior work showed that deposition of histone marks like H3K4me3 and H3K9ac, associated to gene activity, were dependent on the splicing of the first introns in human cells (Bieberstein et al., 2012), establishing a link between splicing, epigenetics and transcription. Curiously, besides the general increase in gene transcription, introns also seem to enhance the transcription rate (Fukaya et al., 2017).

In the opposite scenario, retention of introns containing premature termination codons (PTCs) leads to a surveillance mechanism that selectively degrades these mRNAs called nonsense-mediated decay (NMD) (Lejeune and Maquat, 2005). One interesting example that takes advantage of this mechanism to regulate the expression of 86 genes, can be observed in granulocytes differentiation, where intron retention is associated to the downregulation of splicing factors and increased levels of RNA degradation of those genes by NMD (Wong et al., 2013).

## Nuclear Export and localization

Splicing is also involved in efficient nuclear export of mRNAs. Work done in human culture cells, showed that intron splicing enhances 6 to 10-fold the efficiency of mRNA export (Valencia et al., 2008). In yeast, the TREX multi-complex is important for mRNA nuclear export (Fischer, 2002; Jimeno, 2002). Interestingly, and contrary to yeast, in human cells the recruitment of TREX to mRNA is dependent on splicing (Masuda, 2005).

### Timing of transcription

Intron length can have important roles in the regulation of gene expression. For instance, transcription of the 6o kb long E74A gene takes about 60 min to be expressed after induction by Ecdysone. This expression delay is critical for gene expression coordination during the onset of *Drosophila* metamorphosis (Karim and Thummel, 1991; Thummel et al., 1990). Yet, since the mRNA has only 6 kb, this suggests that it is the large size of its intron that underlies this delay. Likewise, *hairy* and *enhancer of split 7* genes (*Hes7*), contain a 1.8-kb intron that causes a 19-minute delay in protein expression. *Hes7* regulates itself by a negative-feedback loop, and this splicing-dependent delay causes its expression to oscillate, which is an important mechanism for the regulation of somite segmentation in the mouse embryo (Takashima et al., 2011). Interestingly, several genes with related developmental processes, share similar intron lengths (Seoighe and Korir, 2011; Swinburne and Silver, 2008).

### Genome instability prevention

During transcription, nascent mRNAs are known to potentially form stable hybrid structures with the template DNA, the so-called R-loops. These structures are a source of genetic instability, since they can trigger expression defects, double strand breaks or unwanted recombination events (Chan et al., 2014; Santos-Pereira and Aguilera, 2015). Interestingly, introns appear to play an important role preventing the formation of these structures. While deletion of endogenous introns increases R-loop formation, insertion of an intron into an intronless gene suppresses R-loop accumulation, in a process that depends on spliceosome recruitment, but not through splicing *per se* (Bonnet et al., 2017).

## 1.3.3. Cleavage and polyadenylation

Cleavage and polyadenylation are processes that occur, in most of cases, after the Pol II completes the synthesis of the transcription unit, thus resulting in a mRNA released from nascent RNA with a processed 3′end carrying a Poly (A) tail.

Unlike yeast, that uses two different mechanism, metazoan pre-mRNA 3′ end processing is performed by the Cleavage and Polyadenylation (CPA) complex, which is composed by the cleavage and polyadenylation specificity factor (CPSF), cleavage stimulation factor (CSTF) and cleavage factors I and II (CFIm and CFIIm) subcomplexes (Shi and Manley, 2015). In a process that is coupled to transcription, specific sequences in the nascent RNA molecule are recognized by different components of CPA. These sequences are composed by a central motif AAUAAA (or less frequently AUUAAA) polyadenylation (pA) site or (PAS), flanked by auxiliary elements: a U- or GU- rich downstream sequence element (DSE) and a U-rich upstream sequence element (USE) (Proudfoot, 1976, 2011). This way, CPFS73 a subunit of CPSF performs the catalysis of RNA between the pA and DSE, at CA dinucleotide. The resulted upstream fragment is then processed by other CPSF subunits: the nuclear poly(A) binding protein (PABPN) and the poly(A) polymerase (PAP), that will add a poly-A tail (that can vary between 50 and 100nt) (Chang et al., 2014) (Figure 1.5). This process ensures the RNA protection from 3′-5′ degradation and guarantees the proper export from the nucleus and translation of this transcripts in the cytoplasm.

Exceptionally, some small nuclear RNAs, like spliceosome snRNAs, and histone mRNAs 3′-end cleavage process is performed by the Integrator complex using the catalytic RNA endonuclease activity of IntS9/IntS11 hetero-dimer subunits, rather than CPA complex (Baillat and Wagner, 2015; Rienzo and Casamassimi, 2016). In these cases, a conserved sequenced ($GTTTN_{0\text{-}3}AAARNNAGA$) called 3′-box and located 9–19 nucleotides downstream of their 3′-end is recognized by the complex (Hernandez, 1985).

#### 1.3.3.1. Alternative and premature polyadenylation

Besides their function as locals to promote the cleavage and polyadenylation, 3′ untranslated regions (3′UTRs) are also involved in mRNA post-transcriptional regulation processes, that includes: localization and translation of this molecule. Thus, alterations in their length are expected to add an extra layer of gene expression regulation. Accordingly, genome wide studies have shown that over 70% of mammalian genes have more than one poly(A) sites (PAS) (Derti et al., 2012), and a similar number was also found in *Drosophila* 65% (Sanfilippo et al., 2017). Through a process denominated alternative polyadenylation (APA), RNAs result

in distinct 3′ termini, generating different transcript isoforms with distinct functions in a great extent and tissue-specific manner (reviewed in (Tian and Manley, 2017)).

Cleavage and polyadenylation are not restricted to 3′UTR; these processes can also occur in intragenic regions, normally in introns, or even next to the TSS, thus generating truncated proteins. Deregulation of this process is, in many cases, associated to different types of cancers (Yuan et al., 2019). However, in some cases this type of events appear to be essential since they regulate the general levels of gene expression. This effect, was observed over 400 *Drosophila* protein coding genes that showed evidences of premature termination, in a process regulated by the Integrator complex (Tatomer et al., 2019).

On the other side, mechanisms such as telescripting, suppress premature cleavage and polyadenylation events in eukaryotic cells. This process is known to involve the spliceosomal U1 snRNP complex that competes with 3′ termination factors by the intronic PASs, thus repressing premature termination (Berg et al., 2012; Kaida et al., 2010; Venters et al., 2019).

Different mechanisms have been shown to specifically regulate APA. For instance, modulation of the expression levels of CPA components increased usage of the weaker upstream intronic PAS, as it was shown in B cells differentiation (Takagaki et al., 1996). Additional mechanisms involve the regulation of PAS usage through the interaction with other RNA binding proteins. For example, *embryonic lethal abnormal vision (elav)* was shown to mediate neuron-specific 3′ UTR lengthening by suppressing the use of proximal PASs in *Drosophila* (Hilgers et al., 2012).

### 1.4. Transcription Termination

Transcription termination is a process in which, after the Pol II completes the synthesis of the transcription unit, the Pol II molecule is released from DNA template, in order to prevents transcriptional interference between adjacent transcriptional units (Figure 1.5)(Iasillo et al., 2017; Porrua et al., 2016; Proudfoot, 2016) and ensuring as well the recycling of Pol II.

Interactions between upstream RNA terminator sequences, the 3′end processing factors which in your turn are coupled to CTD of elongating Pol II, are known to cause the pause of Pol II over the termination region (Glover-Cutter et al., 2008; Gromak et al., 2006). Although, this is important to help termination (Birse, 1997; Dye and Proudfoot, 2001), this pause is not

sufficient to release the Pol II from the DNA template (Kuehner et al., 2011). Still under discussion in the field, Pol II release from DNA is been supported by two non-mutually exclusive models. The first model states a conformational change of Pol II caused by the 3′end processing factors, resulting on the release of Pol II (allosteric model). The second model proposes that the nascent RNA still bound to Pol II after cleavage, is degraded of by the 5′-3′ exonuclease XRN2, which "pursuits" the Pol II this being sufficient to cause the release of this molecule from DNA template and consequent termination (torpedo model) (Figure 1.5) (Porrua et al., 2016).

Even so, Pol II is being detected in many cases to continually transcribe thousands of base pairs downstream of the PAS (Nojima et al., 2015b; Schwalb et al., 2016). Defects in transcription termination (Kuehner et al., 2011) or alterations in chromatin marks (Grosso et al., 2015) can lead to pervasive transcription, which can lead in many cases to abnormalities in the transcription of downstream locate genes. This effect was proposed to result from Pol II transcription collision between convergent genes (Prescott and Proudfoot, 2002) or interference between tandem genes (Greger, 1998). This raises the question whether this "normal" detected pervasiveness of Poll II may play a second role downstream in transcription or even in other processes.

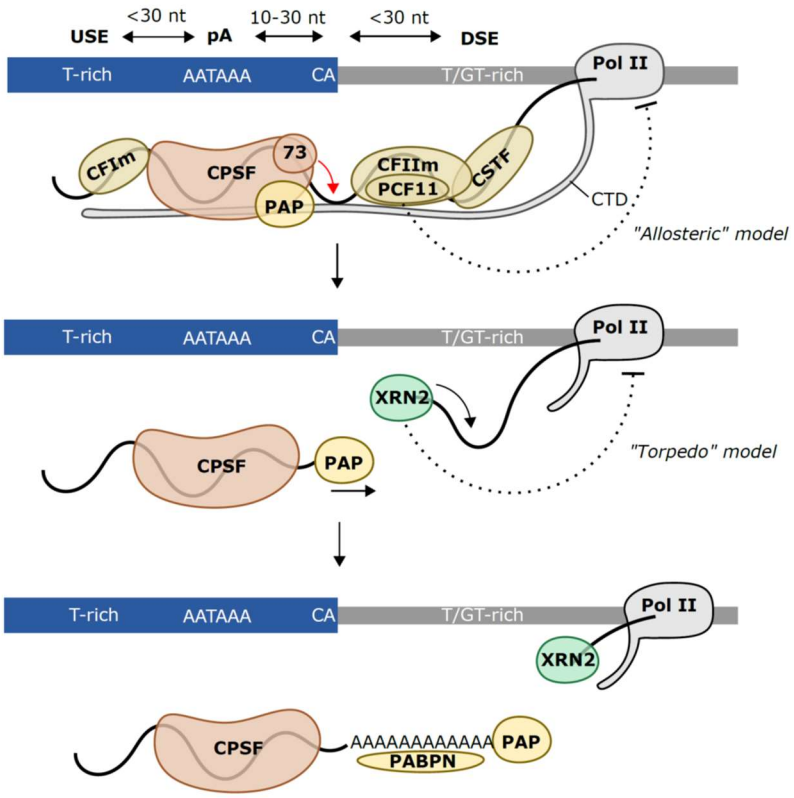



**Figure 1.5. General mechanism of cleavage, polyadenylation and transcription termination in eukaryotes.** The diagram illustrates the assembly and function of CPA complex. Specific sequences on nascent RNA are recognized by the different sub-complexes: CFIm, CFIIm, CPSF and CSTF in a processed coupled to transcription through the interaction with the Pol II CTD. After cleavage by the subunit CPSF73 (73) of CPSF complex, (brown), PAP add the poly (A) tail while PABPN regulates the length (yellow). Current Pol II release models: Pol II conformation change by CPA components "Allosteric model" and Pol II destabilization through nascent RNA degradation by XRN2 "Torpedo model".

## 1.5. *Drosophila* as model system to study co-transcriptional process

With a genome comprised of only 4 pair of chromosomes, sharing many homologous genes with vertebrates, with a short life cycle (10 days) and easy to maintain in the lab, *Drosophila melanogaster* has become one of the most used model systems to study many biological processes. In particular, *Drosophila* embryogenesis has been one of the most widely studied developmental processes, been crucial to understand the mechanisms of gene expression regulation during development.

### 1.5.1. Early *Drosophila* embryonic development

After fertilization, *Drosophila* embryonic development undergoes 13 nuclear divisions without cytokinesis (syncytial blastoderm). The first 9 nuclei divisions occur synchronously, with mitotic cycles alternating between S-phase and M-phase in just ~9 min each. This divisions, considered among the fastest known for any animal embryonic system (Foe and Alberts, 1983), occur under the control of maternally deposited Cyclins, Cdk1 and the tyrosine phosphatase Cdc25/Twine, that ensures the rapid mitotic cycles (Edgar et al., 1994). After cycle 9, most nuclei start to migrate to the periphery of the embryo, and by cycle 10 the migrating nuclei become evenly distributed in a monolayer under the egg surface, forming the so called syncytial blastoderm (Foe and Alberts, 1983). From mitotic division 10 to 13, cycles take place almost synchronously, and start to slow down, taking ~21 min during cycle 13 and 65 min at interphase 14. During this last interphase, cell membrane invaginates around each peripheric nuclei to form a monolayer of cells in a process known as blastoderm cellularization.

Mid-blastula transition (MBT) coincides with a general activation of the zygotic genome activation and degradation of many maternal provided products in a process also known as maternal to zygotic transition (MZT) (Tadros and Lipshitz, 2009) (Figure 1.6). Increased lengthening of the early embryo cell cycles is essential for early zygotic expression as it provides time for zygotic genes to be transcribed. In this context, it has been proposed that activation of zygotic transcription plays an essential role by activating the replication check point, since transcription during DNA replication generates physiological levels of DNA damage due to conflicts between the DNA and RNA polymerase complexes (Blythe and Wieschaus, 2015). This activates the DNA replication checkpoint, which drives Chk1-dependent downregulation of Cdc25 catalytic activity (Edgar and Datar, 1996; Peng, 1997), leading to attenuation of Cdk1 kinase activity, which results in cell-cycle lengthening. Expression of several zygotic genes during interphase 14 dictates the degradation of Twine (Di Talia et al., 2013; Farrell and O'Farrell, 2013), thus resulting on downregulation of Cdk1 activity, which results in the inclusion of a gap phase (G2) during this cell cycle (Farrell et al., 2012) (Figure 1.6).

### 1.5.2. Blastoderm cell fate determination

At the end of interphase 14, the approximately 6000 cells that form the cellular blastoderm are already committed to specific developmental fate (Foe, 1989), and at this stage the embryo starts a series of morphogenetic movements (gastrulation) giving rise to the multi-layered body of the larva (Campos-Ortega and Hartenstein, 1985). This cell fate starts to be orchestrated still during mid-oogenesis when mRNAs, among others, like *bicoid (bcd)* become localized to the anterior, and *oskar (osk)* and *nanos (nos)* to the posterior poles of the oocyte. The consequent translation of *bcd* and *nos,* the so-called maternal effector genes, after fertilization, creates a protein gradient that will regulate the expression of zygotically expressed gap genes like: *knirps (kni), Kruppel (Kr) or giant (gt)*. These will in turn, control the expression of pair-rule genes like: *even-skyped (eve) or fushi-tarazu (ftz)*, through an elaborated network of transcription factors acting as enhancers or repressors, that will define the segments of the *Drosophila* embryo cellular blastoderm (review in (Jaeger, 2011; Lasko, 2012)) (Figure 1.6).

### 1.5.3. *Drosophila* zygotic genome activation

Although the large majority of genes starts to be transcribed during interphase 14 (MBT), there is however a first wave of zygotic genes (pre-MBT) that are already expressed prior to that MBT stage, as early as mitotic cycle 8, (Erickson and Cline, 1993; Pritchard and Schubiger, 1996) (Figure 1.6). Actually, there is evidence that very low levels of transcription can potentially occur even earlier (Ali-Murthy et al., 2013a; ten Bosch et al., 2006). Accordingly, a recent study identified 20 genes to be expressed between nuclei divisions 7 and 9, which gradually increases to 63 during divisions 9 and 10 and that reaches 946 before the end of cycle 13 (Kwasnieski et al., 2019a). In accordance, an intermediary number of genes are known to be transcriptionally active before the end of cycle 12 (Chen et al., 2013). These include, among others, nuclear genes required for sex determination, pattern formation, non-coding RNAs and for blastoderm cellularization (Edgar, 1986; Erickson and Cline, 1993; Lécuyer et al., 2007; Pritchard and Schubiger, 1996). This class of pre-MBT genes show special characteristics compared to late expressed genes. Beside the fact that they are small and intronless (see below), in general, these genes lack the promoter proximal Pol II pausing. This

phenomenon correlates with the fact that most early genes share promoter sequences for the transcription factor Zelda (Chen et al., 2013; Saunders et al., 2013). Zelda has been described as a pioneer transcription factor that binds to enhancers, establishing or maintaining the chromatin open to facilitate the recruitment of Pol II, and thus contributing for the activation of this class of genes (Harrison et al., 2011).

Zygotic totipotent cells require a reprograming process to erase all the epigenetic marks associated to differentiated cell state of the egg or sperm. Thus, like it happens in many other species, *Drosophila* early embryonic chromatin shows a naive state, characterized by the lack of histone marks before the activation of zygotic genome. In a sequential way, acetylated histone residues like H4K8, H3K18, and H3K27 are the first to be observed and associated to pre-MBT chromatin, whereas methylation of H3K4 and H3K36 show up later during MBT (Chen et al., 2013; Li et al., 2014). Besides their marks, the concentration of histone was also shown to be a key player in the regulation of genes whose expression is associated to Zelda transcription factor. Consistently, variation in the concentration of histones affects the length and number of the mitotic cycles (Chari et al., 2019). Curiously, and unlike the majority of species where DNA methylation plays an important role in gene expression regulation, it was shown that *Drosophila* lacks DNA methylation patterns (Raddatz et al., 2013).

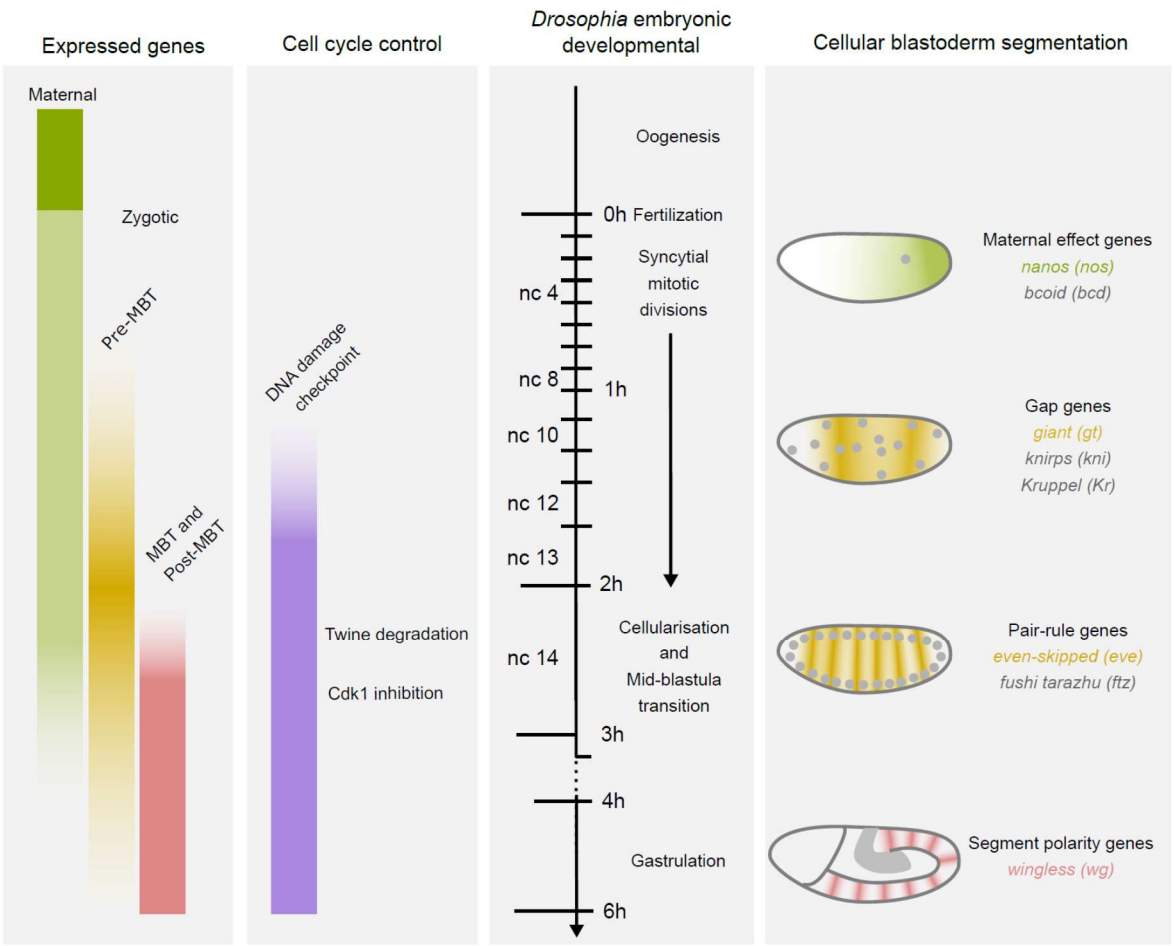



**Figure 1.6. Schematic representation of *Drosophila's* early embryo development main events.** The diagram illustrates the temporal expression of maternal, pre-MBT, MBT and post-MBT genes during *Drosophila* embryonic development (left panel). Maternal genes are transcribed during oogenesis (green) and the resulting mRNAs persist during early embryogenesis (light green). Zygotic transcription starts before mid-blastula transition (pre-MBT genes, yellow). Most genes become transcriptionally active during or after mid-blastula transition (MBT and post-MBT, red). Main events responsible for the regulation of syncytial mitotic cycles length. DNA damage checkpoint response (purple) and Twine degradation followed by Cdk1 inhibition during interphase 14 (second panel). Chronology of the main events occurring during *Drosophila* embryo development. Time and correspondent nuclei cycle (nc) are shown (third panel). Example of genes contributing for the segmentation patterns of *Drosophila's* embryo and their spatial localization in the embryo: maternal effect gene *nanos* (nos) (green); gap gene *giant* (gt) (yellow); pair-rule gene *even-skypped* (eve) (yellow) and segment polarity gene *wingless* (wg) (red).

### 1.5.4. Gene expression constraints and gene architecture in *Drosophila*

Since *Drosophila* early embryo transcription is restricted to interphases (Shermoen and O'Farrell, 1991) and the rates of transcriptional elongation can be as fast as 2.4-3.0 kb per minute (Fukaya et al., 2017), the short length of the syncytial interphases during pre-MBT embryo impose significant constraints to pre-MBT transcription. In fact, transcriptional elongation of long genes can be, either prematurely terminated (McKnight and Miller Jr., 1976; Rothe et al., 1992; Sandler et al., 2018; Shermoen and O'Farrell, 1991) or aborted (Kwasnieski et al., 2019a). This is consistent with the fact that pre-MBT genes are on average shorter than the overall genes expressed at later stages of development (Artieri and Fraser, 2014; Hoskins et al., 2011). Curiously, 70% of this pre-MBT genes are intronless, in contrast to the 20% of intronless genes represented in all genes (De Renzis et al., 2007a). Suggesting that like in transcription, splicing seems to be avoided by pre-MBT genes, a comparison of similarly size genes clearly showed a bias for intronless genes among the pre-MBT dataset (Martinho et al., 2015). This idea was further supported by work developed in our lab ((Guilgur et al., 2014); and presented in the chapter 3.1 of this thesis), and recently reinforced by the fact that, during this particular developmental stage it was observed higher levels of intron retention (Kwasnieski et al., 2019a).

Human genes present a characteristic gene architecture containing in average 4 long introns (3413 bp, on average) flanked by short length exons (51 bp, on average) (Long and Deutsch, 1999), in which around 95% of humans genes show evidence of alternative splicing (Pan et al., 2008; Wang et al., 2008). Although, *Drosophila* has less (2.5, on average) and shorter (564 bp, on average) introns, and longer flanking exons with 141 bp (on average), compared with human, the gene architecture is significantly more diverse, containing many genes with longer introns and shorter exons (Long and Deutsch, 1999). Moreover, *Drosophila* also shows a high degree of alternative splicing, especially in germ cells, muscle and the central nervous system. This clearly indicates that *Drosophila* is not only is a good model system to study alternative splicing, but also to understand how is splicing regulated by gene architecture.

## 1.6 Transcriptomic approaches

In the last twenty years, the development of distinct ground-breaking transcriptomics approaches has unravelled at genome-wide level a complex molecular network responsible for the regulation of gene expression during cell differentiation.

### 1.6.1. Transcript identification and quantification

The RNA field emerged in the decades of 70s and 80s, mainly thanks to the possibility to synthesise artificial oligonucleotides and use isolated thermostable DNA polymerase I and Reverse transcriptase enzymes. Among others, these factors contributed for the development of some of the basic molecular biology techniques still used today in the field. Some examples are: quantification of DNA and RNAs, like Southern and Northern blots, respectively (Alwine et al., 1977; Southern, 1975), as well as the amplification (polymerase chain reaction (PCR)) (Fey et al., 1991) and sequencing (Sanger method) (Sanger et al., 1977), of both DNA and cDNA generated from mRNAs. PCR is still one of the most used techniques, serving as a base for many other techniques given its unique ability to amplify with high specificity small amounts of DNA. Later, PCR became a more quantitative technique, with the development of two different procedures: the Quantitative real time PCR (qPCR), which through the monitorization of DNA amplification during the PCR reaction is able to quantify the relative abundance of the initial molecules analysed (Higuchi et al., 1993), and digital PCR (dPCR), that taking advantage of diluting the nucleic acid sample into numerous individual partitions (that are independently amplified), is able to perform an absolute quantification of the initial molecules (based on Poisson statistics) (Sykes et al., 1992). With the association of these detection techniques to immunoprecipitation of protein cross-linked to chromatin DNA (ChIP), for instance, it was possible to validate the presence of RNA Pol II over *Drosophila* genes (Gilmour and Lis, 1985), as well the presence of histones over the same transcriptionally active genes (Solomon et al., 1988).

In the same period, Nuclear Run-On assay, which consists on the arrest of transcription that is posteriorly resumed *in vitro* in the presence of radiolabelled nucleotides and anionic

detergent sarkosyl (that prevents the transcription re-initiation), allowed to assess the activity of transcriptionally engaged RNA Pol II (Gariglio et al., 1981). Through this technique is was possible to show the first evidence for promoter proximal pausing in *Drosophila Hsp70* gene (Rougvie and Lis, 1988).

### 1.6.2. Transcript visualization

Electron microscopy allowed the observation of the first actively transcribing genes with Miller's spreading technique in amphibian oocytes (Miller and Beatty, 1969). It was also possible to visualise the first evidences of splicing with high formamide spreading technique (Berget et al., 1977), and co-transcriptional splicing (Beyer and Osheim, 1988). By hybridizing oligonucleotide labelled probes to DNA or RNA in cells (*in situ* hybridization), it was possible to visualize not only the spatial localization of genes or produced transcripts, but also its expression pattern in tissues. This technique, greatly contributed to the better understanding of gene regulation pathways during development (Gall and Pardue, 1969; John et al., 1969). With the introduction of fluorescent detection *in situ* hybridization (FISH) technique, it became a more sensible procedure (Bauman et al., 1981), allowing the detection of single RNA molecules (Femino, 1998). Recently, a FISH based technique was able to identify 10,421 genes at their nascent transcription active sites in single cells (Shah et al., 2018).

However, almost all of these techniques used fixed cells, not allowing the observation of these processes live cell kinetics. By taking advantage of the ability of MS2 binding protein to bind stem-loop structures that contain MS2 binding sites, and by tagging a GFP to MS2 protein, it was possible to perform live cell imaging of mRNAs that containing the MS2 binding sites (Bertrand et al., 1998). By taking advantage of this imaging system, and by incorporating these stem-loop structures into introns, it was possible to measure splicing kinetics with single RNA molecule resolution in human beta-globin gene (Martin et al., 2013).

### 1.6.3. Transcriptomics

Although crucial to reveal the fundamental mechanisms of gene expression, the methodologies described before were typically applied to a small number of genes, lacking therefore a genome-wide perspective of the molecular mechanisms occurring within a cell.

This gap starts to be filled in the 90´s with the emergence of automated Sanger based techniques, like Serial Analysis of Gene Expression (SAGE) (Velculescu et al., 1995), and complementary probe hybridization based techniques like the Microarrays (Schena et al., 1995). Second-generation sequencing technologies started to sequence over 200,000 reads of 110 bp length, but rapidly evolved to sequence millions of reads with Illumina HiSeq technology, becoming nowadays the most popular technique to sequence cDNA from mRNAs (RNA-seq) (Figure 1.7A) (reviewed in (McGettigan, 2013)). Moreover, by sequencing both paired-end reads, it was possible to have higher levels of reads mappability and a better identification and mapping of spliced reads (Katz et al., 2010).

As expected, previous referred techniques, rapidly took advantage of the capabilities of high throughput sequencing. For instance, ChIP based sequencing (ChIP-seq) (Johnson et al., 2007), and Individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP), that allow to analyse protein-RNA interactions with nucleotide resolution (Huppertz et al., 2014), started to emerge. Methodologies to investigate newly synthesized RNA molecules involving metabolic labelling of nascent transcripts (Fuchs et al., 2015) include global run on sequencing (GRO-seq) (Figure 1.7B)(Core et al., 2008a), precision nuclear run on sequencing (PRO-seq) (Kwak and Lis, 2013), 4sUDBR-seq (Fuchs et al., 2014) and transient transcriptome sequencing (TT-seq) (Schwalb et al., 2016).

To overcome the use of cross-linking, to increase sequence resolution and avoid the limitations of *in vitro* labelling procedures, Native Elongating Transcript sequencing (NET-seq) (Figure 1.7C) was developed by Churchman and Weissman (Churchman and Weissman, 2011a). NET-seq monitors transcriptionally active Pol II with single nucleotide resolution by specifically sequencing the 3′-end of nascent RNA molecules bound to Pol II. Without the need of cross-link procedures, the Pol II-DNA-RNA ternary complex is solubilized from the chromatin with DNAse (Churchman and Weissman, 2011a) or Micrococcal nuclease (Nojima et al., 2015a), and consequently immunoprecipitated using antibodies for Pol II. After RNA purification, Pol II position can be identified by the last incorporated nucleotide. Using specific

antibodies for distinct phosphorylated isoforms of Pol II CTD, it was also possible to understand in more detail the link between CTD phosphorylation, Pol II position along the different stages of transcription, and distinct co-transcriptional processes, in humans (Nojima et al., 2015a, 2018), yeast (Harlen et al., 2016) and plants (Zhu et al., 2018).

More recently, and thanks to SMRT long read sequencing developed by Pacific Biosciences, that manages to sequence reads 15000bp length on average, chromatin RNA has been used to associated Pol II position along the gene, with co-transcriptional splicing in yeast (Carrillo Oesterreich et al., 2016). On the other hand, and using a completely different technology, Nanopore sequencing allows the direct sequencing of both DNA and RNA molecules with no need for PCR amplification, thus reducing any potential amplification bias (Clamer et al., 2014; Garalde et al., 2018). By taking advantage of this direct RNA sequencing approach, and 4sU labelling during transcription, it was also possible to correlate splicing with Pol II position (Drexler et al., 2020). However, and due the low coverage provided by these techniques, this type of analysis is either reduced to a few genes, or many genes with low read coverage per gene.

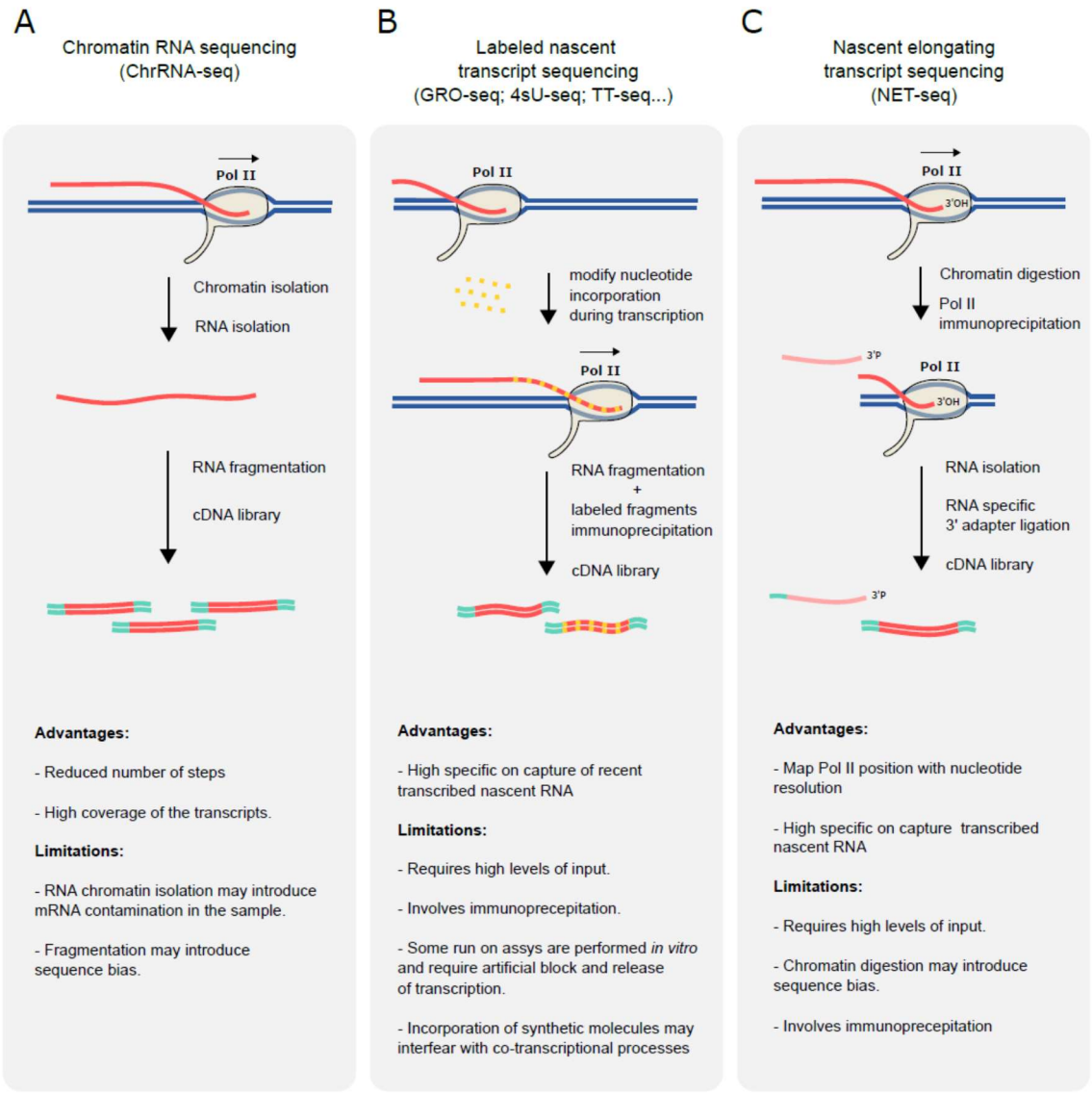



**Figure 1.7. General overview of transcriptomic sequencing methodologies to analyse nascent RNAs. (A)** Chromatin RNA sequencing. **(B)** Methodologies involving incorporation of modify nucleotides during transcription. RNAs containing the modify nucleotides are specifically capture by immunoprecipitation, which can occur before the fragmentation step (GRO-seq and 4sU-seq) or after (TT-seq). **(C)** Nascent elongating transcript sequencing (NET-seq) involves the digestion of chromatin followed by immunoprecipitation of solubilized Pol II-DNA RNA tricomplex. Nascent RNA bound to the Pol II is specifically ligated to the adapters. DNA (blue), RNA (red), modify nucleotides (yellow dots), adapters (green).

# 2. Objectives

The vast majority of introns are removed co-transcriptionally. However, how splicing and transcription processes influence each other and are influenced by other factors, is largely unknown. Having this in mind and since many of these questions are normally addressed *in vitro* or taking advantage of cell cultures, in this thesis out main goal was to understand **how transcription and splicing influence each other during development**. To answer this question, we decided to focus our analysis during *Drosophila* early embryonic development, where the fast nuclei divisions impose significant constrains on early zygotic transcription and gene architecture (Rothe et al., 1992).

Due to the high frequency of intronless genes, **we first hypothesized that there are significant constraints to splicing in the early embryo when compared to other developmental stages**. To test this hypothesis, we took advantage of a splicing factor mutant allele that specifically impair splicing during early embryonic development.

Next, we **hypothesized that co-transcriptional splicing efficiency varies during development**. In order to measure genome-wide splicing kinetics during development, we adapted native elongating transcript sequencing (NET-seq) to the developing Drosophila embryo, in order to correlate splicing with the position of RNA polymerase II (Pol II).

By taking advantage of NETseq and *Drosophila* diverse gene architecture, we **tested the hypothesis whether splicing tend to occur immediately after Pol II transcribes the 3′ss in *Drosophila* embryos,** like was previously observed in yeast (Carrillo Oesterreich et al., 2016). and **how gene architecture and other genomic features influence the moment of splicing during transcription and transcriptional termination**.

Although most early zygotic genes are intronless and constraints to splicing are likely to exist, a small subset of these genes however contains introns. Since early zygotic intron-containing genes frequently show a complex pattern of gene expression (e.g. patterning genes), and these introns can be evolutionary conserved, we **hypothesized that they are important for early embryo gene expression.** To test this hypothesis, we generated bacterial artificial chromosome (BAC)-based transgenic construct of even skipped (eve) with or without its intron, and investigated the ability of this construct to complement a eve loss-of-function mutant allele.

# 3. Results

## 3.1. Requirement for highly efficient pre-mRNA splicing during *Drosophila* early embryonic development

The results presented below were published in the eLife peer-reviewed journal in 2014. The article in the publication format can be found in the Appendix of this thesis.

Leonardo Gastón Guilgur [1,2,3], Pedro Prudêncio [1,2,3], Daniel Sobral [1], Denisa Liszekova [1], André Rosa [1], Rui Gonçalo Martinho [1,2,3].

[1] Instituto Gulbenkian de Ciência, Oeiras, Portugal;

[2] Departamento de Ciências Biomédicas e Medicina, Universidade do Algarve, Faro, Portugal

[3] IBB-Institute for Biotechnology and Bioengineering, Centro de Biomedicina Molecular e Estrutural, Universidade do Algarve, Faro, Portugal

**Author contribution**

Leonardo Gastón Guilgur, Pedro Prudêncio and Rui Gonçalo Martinho designed the experiments and wrote the manuscript. Leonardo Gastón Guilgur, Pedro Prudêncio performed most of the experiments and data analysis. Denisa Liszekova and André Rosa were involved in part of the mutant allele characterization. Daniel Sobral analysed the RNA-seq data. All authors revised part or the entire manuscript.

### 3.1.1. Overview

*Drosophila* syncytial nuclear divisions limit transcription unit size of early zygotic genes. As mitosis inhibits not only transcription, but also pre-mRNA splicing, we reasoned that constraints on splicing were likely to exist in the early embryo, being splicing avoidance a possible explanation why most early zygotic genes are intronless. We isolated two mutant alleles for a subunit of the NTC/Prp19 complexes, which specifically impaired pre-mRNA splicing of early zygotic but not maternally encoded transcripts. We hypothesized that the requirements for pre-mRNA splicing efficiency were likely to vary during development. Ectopic maternal expression of an early zygotic pre-mRNA was sufficient to suppress its splicing defects in the mutant background. Furthermore, a small early zygotic transcript with multiple introns was poorly spliced in wild-type embryos. Our findings demonstrate for the first time the existence of a developmental pre-requisite for highly efficient splicing during *Drosophila* early embryonic development and suggest in highly proliferative tissues a need for coordination between cell cycle and gene architecture to ensure correct gene expression and avoid abnormally processed transcripts.

### 3.1.2. *Drosophila* Fandango/Xab2 is required for blastoderm cellularization

Previously we isolated a collection of maternal mutants defective in blastoderm cellularization and/or germ-band extension (Pimenta-Marques et al., 2008). Complementation group 7 contained two different mutant alleles with similar defects in blastoderm cellularization. Through deficiency mapping and a candidate gene approach we concluded that both were allelic to the uncharacterized coding gene CG6197 (Flybase). To confirm the mutants' identity, we rescued their zygotic lethality, female sterility (germ-line clones), and blastoderm cellularization defects (maternal mutant embryos) using a genomic fragment construct that contained a wild-type copy of CG6197 (Figure 3.1—figure supplement 1A, data not shown). Both isolated alleles of CG6197 showed identical phenotypes: maternal mutant embryos (hereafter referred to as mutant embryos) showed normal syncytial nuclear divisions (Figure 3.1A,B) but subsequently failed to elongate the cortical nuclei, which became mislocalized during blastoderm cellularization (Figure 3.1C–F, quantification in Figure 3.1G). The blastoderm cellularization phenotype was remarkably similar to that described for

kugelkern/Charleston mutant embryos (Brandt et al., 2006; Pilot, 2006). Based on the observed phenotypes, we named the corresponding gene *fandango*, after the Iberian folk dance. *fandango* encodes the *Drosophila* ortholog of yeast SYF1 (synthetic lethal with cdc41) (Russell et al., 2000) and human XAB2 (XPA binding protein 2) (Kuraoka et al., 2008; Nakatsu et al., 2000). These proteins were described as subunits of the NTC/Prp19 complexes, which are important for spliceosome stabilization and activation (Chan, 2003; Chang et al., 2009; Hogg et al., 2010). Fandango protein has multiple tetratricopeptide repeat (TPR) motifs, which is a protein–protein interaction module (Zeytuni and Zarivach, 2012). Sequencing both alleles of *fandango* (*fand*1 and *fand*2) revealed distinct mutations within the *fandango* open reading frame (ORF). *fand*1 contained a missense point mutation in a highly conserved residue within TPR domain VII (from an alanine to a valine; A401V), whereas *fand*2 contained a microdeletion of 18 nucleotides within TPR domain VI, which deleted six conserved amino acids from position 355 to 360 (Figure 3.1—figure supplement 1B). In total protein extracts, both *fand*1 and *fand*2 mutant embryos showed a significant reduction in Fandango protein levels compared to control (Figure 3.1I). *fandango* mRNA levels, analyzed by realtime qPCR, were similar between control and *fand*1 mutant embryos (Figure 3.1J), suggesting that the mutation did not alter the stability of the encoding pre-mRNA.

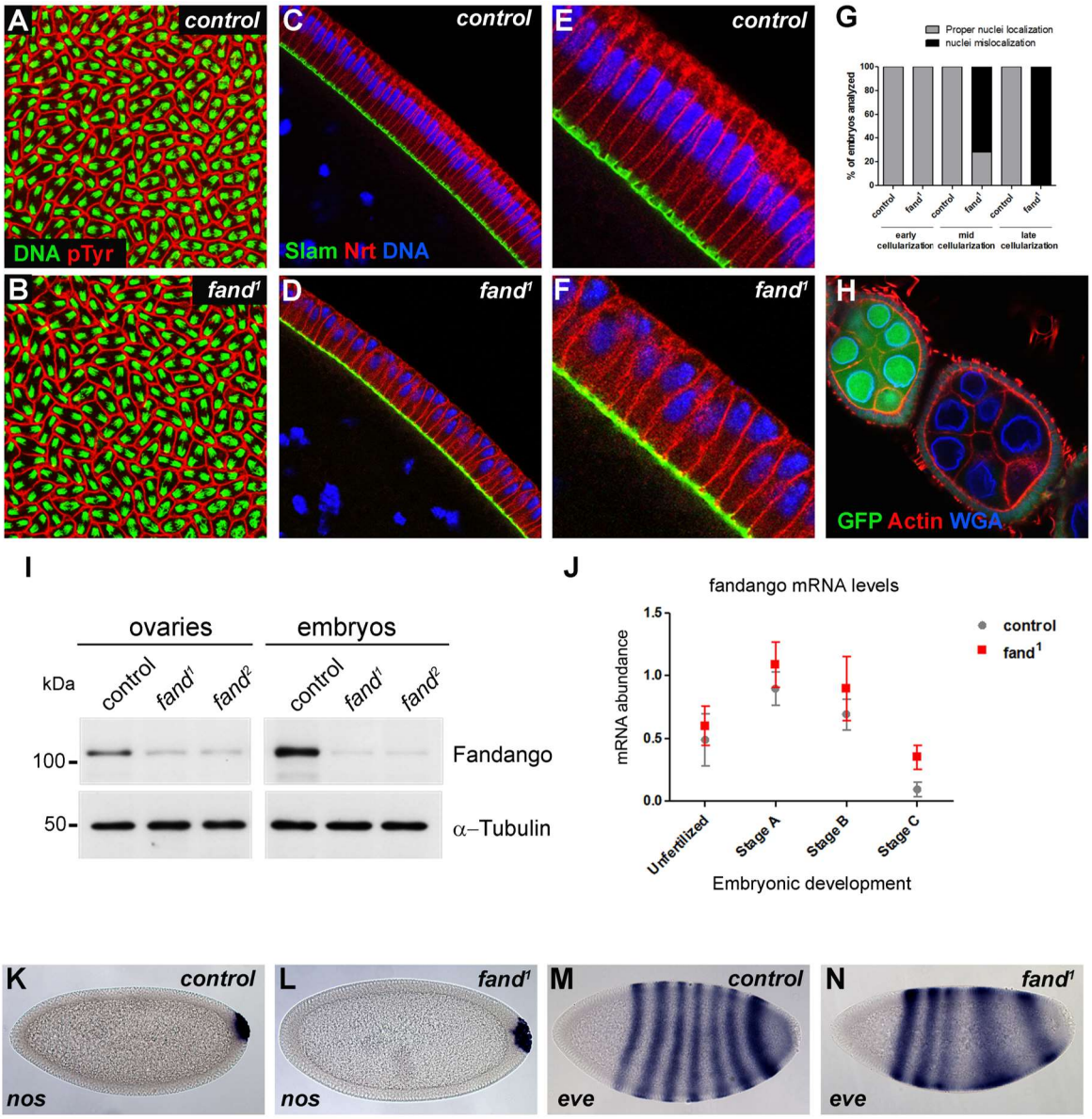



**Figure 3.1. *Drosophila* Fandango/Xab2 is required for blastoderm cellularization.** (A and B) Panels show embryos with normal syncytial blastoderm nuclear divisions in control embryos (hs-FLP; FRT42B) (A) and *fand*1 germ-line clone embryos (hs-FLP; FRT42B *fand*1, maternal mutant) (B). Embryos were stained for DNA (green) and p-Tyrosine (red). (C–F) Panels show blastoderm cellularized embryos. Control embryos showed normal epithelial architecture with elongated nuclei and columnar cell shape (C). *fand*1 germ-line clone mutant embryos showed abnormal epithelial architecture, the cortical nuclei failed to elongate and became mislocalized (D). (E and F) Magnification of C and D, respectively. Embryos were stained for Slam (green), Neurotactin (red), and DNA (blue). (G) Quantification of *fandango* maternal mutant embryo phenotype during blastoderm cellularization. Early cellularization: control: 100% normal (n = 44), *fand*1: 100% normal (n = 49); mid cellularization: control: 100% normal (n = 25), *fand*1: 28% normal (n = 21); late cellularization: control: 100% normal (n = 42), *fand*1: 0% normal (n = 38). (H) Maternally controlled oogenesis was normal in *fandango* mutant clones. Absence of endogenous nGFP (green) indicated that the cells were homozygous for *fand*1 mutation. Ovaries were stained for F-actin (red) and WGA (blue). (I) Western blot of whole protein extracts from embryos and ovaries mutant for *fand*1 and *fand*2 alleles (germ-line clones) showed a clear reduction in Fandango protein levels compared to

control tissues. It should be noticed that due to experimental constraints the total protein extracts from mutant ovaries included not only signal from mutant germ-line cells (homozygous for *fand*1), but also the tightly associated heterozygote somatic follicle cells. α-Tubulin was used as a loading control. (J) Real-time qPCR analysis showed no significant differences in *fandango* mRNA levels between control and *fand*1 embryos during development (Two-way ANOVA p>0.05 ns.). *fandango* mRNA levels were normalized with β-actin mRNA levels. (K–N) in situ hybridization for nanos RNA (maternal) and even-skipped RNA (early zygotic) in blastoderm cellularized embryos. Both control (K) and *fand*1 mutant (L) embryos showed normal nos localization pattern in the pole cells. *fand*1 embryos (N) showed A–P patterning defects of eve compared to control embryos (M).

### 3.1.3. *Drosophila* Fandango/Xab2 is differentially required for splicing of maternal and early zygotic pre-mRNAs

As noted above *fandango* maternal mutant embryos and kugelkern (kuk) mutant embryos showed remarkably similar blastoderm cellularization defects (Brandt et al., 2006; Pilot, 2006). Since *fandango* encodes a protein whose yeast and human orthologs are required for efficient spliceosome activity, we hypothesized that Fandango was required for splicing of kuk transcripts. kuk encodes two different transcripts, which vary in intron size (Figure 3.2A). Both transcripts are predicted to encode the same protein. Analysis of publicly available modENCODE transcriptome datasets (Graveley et al., 2011) suggested that the large kuk transcript was maternally expressed, whereas the small kuk transcript was only expressed zygotically. Through RT-PCR analysis we confirmed that similarly to control maternal genes (nanos and oskar) the large kuk transcript was maternally expressed (being present in unfertilized eggs), whereas the small kuk transcript was exclusively zygotically expressed (being present only in fertilized eggs) as the case of well-known early zygotic genes (even-skipped and krüppel) (Figure 3.2—figure supplement 1A). To investigate by RT-PCR whether Fandango was required for splicing of kuk pre-mRNAs, specific sets of primers (exon–exon, e–e; intron–exon, i–e) were designed for each kuk transcript, taking advantage of a longer 3′UTR in the small kuk transcript (Figure 3.2A). Surprisingly, whereas *fandango* embryos showed significant splicing defects of the small zygotic kuk transcript, the large maternal kuk transcript was correctly spliced (Figure 3.2B; Figure 3.2—figure supplement 1B). Splicing defects were fully rescued by a genomic fragment construct that contained a wild type copy of *fandango* (Figure 3.2—figure supplement 1C). The differential requirement of Fandango for

splicing of kuk transcripts prompted us to investigate more than 20 other maternal and early zygotic genes. RT-PCR analysis of *fandango* embryos invariably showed splicing defects of early zygotic but not maternally encoded transcripts (Figure 3.2C, data not shown). High-throughput transcriptome sequencing (RNAseq) confirmed that splicing of early zygotic but not maternally encoded gene products was affected in *fandango* embryos (Figure 3.2D, Figure 3.2—figure supplement 2A). Maternal transcripts, whose intron size was equivalent to those observed in early zygotic transcripts, were unaffected (Figure 3.2—figure supplement 2B), which showed that Fandango was not specifically rate limiting for splicing of small introns. Comparison analysis of 5′ and 3′ splice site consensus sequences between maternal and zygotic pre-mRNA transcripts showed no significant differences (Figure 3.2—figure supplement 2C) and the two populations of transcripts displayed a similarly heterogeneous exon–intron structure (Figure 3.2—figure supplement 2D). RT-PCR analysis of maternally encoded transcripts from wild-type and *fandango* mutant ovaries (germ-line clones) also failed to detect splicing defects (Figure 3.2—figure supplement 1D). This suggested that the absence of splicing defects of maternally encoded transcripts in *fandango* embryos was not due to specific degradation of unspliced transcripts during oogenesis. The differential requirement of Fandango for splicing of early zygotic encoded transcripts is fully consistent with the observation that maternally controlled oogenesis, primordial germ-cell formation, and syncytial nuclear divisions were normal in *fandango* mutants (Figure 3.1A,B,H,K,L), whereas the first detectable phenotype only occurred during zygotically controlled blastoderm cellularization (Figure 3.1C–F). Despite the fact that clonal analysis of the female germ line for both alleles of *fandango* showed normal oogenesis and egg laying (Figure 3.1H) (data not shown), Fandango protein levels were significantly reduced in the mutant ovaries (germ-line clones) (Figure 3.1I). *ango* embryos also failed to initiate germ-band extension after blastoderm cellularization (data not shown). It was previously shown that anterior–posterior (A–P) patterning is required for germ-band extension (Zallen and Wieschaus, 2004). Consistently, *fandango* embryos showed A–P patterning defects in the early zygotic pair-rule gene even-skipped (Figure 3.1M,N).

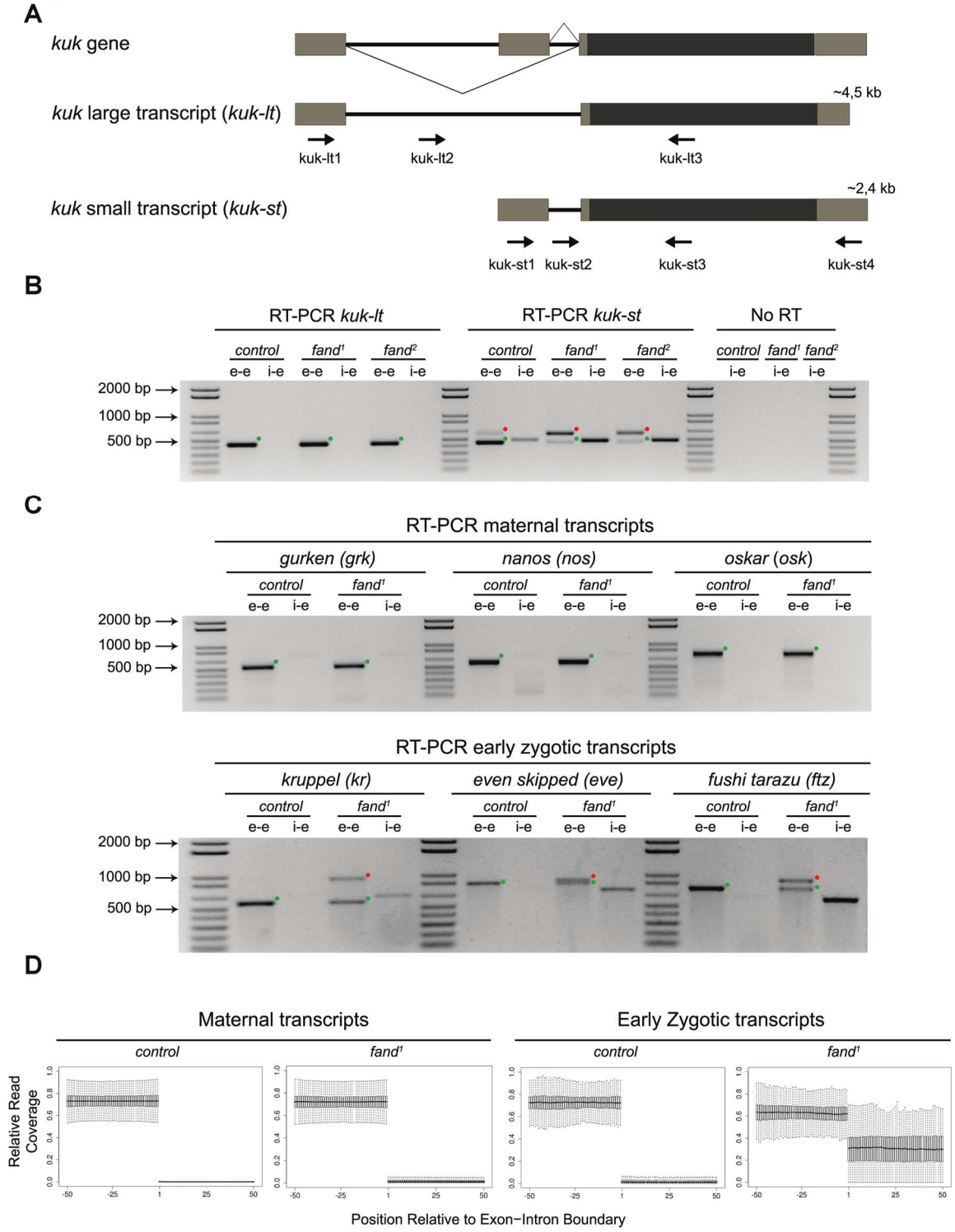



**Figure 3.2. Splicing of early zygotic but not maternally encoded pre-mRNAs is affected in *fandango* mutants.** (A) The kugelkern (kuk) locus encodes two transcripts of different size, kuk-lt containing a large intron and kuk-st with a short intron. Orientation and position of primers used for splicing analysis is indicated (arrows). (B) RT-PCR analysis of kuk transcripts. Control embryos yielded PCR products in the size predicted for the properly spliced forms of both kuk transcripts using exon–exon (e–e) primers (green dots, kuk-lt: 431 bp and kuk-st: 437 bp). *fandango* maternal mutant embryos Figure 2. (*fand*1 and *fand*2 alleles) showed splicing defects only in the kuk-st transcript; PCR products were detected by e–e primers in

the size expected for intron retention (red dots, kuk-st: 596 bp) and by intron–exon (i–e) primers (kuk-st: 474 bp). Splicing of the kuk-lt was not affected in *fandango* mutant background; PCR products were only detected with e–e primers in the predicted size for the correctly spliced pre-mRNA (green dots, kuk-lt: 431 bp). 'No RT' controls (only total RNA as template) yielded no amplification, meaning there was no contamination with genomic DNA in the samples tested. (C) RT-PCR analysis of maternal and early zygotic genes. Maternal transcripts were properly spliced, in both, control and *fand*1 mutant embryos; PCR products were only detected using e–e primers (green dots, grk: 527, nos: 581, osk: 762 bp). In contrast, early zygotic transcripts were correctly spliced only in control embryos (green dots, kr: 559, eve: 828, ftz: 753 bp). *fand*1 mutant embryos yielded PCR products in the size predicted for intron retention with e–e primers (red dots, kr: 932, eve: 899, ftz: 900 bp) and with i–e primers (kr: 629, eve: 720, ftz: 595 bp). All PCR bands showed in the panels were cloned and sequenced to confirm their identity. Green dots indicate correctly spliced transcripts, red dots indicate unspliced transcripts (intron retention). (D) RNA-Seq data confirmed that zygotic but not maternally encoded transcripts displayed a large fraction of splicing defects (intron retention) in *fand*1 mutant embryos. The panel shows box plot of the distribution of numbers of reads per bp relative to the total number of reads falling inside a 100 bp window centered around the 5′splice sites of zygotic (n = 408 splice sites from 270 genes) or maternal genes (n = 5876 splice sites from 2048 genes).

### 3.1.4. Fandango is similarly associated with the NTC/Prp19 complexes during oogenesis and early embryonic development

The highly conserved NTC/Prp19 and NTC/Prp19-related complexes are essential for pre-mRNA splicing as they facilitate the formation and progression between distinct spliceosome conformations during the splicing reaction (Chan, 2003; Hogg et al., 2010). Endogenous Fandango and Prp19 physically interacted in the early embryo (Figure 3.3A). Moreover, both endogenous Fandango and Prp19 physically interacted with endogenous ISY1 and CDC5L (Figure 3.3A), confirming that Fandango is a bona fide subunit of *Drosophila* NTC/Prp19 complexes. Immunoprecipitation of Myc-tagged Fandango and Myc-tagged Prp19 from embryonic protein extracts also identified an identical group of interacting proteins (Table 3.1; Supplementary file 1). Whereas Myc-Fandango mostly interacted with the NTC/Prp19-related complex subunits, Myc-Prp19 interacted principally with the NTC/Prp19 complex subunits. This illustrated that, as in humans, distinct but interacting NTC/Prp19 complexes exist in *Drosophila*, in agreement with the recent suggestion that a remarkable degree of conservation of distinct splicing complexes exists among metazoans (Herold et al., 2009). The differential requirements of Fandango for pre-mRNA splicing of maternal and early zygotic transcripts potentially suggest distinct interactions between Fandango and other splicing proteins during oogenesis and early embryonic development. Nevertheless,

immunoprecipitation of Myc- Fandango specifically expressed in the female germ line during oogenesis and in the early embryo identified a virtually identical group of interacting proteins: mostly subunits of the NTC/Prp19-related complex, and to a lesser extent, subunits of the NTC/Prp19 complex (Table 3.1; Supplementary file 1). These results showed that Fandango physically interacts with a similar group of splicing proteins during oogenesis and in the early embryo. To better understand the splicing defects observed in *fandango* embryos, we investigated if the integrity of NTC/Prp19 complexes was affected in this mutant. Size-exclusion chromatography showed detectable changes in the integrity of NTC/Prp19 complexes in *fandango* embryos (Figure 3.3B), with a significant reduction in the levels of ISY1 protein (Figure 3.3C). ISY1 is a NTC/Prp19-related complex subunit (Figure 3.3A). The loss of integrity of the ISY1-positive ~600–800 kDa NTC/Prp19 complex (Figure 3.3B) and concomitant reduction in the stability of some of their subunits, most likely impaired efficient activation of the spliceosome (Villa, 2005) and were likely explanations for the splicing defects observed in *fandango* embryos. In agreement with the suboptimal spliceosome activation hypothesis, intron retention was the main splicing defect of early zygotic transcripts in *fandango* embryos (Figure 3.2B,C, Figure 3.2—figure supplement 2B; data not shown). Levels of ISY1 were similarly affected in *fandango* mutants during oogenesis and in the early embryo (Figure 3.3C), suggesting this decrease did not explain the differential requirements of Fandango for splicing of early zygotic and maternally encoded transcripts. Mutant clonal analysis of a stronger allele of *fandango* (nonsense mutation), showed a complete loss of the female germ line in adult ovaries (data not shown). This demonstrated that the two isolated alleles of *fandango* are hypomorphic and suggested that Fandango was required, albeit at lower levels, for splicing of maternal transcripts. We concluded it is unlikely that a differential expression and/or association of core components of the spliceosome could potentially explain the differential requirements for Fandango between oogenesis and the early embryo. The most likely explanation is that Fandango is quantitatively (but not qualitatively) differentially required during early embryonic development.

**Table 1. LC-MS analysis of co-immunoprecipitation assays from ovaries and embryos.**
Co-immunoprecipitations were performed using total protein extracts from the different tissues expressing Myc-tagged Fandango or Myc-tagged-Prp19. Human and yeast homologues and the different sub-complexes are shown as described in (Herold et al., 2009). (−), (+), (++), (+++) correspond to 0, 1–9, 10–19, and >20 non-repeated peptides respectively. None of the proteins shown were detected in the negative controls (for detailed LC-MS analysis see Supplementary file 1)

| *Drosophila* | | *Human/ yeast* | Fandango-myc | | | | Prp19-myc | |
|---|---|---|---|---|---|---|---|---|
| CG | gene | | ovaries | | embryos | | embryos | |
| | | | rep1 | rep2 | rep1 | rep2 | rep1 | rep2 |
| **prp19 complex** | | | | | | | | |
| CG5519 | prp19 | PRP19 / Prp19 | + | + | + | + | +++ | ++ |
| CG6905 | cdc5-like | CDC5L / Cef1 | + | + | + | + | +++ | ++ |
| CG1796 | Tango4 | PLRG1 / Prp46 | + | + | + | + | + | + |
| CG4980 | - | BCAS2 / Snt309 | - | + | - | - | + | + |
| CG12135 | c12.1 | CWC15 / cwc15 | + | - | - | - | - | - |
| **prp19 related** | | | | | | | | |
| CG6197 | Fandango | Xab2 / Syf1 | +++ | +++ | +++ | +++ | + | + |
| CG31368 | - | AQR / - | +++ | +++ | +++ | +++ | + | + |
| CG4886 | cyp33 | PPIE / - | ++ | ++ | + | ++ | + | + |
| CG9667 | - | ISY1 / ISY1 | + | + | + | + | + | + |
| CG8264 | Bx42 | SNW1 / Prp45 | + | + | + | + | + | + |
| CG14641 | - | RBM22 / Cwc2 | - | + | + | + | + | - |
| CG3193 | crn | CRNKL1 / Clf1 | - | + | - | - | + | - |
| CG13892 | cypl | PPIL1 / - | - | - | - | + | - | - |
| CG1639 | l(1)10Bb | BUD31 / Bud31 | - | - | - | - | + | - |

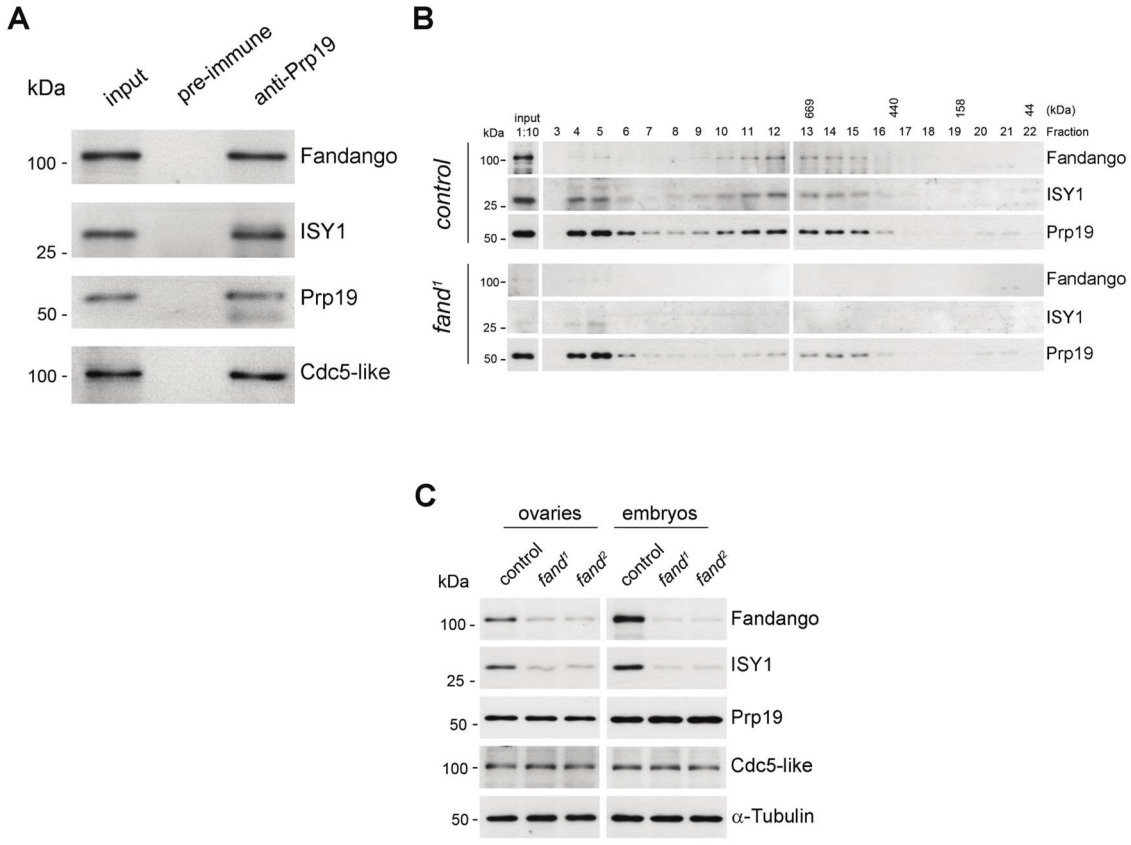



**Figure 3.3. Fandango physically interacts with a similar group of splicing proteins during oogenesis and embryogenesis.** (A) Pull down assay from nuclear-enriched protein extracts using a polyclonal antibody of Prp19. Endogenous Prp19 interacts physically with Fandango and other subunits of the NTC/Prp19 complexes (ISY1 and CDC5L). Pre-immune serum was used in the control. (B) Size-exclusion chromatography of control and *fand*1 mutant protein

extracts from 0–3 hr embryo collections using a Superose 6 10/300 column. After separation, each fraction was analyzed by Western blot. NTC/Prp19 complexes subunits (Prp19, Fandango, and ISY1) were part of a ∼600–800 kDa complex and also co-purified in a significantly larger complex (fraction 4 and 5). *fand*1 mutant protein extracts showed a significant reduction in levels of Fandango and ISY1 subunits and a size reduction of the Prp19-positive ∼600–800 kDa complex. (C) Western-blot analysis of total protein extracts from ovaries (left) and 0–3 hr embryos (right) from control and both *fandango* alleles, showed a reduction of Fandango and ISY1 protein levels in both tissues. Protein levels of Prp19 and CDC5L were not affected. α-Tubulin was used as loading control. Fandango Western blot is the same as shown in Figure 3.1I.

### 3.1.5. Reduction in Fandango levels affects mainly its splicing function

Transcriptional elongation can affect co-transcriptional splicing (Ip et al., 2011; de la Mata et al., 2003). It was recently shown that Syf1, the yeast ortholog of Fandango, is also important for Pol II transcriptional activity (Chanarat et al., 2011; David et al., 2011), therefore we decided to investigate transcription in *fandango* embryos. Three intronless early zygotic genes (nullo, snail, and scute) and two early zygotic genes with introns (even-skipped and tailless) were selected for further analysis by real-time qPCR. During mid/late-syncytial blastoderm (stage B) (Figure 3.4A, 'Materials and methods'), no significant differences in transcript abundance were observed between control and *fandango* (Figure 3.4—figure supplement 1A), whereas embryos mutant for grapes showed the expected reduction of transcript levels (Figure 3.4—figure supplement 1A; (Sibon et al., 1997). During transcriptional elongation, Pol II is specifically phosphorylated on the Ser2 residue of its carboxy-terminal domain (CTD) (Hsin and Manley, 2012). In agreement with the onset of early zygotic transcription, we observed a significant increase in Pol II CTD Ser2 phosphorylation as the embryo developed from early/mid-syncytial blastoderm (stage A), into mid/late-syncytial blastoderm (stage B), and blastoderm cellularization (interphase 14) (stage C) (Figure 3.4A,B). Both control and *fandango* embryos showed a similar increase in global levels of Pol II CTD Ser2 phosphorylation (Figure 3.4B,C). As transcriptional regulation during interphase 14 (stage C) relies on correct expression of early zygotic genes and degradation of many maternal RNAs (MZT) (Tadros and Lipshitz, 2009), we concluded that transcriptional changes at this stage (Figure 3.4—figure supplement 1A) were most likely a consequence of the widespread defects occurring during mid/late-syncytial blastoderm. Altogether, we concluded that the observed reduction in Fandango levels affects mainly its splicing function.

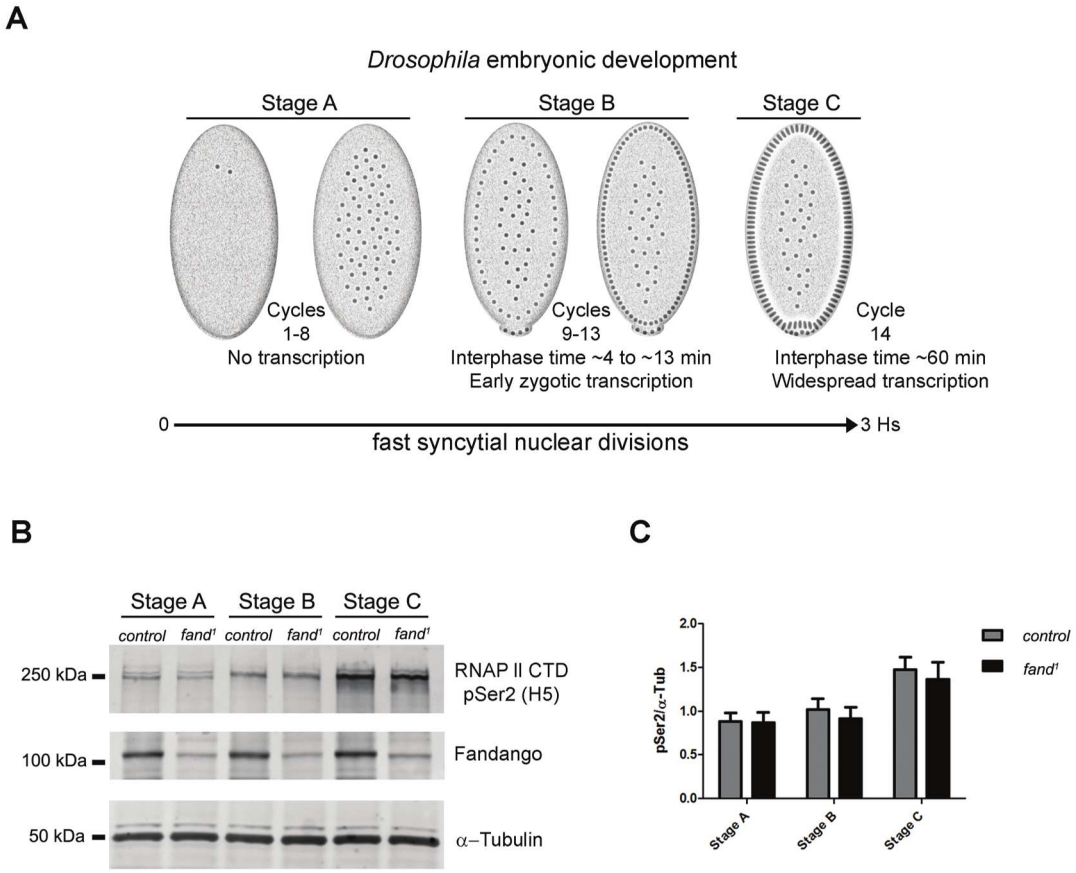



**Figure 3.4. Early zygotic transcription is not affected during mid/late-syncytial blastoderm in *fandango* mutants**. (A) Embryos were divided into three different groups according to developmental stage ('Materials and methods'), stage A: early/mid-syncytial blastoderm embryos, stage B: mid/late-syncytial blastoderm embryos, and stage C: blastoderm cellularization embryos. (B) Western blot for Pol II CTD Ser2 phosphorylation levels. Control and *fand*1 embryos showed a similar increase in the global levels of Pol II CTD Ser2 phosphorylation over the course of early embryonic development. α-Tubulin was used as a loading control. (C) Quantification of the CTD Ser2 phosphorylation from five independent western blot assays showed no significant difference at any of the embryonic developmental stages analyzed (Two-way ANOVA $p>0.05$ ns.).

### 3.1.6. Ectopic maternal expression of an early zygotic transcript in the mutant background was sufficient to suppress its splicing defects

To investigate if the differential requirement of Fandango for splicing of early zygotic and maternally encoded transcripts potentially resulted from distinct transcript sequences, we generated an early zygotic kuk transcript (kuk-LacZ) under the control of an UAS/Gal4 inducible promoter, where the open reading frame (ORF) was replaced by LacZ (Figure 3.5A,

see ‘Materials and methods’ for more details). As expected, when this construct was expressed zygotically, it was correctly spliced in control but not in *fandango* embryos (Figure 3.5B). In contrast, splicing of the kuk-LacZ construct occurred normally in both control and *fandango* mutants when it was expressed maternally (Figure 3.5B). Since maternal expression of an early zygotic transcript, in a *fandango* mutant background, was enough to suppress its splicing defects, we concluded that the differential requirement of Fandango for splicing of early zygotic transcripts was most likely due to the developmental context of gene expression and not a particularity in the early zygotic pre-mRNA sequences. Consistently, we failed to detect differences related to intron size, splice sites consensus, and exon–intron structure between maternal and zygotic transcripts (Figure 3.2—figure supplement 2B–D).

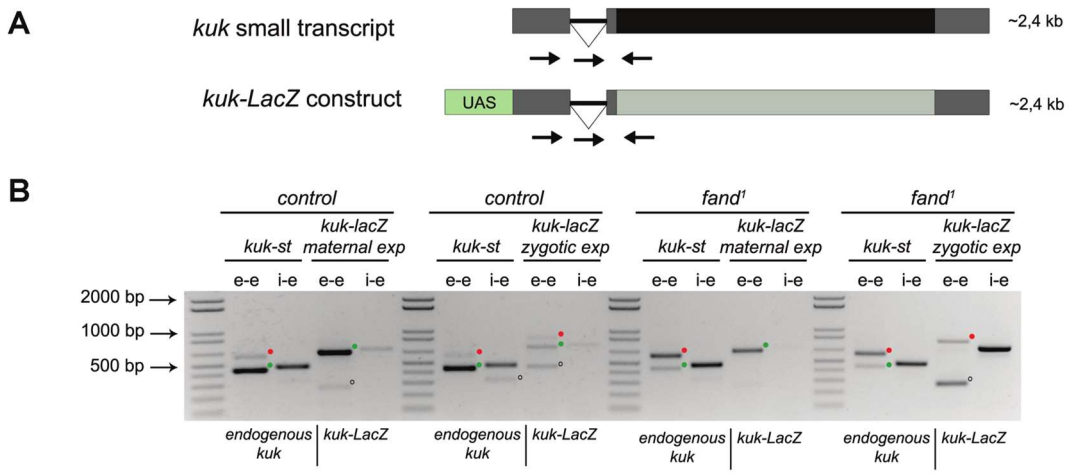



**Figure 3.5. Ectopic maternal expression of an early zygotic transcript in the mutant background is sufficient to suppress its splicing defects**. (A) The kuk-LacZ construct was built using the 5′UTR, the intron and the 3′UTR of the kuk small transcript (dark gray), and replacing the kuk ORF (black) by the LacZ coding sequence (light gray). To induce the expression of this construct it was put under the control of the UAS promoter (green) to drive the tissue specific expression with GAL4 drivers. Orientation and position of primers used for splicing analysis is indicated (arrows). (B) RT-PCR analysis of the kuk-LacZ construct. When it was zygotically expressed, it was correctly spliced in control but not in *fand*1 embryos (similarly to the endogenous small kuk transcript). Intron retention with e–e primers (red dots, kuk-st: 596 bp and kuk-LacZ: 869 bp) and a PCR product with i–e primers (751 bp) were observed in the mutant. When it was maternally expressed, kuk-LacZ construct was correctly spliced both in control and *fand*1 embryos, being detected just the spliced form of the construct (green dots, kuk-st: 437 bp and kuk-LacZ: 713 bp). In contrast, the endogenous zygotically expressed small kuk transcript (kuk-st) is still poorly spliced in *fand*1 embryos carrying the kuk-LacZ construct. Open circles indicate unspecific PCR products (confirmed by sequencing). Green dots indicate correctly spliced transcripts, whereas red dots indicate unspliced transcripts (intron retention).

### 3.1.7. A small early zygotic transcript with multiple introns was poorly spliced in wild-type embryos

*fandango* mutants showed a significant reduction in Fandango and ISY1 protein levels (Figure 3.3C), which most likely impaired efficient activation of the spliceosome (Villa, 2005). Since mitosis inhibits splicing (Shin and Manley, 2002), pre-mRNA splicing of early zygotic transcripts needs to be highly efficient for these genes to be correctly expressed. This suggests the existence of a developmental pre-requisite for highly efficient splicing, so that a suboptimal activation of the spliceosome would specifically impair pre-mRNA splicing of early zygotic but not maternal transcripts. Wildtype embryos already showed a detectable amount of intron retention in early zygotic transcripts (Figure 3.2B,D, Figure 3.2—figure supplement 2B), which was dramatically exacerbated in *fandango* embryos (Figure 3.2B,D, Figure 3.2—figure supplement 2B). We hypothesized that regardless of transcript size, there was also a constraint on pre-mRNA splicing of early zygotic transcripts in wild-type embryos. We generated a gene where the 5′UTR sequence including the intron of the small zygotic kuk transcript was quadruplicate to test this hypothesis (Figure 3.6A,D see 'Materials and methods' for more details). Quadruplicate introns were linked by in-frame LacZ coding sequences, and the entire construct (4x intron kuk-LacZ) was under the control either of an endogenous early zygotic minimal promoter (nullo-4x intron kuk- LacZ, ~2.5 Kb) (Figure 3.6A) or an inducible UAS/Gal4 promoter (UAS-4x intron kuk-LacZ, ~2.5 Kb) (Figure 3.6D). The total size of the encoded pre-mRNAs was comparable to many other endogenous early zygotic genes (e.g., kugelkern, runt, krupple). As a control, we introduced point mutations in the splice sites of these constructs to generate comparable intronless transcripts (no intron kuk-LacZ) (Figure 3.6A,D).

Only the first intron (int1) of the 4x intron kuk-LacZ construct was correctly spliced when it was zygotically expressed in wild-type embryos under the control of an endogenous early zygotic minimal promoter (Figure 3.6B). Likewise, when the 4x intron kuk-LacZ construct was early zygotically expressed under the control of the inducible promoter UAS/Gal4 there were similar splicing defects (intron retention) (Figure 3.6E). Measurement of in vivo kinetics of mRNA splicing showed that half-lives for splicing reactions are <1 min for the first intron, but 2–8 min for both second and third introns (Audibert et al., 2002). Hence, splicing of two or more introns requires more time than transcription and becomes rate limiting. Consistent with the hypothesis of a temporal constraint on pre-mRNA splicing, when the 4x

intron kuk-LacZ construct was zygotically expressed, the splicing defects of the firstly transcribed 5′-localized introns (Int1 and Int2) were significantly weaker than those observed in the later transcribed 3′-localized introns (Int3 and Int4) (Figure 3.6E). Importantly, maternal expression of this construct was sufficient to significantly suppress its splicing defects (Figure 3.6E). Real-time qPCR analysis showed that these constructs were equivalently zygotically and maternally expressed (Figure 3.6C,F). This suggested that splicing did not quantitatively impair early zygotic transcription, which was consistent with the observation that the rates of transcriptional elongation proceed independently of splicing (Brody et al., 2011).

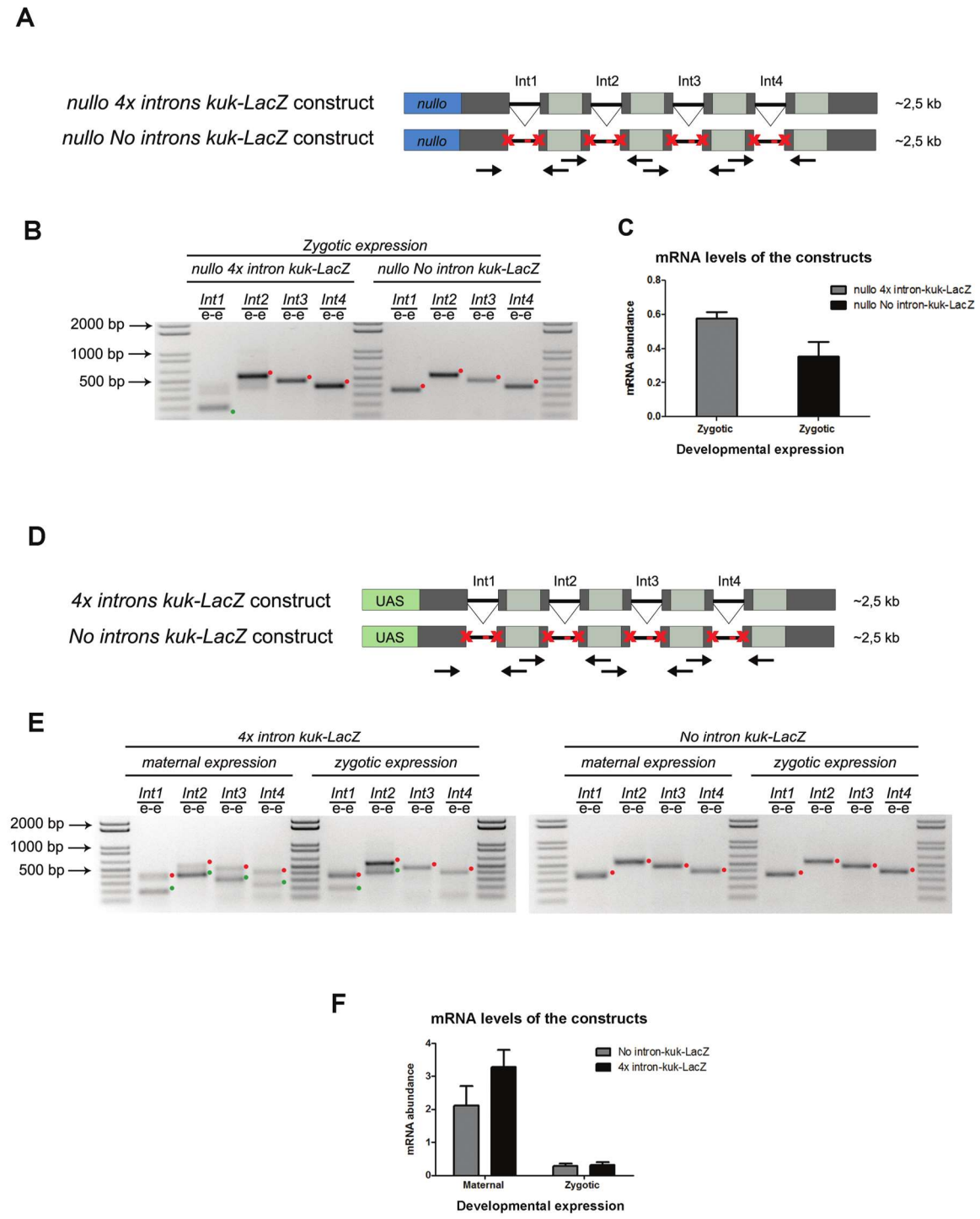



**Figure 3.6. A small early zygotic transcript containing four introns is poorly spliced in wild-type embryos.** (A and D) The 4x intron kuk-LacZ construct was a variant of the kuk-LacZ that contains four copies of kuk small transcript intron (dark gray). Each intron is separated by 201 nucleotides of an in frame Lac-Z sequence (light gray). The no intron kuk-LacZ construct has all splice sites present in the 4x intron kuk-LacZ construct mutated to thymidines. The constructs were fused to a nullo minimal promoter (blue) (A), or fused to an inducible UAS promoter (green) (D). Orientation and position of primers used for splicing analysis is indicated (arrows). (B) RT-PCR analysis showed significant splicing defects (intron

retention) of the 4x intron kuk-LacZ construct when expressed under the control of an endogenous early zygotic promoter (nullo promoter). The first intron was correctly spliced, being detected mainly the PCR product corresponding to the spliced form (green dot). The remaining introns (second, third, and fourth) were completely unspliced (red dots, intron retention). In the intronless (no intron kuk-LacZ) construct, under the control of the same nullo promoter were only observed PCR bands whose sizes correspond to unspliced forms (red dots, intron retention). (C) Real-time qPCR analysis showed that the 4x intron kuk-LacZ and no intron kuk-LacZ constructs were expressed to the same extent when under the control of the nullo minimal promoter (t test $p>0.05$ ns.). (E) RT-PCR analysis of the 4x intron kuk-LacZ construct showed significant splicing defects (intron retention) when zygotically expressed in wild-type embryos under the control of an inducible UAS promoter. Although the most 5′-localized introns (first and second) were still partially spliced, being observed two PCR bands corresponding to the spliced (green dots, int1: 191 and int2: 385 bp), and unspliced forms (red dots, int1: 347 and int2: 541 bp). The furthest 3′-localized introns (third and fourth) were completely unspliced, being only observed one PCR band with the size corresponding to intron retention (red dots, int3: 463 and int4: 385 bp). Maternal expression of the 4x intron kuk-LacZ construct was sufficient to significantly suppress splicing defects in the four introns analyzed (green dots, spliced forms: int1: 191, int2: 385, int3: 307, int4: 229 bp; red dots, unspliced forms: int1: 347, int2: 541, int3: 463, int4: 385 bp). Zygotic and maternal expression of the no intron kuk-LacZ construct only showed PCR bands with sizes corresponding to unspliced forms (red dots, intron retention). (F) Real-time qPCR analysis showed that the 4x intron kuk-LacZ and no intron kuk-LacZ constructs were expressed to the same extent both maternally (Two-way ANOVA $p>0.05$ ns.) and zygotically ($p>0.05$ ns.) in wild-type embryos. All PCR bands shown in these panels were cloned and sequenced to confirm their identity. Green dots indicate correctly spliced transcripts, red dots indicate unspliced transcripts (intron retention).

**A**

| Cross | | Offspring | | | | | |
|---|---|---|---|---|---|---|---|
| | | CyO flies | | | CyO⁺ flies (*w; fand¹/fand²*) | | |
| | F1 ♂ | n | % | % exp | n | % | % exp |
| ♀ w/w; *fand¹*/CyO x ♂ *wt-fandango* ; *fand²*/CyO | | 382 | 37 | 33 | 0 | 0 | 17 |
| | | CyO flies | | | CyO⁺ flies (w/*wt-fandango* ; *fand¹/fand²*) | | |
| | F1 ♀ | n | % | % exp | n | % | % exp |
| | | 448 | 43 | 33 | 187 | 18 | 17 |

| Cross | | Offspring | | | | | |
|---|---|---|---|---|---|---|---|
| | | CyO flies | | | CyO⁺ flies (*wt-fandango* ; *fand¹/fand²*) | | |
| | F1 ♂ | n | % | % exp | n | % | % exp |
| ♀ *wt-fandango* ; *fand²*/CyO x ♂ w/w ; *fand¹*/CyO | | 176 | 33 | 33 | 82 | 15 | 17 |
| | | CyO flies | | | CyO⁺ flies (w/*wt-fandango* ; *fand¹/fand²*) | | |
| | F1 ♀ | n | % | % exp | n | % | % exp |
| | | 182 | 34 | 33 | 93 | 17 | 17 |

**B**

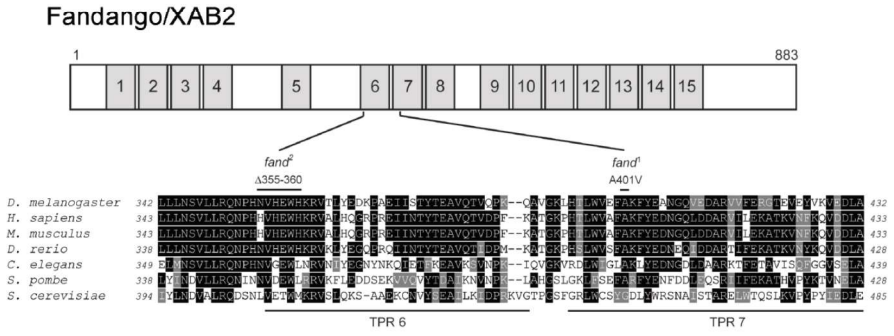



**Figure 3.1 - figure supplement 1. *Fanango* mutant alleles contain changes in highly conserved amino acids.** (A) Panels show the rescue of the transheterozygous zygotic lethality of *fand*$^{1}$ and *fand*$^{2}$ alleles. Only transheterozygous mutant females carrying the wild-type copy of *fandango* (*wt-ango*; FRT42B, *fand*$^{1/2}$) were viable (dark gray box). Transheterozygous mutant males without receiving the genomic fragment (FRT42B, *fand*$^{1/2}$) died (light gray box). In the reciprocal cross, both females and males transheterozygous mutants carrying the wild-type copy of *fandango*, were viable (dark gray box). (B) Fandango has multiple copies of a tetratricopeptide repeat (TPR) motif, a protein–protein interaction module found in a number of functionally different proteins. A scheme displaying the distribution of conserved TPR protein domains in Fandango (top). Mutations of *fand*$^{1}$ and *fand*$^{2}$ alleles affect highly conserved amino acids of the TPR domains 7 and 6, respectively. *fand*$^{1}$ contained a missense point mutation changing an alanine to a valine at aminoacid position 401 (A401V) and *fand*$^{2}$ contained a microdeletion which resulted in loss of six conserved amino acids from position 355 to 360 (Δ355–360). Partial alignment of Fandango (*Drosophila*

*melanogaster* CG6197, ref.NP_610891.1) with orthologous Xab2 (*Homo sapiens*, ref. NP_064581.2), Xab2 (*Mus musculus*, ref. NP_080432.1), Xab2 (*Danio rerio*, ref.NP_001038248.1), C50F2.3 (*Caenorhabditis elegans*, ref. NP_491250.1), cwf3 (*Schizosaccharomyces pombe*, ref. NP_596612.1), and SYF1 (*Saccharomyces cerevisiae*, ref. NP_010704.1) (bottom).

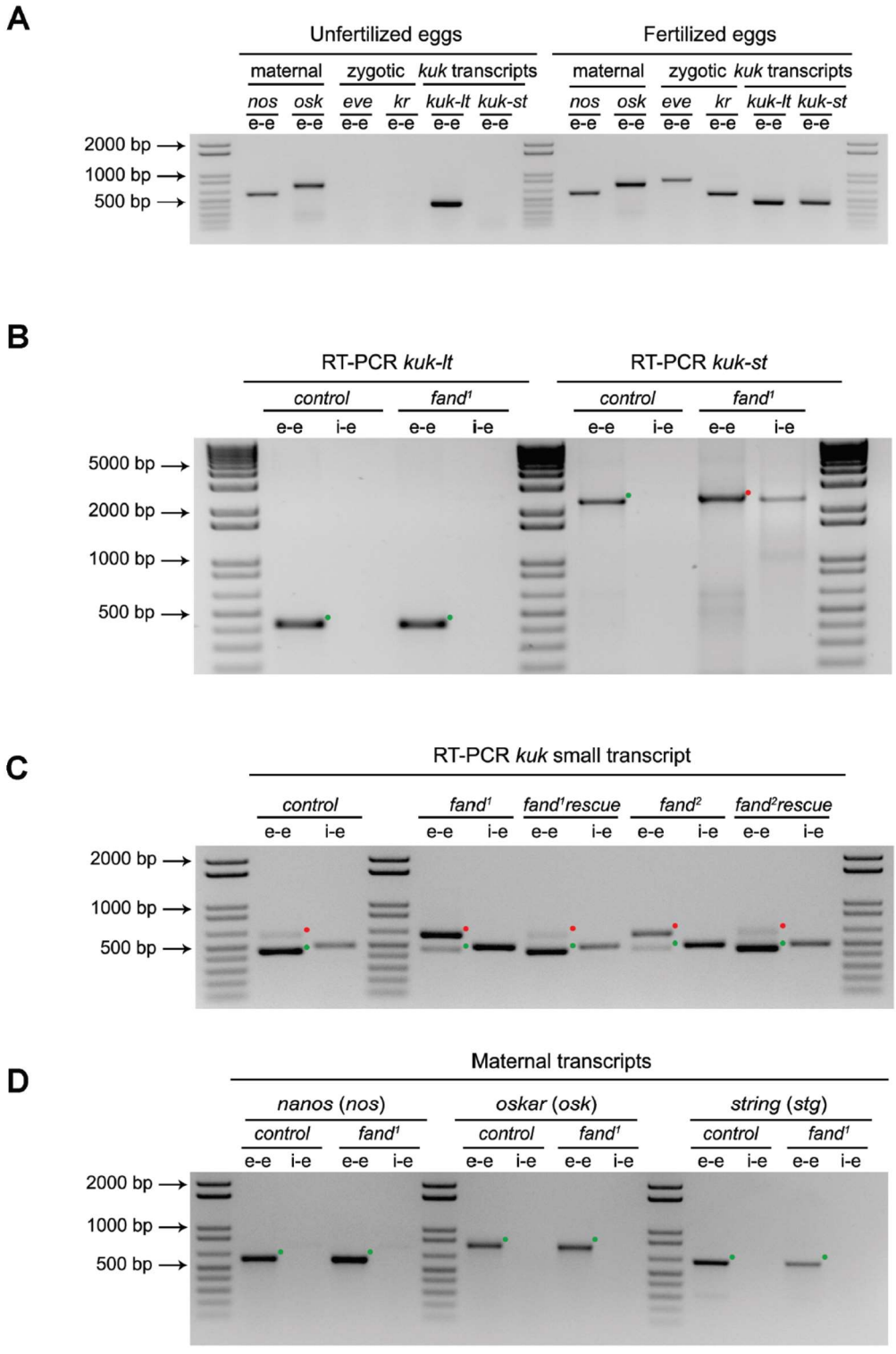


**Figure 3.2 - figure supplement 1. Splicing of early zygotic but not maternally encoded pre-mRNAs is affected in *fandango* mutants.** (A) RT-PCR analysis of *kuk* transcripts from unfertilized and fertilized eggs. *kuk-lt* is maternal and *kuk-st* only zygotically expressed. *kuk-st* (437 bp) transcripts are only detected in fertilized eggs, as are other well-known early zygotic transcripts (*kr*: 559, *eve*: 828 bp). *kuk-lt* (431 bp) is amplified from both fertilized and unfertilized eggs, as are other known maternal transcripts (*nos*: 581, *osk*: 762 bp). (B) RT-PCR analysis of *kuk* transcripts with a specific reverse primer for *kuk-st* (kuk-st4). Control embryos yielded PCR products with the size expected for properly spliced *kuk* transcripts using e–e primers (green dots, *kuk-lt*: 431 bp and *kuk-st*: 2257 bp). *fand*$^{1}$ maternal mutant embryos showed splicing defects in *kuk-st*; PCR products were detected in the size expected for intron retention with e–e primers (red dot, *kuk-st*: 2413 bp) and by i–e primers (*kuk-st*: 2293 bp). Splicing of the *kuk-lt* was not affected in *fandango* mutant background; a PCR product was only detected with e–e primers in the expected size for correctly spliced pre-mRNAs (green dot, *kuk-lt*: 431 bp). (C) RT-PCR analysis of *kuk-st* showed the rescue of splicing defects observed in both *fandango* alleles by a genomic fragment construct derived from the third chromosome that contained a wild-type copy of *fandango*. Embryos analyzed were laid by GLC females FRT42B *fand*$^{1}$/CyO; *wt-fandango* or FRT42B *fand*$^{2}$/CyO; *wt-fandango*. GLC FRT42B and mutant GLC *fand*$^{1}$ and *fand*$^{2}$ embryos were used as controls. (D) RT-PCR analysis of maternally encoded transcripts from wild-type and *fandango* mutant ovaries (germ-line clones) failed to detect any splicing defects. The PCR products detected in both samples were of the size predicted for properly spliced pre-mRNAs (green dots, *nos*: 581, *osk*: 762, *stg*: 614 bp). All PCR products shown in these panels were cloned and sequenced to confirm their identity. Green dots indicate correctly spliced transcripts and red dots indicate unspliced transcripts (intron retention).

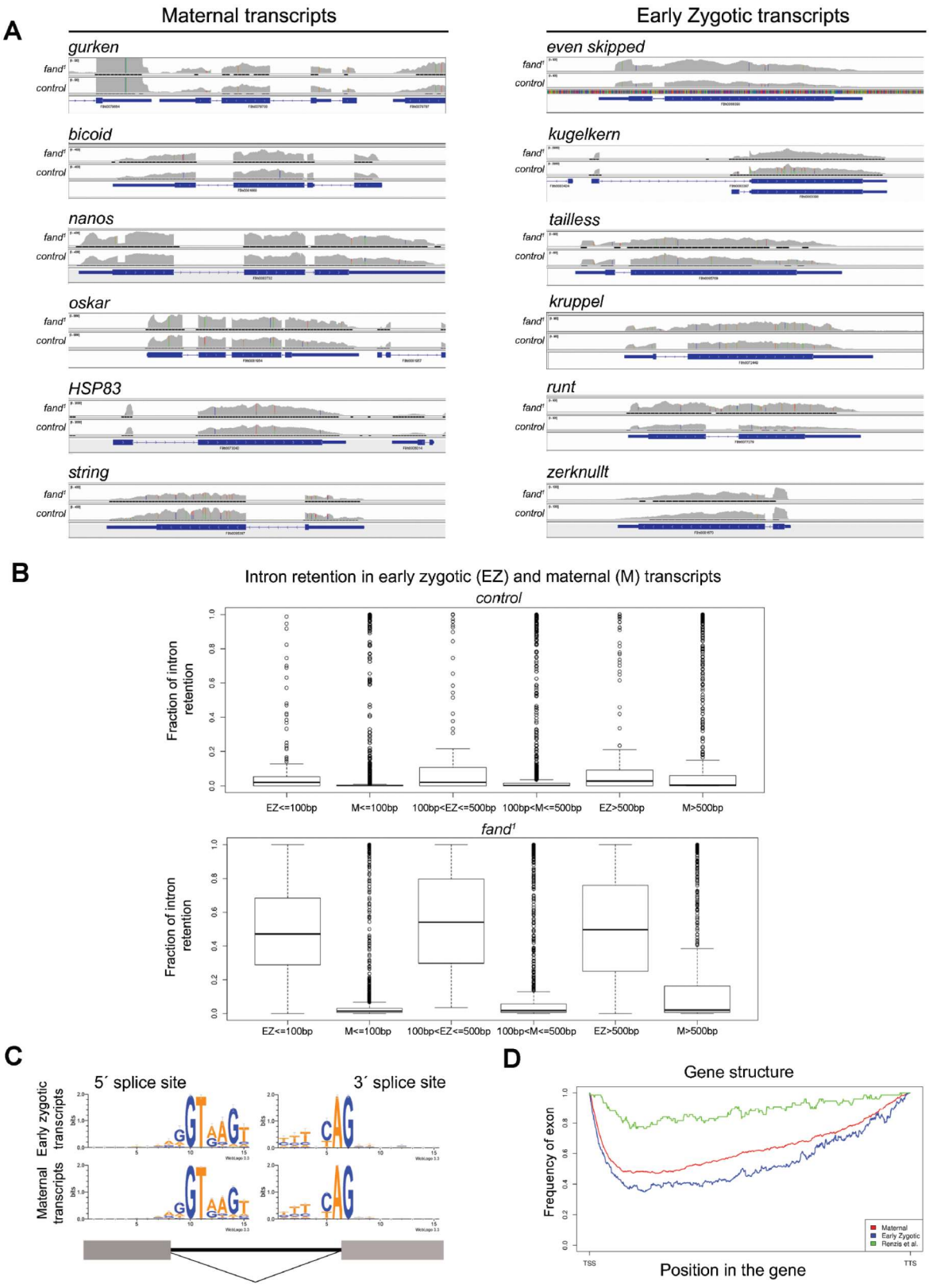



**Figure 3.2 - figure supplement 2. Early zygotic but not maternally encoded pre-mRNAs shows significant intron retention in *fandango* mutants.** (A) Zygotic genes display intron retention in the *fand*$^{1}$ mutant, while maternal genes do not. This panel shows Integrative Genomics Viewer (IGV) screenshots of RNA-Seq read coverage from tophat alignments of a selected set of known early zygotic and maternally expressing genes. Within each gene, the

image scale is identical for both *fand*$^{1}$ and control. (B) Zygotic genes in *fand*$^{1}$ mutant present clear evidence of intron retention, independent of intron size. The panel displays box plot distributions of percentage intron retention of exon–intron boundaries of early zygotic and maternal genes in *fand*$^{1}$ and control. Exon–intron boundaries were divided in bins, by intron size (less than 100 bp; from 100 to 500 bp; and greater than 500 bp). Sizes were selected empirically to have comparable dataset sizes in each bin (150–200 boundaries for zygotic genes, 1000–3000 for maternal genes). The frequency of intron retention was determined by comparing the number of unsplit reads overlapping the splice site with the number of reads that showed an exon–exon split (see 'Materials and methods' for more details). (C) Splice sites in zygotic and maternal genes presented the same characteristic sequence pattern (5′ GT; 3′ AG). (D) Zygotic genes and maternal genes (see 'Materials and methods' for more details) did not reveal any distinguishing features in terms of exon–intron structure. The exon frequency is close to that expected from a random distribution, roughly 50% (with the obvious exception of gene endings—TSS and TTS). For comparison, the 59 early zygotic genes described by (De Renzis et al., 2007), 70% of which are intronless, display a very distinct, non-random, pattern.

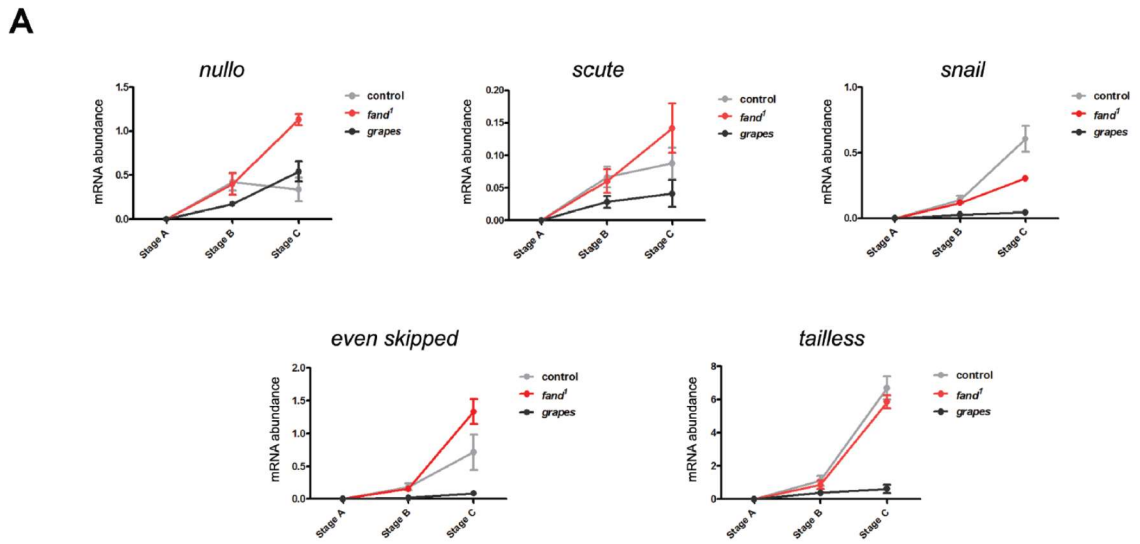



**Figure 3.4 - figure supplement 1. Early zygotic transcription is not affected during mid/late-syncytial blastoderm in *fandango* mutants**. (A) Real-time qPCR analysis to measure levels of early zygotic transcripts from intronless (*nullo*, *snail,* and *scute*) and containing introns genes (*even-skipped* and *tailless*) during embryonic development. mRNA levels from early zygotic genes were normalized with β-actin mRNA levels. At stage B, when early zygotic genes are transcribed, there was no significant differences in mRNA abundance for any of the genes analyzed in either control or *fandango* samples (Two-way ANOVA p>0.05 ns.).

Detailed LC-MS analysis of co-immunoprecipitation assays from ovaries and embryos.

| Reproducible peptides | | | | Ovaries Fandango-Myc | | | | | | Embryos Fandango-Myc | | | | | | Embryos Prp19-Myc | | | | | | Ovaries Control (-) | | | | | | Embryos Control (-) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Rep 1 | | | Rep 2 | | | Rep 1 | | | Rep 2 | | | Rep 1 | | | Rep 2 | | | Rep 1 | | | Rep 2 | | | Rep 1 | | | Rep 2 | | |
| CG | *Drosophila* | *Human* | Mass (kDa) | Score | Matches | Sequences | Score | Matches | Sequences | Score | Matches | Sequences | Score | Matches | Sequences | Score | Matches | Sequences | Score | Matches | Sequences | Score | Matches | Sequences | Score | Matches | Sequences | Score | Matches | Sequences | Score | Matches | Sequences |
| **Prp19 related complex** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| CG6197 | Fandango | XAB2 | 103,83 | 1188 | 30 | 24 | 1506 | 38 | 27 | 1201 | 30 | 23 | 1223 | 28 | 24 | 57 | 2 | 2 | 39 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG31368 | - | AQR | 172,98 | 2576 | 66 | 47 | 3319 | 75 | 53 | 2579 | 62 | 47 | 1640 | 38 | 30 | 214 | 8 | 8 | 30 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG4886 | cyp33 | PPIE | 33,65 | 843 | 17 | 10 | 850 | 18 | 13 | 478 | 10 | 9 | 959 | 18 | 12 | 108 | 1 | 1 | 126 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG9667 | - | ISY1 | 31,45 | 653 | 16 | 7 | 541 | 14 | 8 | 511 | 12 | 6 | 621 | 16 | 9 | 123 | 2 | 2 | 163 | 3 | 2 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG8264 | Bx42 | SNW1 | 61,18 | 92 | 1 | 1 | 141 | 2 | 2 | 72 | 1 | 1 | 87 | 1 | 1 | 96 | 1 | 1 | 136 | 2 | 2 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG14641 | - | RBM22 | 47,94 | - | - | - | 48 | 1 | 1 | 60 | 2 | 2 | 68 | 2 | 2 | 53 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG3193 | crn | CRNKL1 | 84,49 | - | - | - | 87 | 3 | 3 | - | - | - | - | - | - | 108 | 2 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| **Prp19 complex** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| CG5519 | Prp19 | PRPF19 | 55,56 | 348 | 9 | 7 | 441 | 11 | 8 | 389 | 12 | 9 | 287 | 9 | 7 | 10498 | 188 | 20 | 2606 | 65 | 16 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG6905 | cdc5-like | CDC5L | 93,26 | 108 | 2 | 2 | 259 | 5 | 3 | 266 | 5 | 4 | 256 | 6 | 6 | 1466 | 30 | 23 | 907 | 19 | 16 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG1796 | Tango4 | PLRG1 | 53,06 | 162 | 1 | 1 | 211 | 2 | 2 | 176 | 2 | 2 | 211 | 2 | 2 | 642 | 8 | 6 | 321 | 4 | 4 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG4980 | - | BCAS2 | 31,44 | - | - | - | 30 | 1 | 1 | - | - | - | - | - | - | 393 | 7 | 7 | 138 | 3 | 3 | - | - | - | - | - | - | - | - | - | - | - | - |
| **Spliceosome related** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| CG5931 | l(3)72Ab | SNRNP200 | 245,78 | 151 | 2 | 2 | 65 | 1 | 1 | 102 | 2 | 2 | 96 | 2 | 2 | 186 | 2 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG8877 | Prp8 | PRPF8 | 280,71 | 81 | 3 | 3 | 148 | 3 | 3 | 79 | 2 | 2 | - | - | - | 265 | 6 | 6 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG4849 | - | EFTUD2 | 111,61 | 60 | 1 | 1 | 69 | 1 | 1 | 132 | 2 | 2 | 78 | 3 | 3 | 100 | 3 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG13900 | - | SF3B3 | 138,01 | - | - | - | - | - | - | 51 | 1 | 1 | - | - | - | 118 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG16941 | - | SF3A1 | 88,02 | - | - | - | - | - | - | - | - | - | - | - | - | 105 | 3 | 3 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG2807 | - | SF3B1 | 150,28 | - | - | - | - | - | - | - | - | - | - | - | - | 63 | 2 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG7971 | - | SRRM2 | 122,87 | - | - | - | 62 | 1 | 1 | 182 | 4 | 4 | 162 | 3 | 3 | 202 | 4 | 4 | 238 | 5 | 5 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG11274 | SRm160 | SRRM1 | 107,76 | - | - | - | - | - | - | - | - | - | 72 | 1 | 1 | 153 | 2 | 2 | 84 | 2 | 2 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG8427 | SmD3 | SNRPD3 | 15,69 | - | - | - | - | - | - | - | - | - | - | - | - | 50 | 2 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG5119 | pAbp | PABPC4 | 70,11 | 170 | 3 | 3 | 185 | 3 | 3 | 150 | 3 | 3 | 405 | 6 | 5 | 375 | 7 | 7 | 356 | 6 | 5 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG12343 | - | SYF2 | 26,52 | 100 | 1 | 1 | 91 | 1 | 1 | 87 | 1 | 1 | 104 | 2 | 2 | 107 | 1 | 1 | 121 | 2 | 2 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG10753 | SmD1 | SNRPD1 | 13,79 | - | - | - | 47 | 1 | 1 | - | - | - | - | - | - | 95 | 2 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG6015 | - | CDC40 | 65,96 | - | - | - | - | - | - | - | - | - | 123 | 2 | 1 | 104 | 2 | 2 | 70 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG2063 | - | SAP30BP | 36,25 | - | - | - | - | - | - | - | - | - | - | - | - | 50 | 1 | 1 | 56 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |
| **Miscellaneous** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| CG2175 | dec-1 | - | 179,65 | 1753 | 26 | 16 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG4466 | Hsp27 | CRYAB | 23,66 | 422 | 8 | 5 | 377 | 7 | 5 | 56 | 1 | 1 | - | - | - | 54 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG13804 | yellow-g2 | - | 43,18 | 123 | 3 | 3 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG1404 | ran | RAN | 24,92 | 36 | 1 | 1 | 49 | 1 | 1 | - | - | - | - | - | - | 36 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG10686 | tral | LSM14A/LSM14B | 69,47 | - | - | - | 122 | 2 | 2 | - | - | - | - | - | - | 102 | 2 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG9775 | - | CHTOP | 46,10 | - | - | - | - | - | - | 90 | 1 | 1 | 91 | 1 | 1 | 204 | 3 | 2 | 91 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG1683 | Ant2 | SLC25A4 | 34,01 | - | - | - | - | - | - | 61 | 1 | 1 | 53 | 2 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG4147 | Hsc70-3 | HSPA5 | 72,33 | - | - | - | - | - | - | 45 | 1 | 1 | 49 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG12750 | ncm | CWC22 | 152,00 | - | - | - | - | - | - | - | - | - | 140 | 2 | 1 | 113 | 2 | 2 | 164 | 2 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG8308 | alphaTub67C | TUBAL3 | 51,95 | - | - | - | - | - | - | - | - | - | 110 | 2 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG1341 | Rpt1 | PSMC2 | 48,91 | - | - | - | - | - | - | - | - | - | 63 | 2 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG8351 | Tcp-1eta | CCT7 | 59,98 | - | - | - | - | - | - | - | - | - | - | - | - | 199 | 4 | 4 | 83 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG7235 | Hsp60C | HSPD1 | 61,89 | - | - | - | - | - | - | - | - | - | - | - | - | 136 | 3 | 3 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG8258 | - | CCT8 | 59,80 | - | - | - | - | - | - | - | - | - | - | - | - | 133 | 2 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG8977 | Cctgamma | CCT3 | 59,93 | - | - | - | - | - | - | - | - | - | - | - | - | 100 | 2 | 2 | 87 | 2 | 2 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG8231 | Tcp-1zeta | CCT6A | 58,50 | - | - | - | - | - | - | - | - | - | - | - | - | 89 | 1 | 1 | 55 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG7033 | - | CCT2 | 58,48 | - | - | - | - | - | - | - | - | - | - | - | - | 79 | 2 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG17949 | His2B:CG17949 | HIST3H2BB/HIST1H2BO | 13,69 | - | - | - | - | - | - | - | - | - | - | - | - | 66 | 2 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG12101 | Hsp60 | HSPD1 | 60,89 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 292 | 5 | 4 | - | - | - | - | - | - | - | - | - | - | - | - |
| **Ribosomal** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| CG2746 | RpL19 | RPL19 | 24,15 | 99 | 1 | 1 | - | - | - | - | - | - | - | - | - | 169 | 2 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG7726 | RpL11 | RPL11 | 21,27 | 44 | 1 | 1 | - | - | - | 56 | 1 | 1 | 41 | 1 | 1 | - | - | - | 38 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG1524 | RpS14a | RPS14 | 16,31 | 39 | 1 | 1 | - | - | - | - | - | - | - | - | - | 71 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG5920 | RpS2 | RPS2 | 29,11 | 35 | 1 | 1 | - | - | - | - | - | - | - | - | - | 41 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG8857 | RpS11 | RPS11 | 18,26 | 32 | 1 | 1 | 51 | 1 | 1 | 41 | 1 | 1 | - | - | - | 53 | 2 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG10305 | RpS26 | RPS26 | 13,54 | 31 | 1 | 1 | - | - | - | - | - | - | - | - | - | 53 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG7808 | RpS8 | RPS8 | 23,86 | - | - | - | 144 | 2 | 2 | 133 | 2 | 2 | 70 | 1 | 1 | 171 | 2 | 2 | 134 | 2 | 2 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG6779 | RpS3 | RPS3 | 27,68 | - | - | - | 72 | 1 | 1 | 83 | 1 | 1 | - | - | - | 76 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG7014 | RpS5b | RPS5 | 25,99 | - | - | - | 59 | 1 | 1 | - | - | - | 50 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG3661 | RpL23 | RPL23 | 15,04 | - | - | - | 59 | 1 | 1 | - | - | - | 92 | 1 | 1 | 167 | 2 | 1 | 82 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG9091 | RpL37a | RPL37 | 10,87 | - | - | - | 44 | 1 | 1 | - | - | - | - | - | - | 39 | 1 | 1 | 55 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG10423 | RpS27 | RPS27 | 9,70 | - | - | - | 31 | 1 | 1 | - | - | - | - | - | - | 50 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG9282 | RpL24 | RPL24 | 17,62 | - | - | - | - | - | - | 86 | 1 | 1 | - | - | - | 88 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG3922 | RpS17 | RPS17 | 15,33 | - | - | - | - | - | - | - | - | - | 64 | 1 | 1 | 53 | 1 | 1 | 62 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG7977 | RpL23A | RPL23A | 29,45 | - | - | - | - | - | - | - | - | - | 58 | 1 | 1 | 92 | 2 | 2 | 79 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG4651 | RpL13 | RPL13 | 25,05 | - | - | - | - | - | - | - | - | - | 43 | 1 | 1 | 48 | 1 | 1 | 57 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG3314 | RpL7A | RPL7A | 30,88 | - | - | - | - | - | - | - | - | - | 36 | 1 | 1 | 47 | 1 | 1 | 59 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG4897 | RpL7 | RPL7 | 29,59 | - | - | - | - | - | - | - | - | - | - | - | - | 96 | 2 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG7434 | RpL22 | RPL22 | 30,65 | - | - | - | - | - | - | - | - | - | - | - | - | 88 | 2 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CG5502 | RpL4 | RPL4 | 45,11 | - | - | - | - | - | - | - | - | - | - | - | - | 73 | 2 | 2 | 51 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |
| CG11276 | RpS4 | RPS4X | 29,23 | - | - | - | - | - | - | - | - | - | - | - | - | 55 | 2 | 2 | 31 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |

**Supplementary table 3.1 Complete list of proteins specifically co-immunoprecipitating with Myc-tagged Fandango and Myc-tagged Prp19.** List of proteins co-immunoprecipitating with Myc-tagged Fandango in ovaries (germ-line) and embryos, and with Myc-tagged Prp19 in embryos. Protein extracts from embryos and ovaries not expressing the Myc-tagged proteins were used as negative controls (−). Two replica experiments were performed for each condition. Proteins were identified by LC-MS-MS and data were blasted against the Flybase database using Mascot Search (see 'Materials and methods' for detail). The presence of a protein was measured from the total number of peptides (Matches), the number of non-repeated sequence peptides (Sequences) and the respective Score according to Mascot Search. Predictions of Human protein orthologous from *Drosophila* proteins were made using DIOPT (see 'Materials and methods' for detail). Proteins were grouped in NTC/Prp19-related complex (light gray), NTC/Prp19 complex (dark gray), general spliceosomal related proteins, miscellaneous, and ribosomal proteins. Non-reproducible (only 1 peptide in one replica) and non-specific proteins (more than 1 peptide in negative controls [−]) are listed separately. (*) Spliceosomal proteins in non-reproducible table, (**) NTC/Prp19 complex or NTC/Prp19-related complex subunits present in non-reproducible and non-specific tables.

## 3.2 Intronic sequences requirements during *Drosophila* early development.

Pedro Prudêncio [1, 2], Rosina Savisaar [1], Rui Gonçalo Martinho[1, 2, 3]

[1] Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal.

[2] Center for Biomedical Research, Universidade do Algarve. Faro, Portugal.

[3] iBiMED, Departamento de Ciências Médicas, Universidade de Aveiro. Aveiro, Portugal

**Author contribution**

Pedro Prudêncio and Rui Gonçalo Martinho designed the experiments. Pedro Prudêncio performed most of the experiments and data analysis. Rosina Savisaar helped with intron conservation analysis.

### 3.2.1. Overview

Evidences from the last chapter of this thesis (Guilgur et al., 2014) suggest that during *Drosophila* early embryonic development there are constraints not only to transcriptional elongation but also to splicing, which together could explain why approximately 70% of early zygotic genes are short and intronless (De Renzis et al., 2007a). However, there is a fraction of early zygotic genes (approximately 30%) that contain introns. Since early zygotic intron-containing genes frequently show a complex pattern of gene expression (e.g. embryonic patterning genes), we hypothesized that these pre-MBT introns might be functionally relevant for gene expression.

To test this hypothesis, we decided to investigate the functional requirements of two highly conserved introns within two well characterized embryonic patterning pre-MBT genes: knirps (kni) and even skipped (eve). Both kni and eve show a highly complex expression pattern during early embryonic development, being their transcription dynamically regulated by distinct transcription factors (Bothma et al., 2014; Jaeger, 2011; Small et al., 1991). The function of kni and eve is crucial for anterior-posterior segmentation of the developing *Drosophila* embryos. In order to investigate the functional requirement of kni and eve introns, we investigated if loss-of-function mutant alleles of kni and eve could be rescued by genomic constructs of kni and eve, with or without their introns.

Although we successfully obtained transgenic flies carrying a genomic transgene of eve, with or without its intron, we failed however to obtain flies carrying genomic transgenes of kni. Since both eve transgenes (with or without the intron) rescued the embryonic patterning defects of eve mutant embryos, the first intron of eve is not likely to be functionally relevant during early embryonic development. In conclusion, whereas the eve results do not support our hypothesis, further work is nevertheless still needed to fully investigate the functional requirement of pre-mMBT introns during early embryonic development, and clarify why a subset of pre-MBT expressed genes have introns.

### 3.2.2. The position of the first intron of knirps and even-skipped is conserved among arthropods

We focused our analysis on evolutionary conserved pre-MBT introns, since this is good indicator of functional relevance. To identify such introns, we aligned the amino acid sequence from several well characterized embryonic patterning genes between *Drosophila* melanogaster and the correspondent orthologs within distinct arthropods species, in order to obtain the coding sequences alignment that are present in the exons. This way, and based on the sequence provided by the correspondent transcripts, it was possible to identify introns whose position is conserved in all species, independently of their length. For even-skipped (eve) and knirps (kni) (Figure 3.7), it was possible to see that the position of several introns was not conserved between species, getting loss or gained during evolution. However, in both genes, the position of the first intron was conserved in all analysed species (red arrow on Figure 3.7). This was previously reported for knirps (Naggan Perl et al., 2013), but not for even skipped.

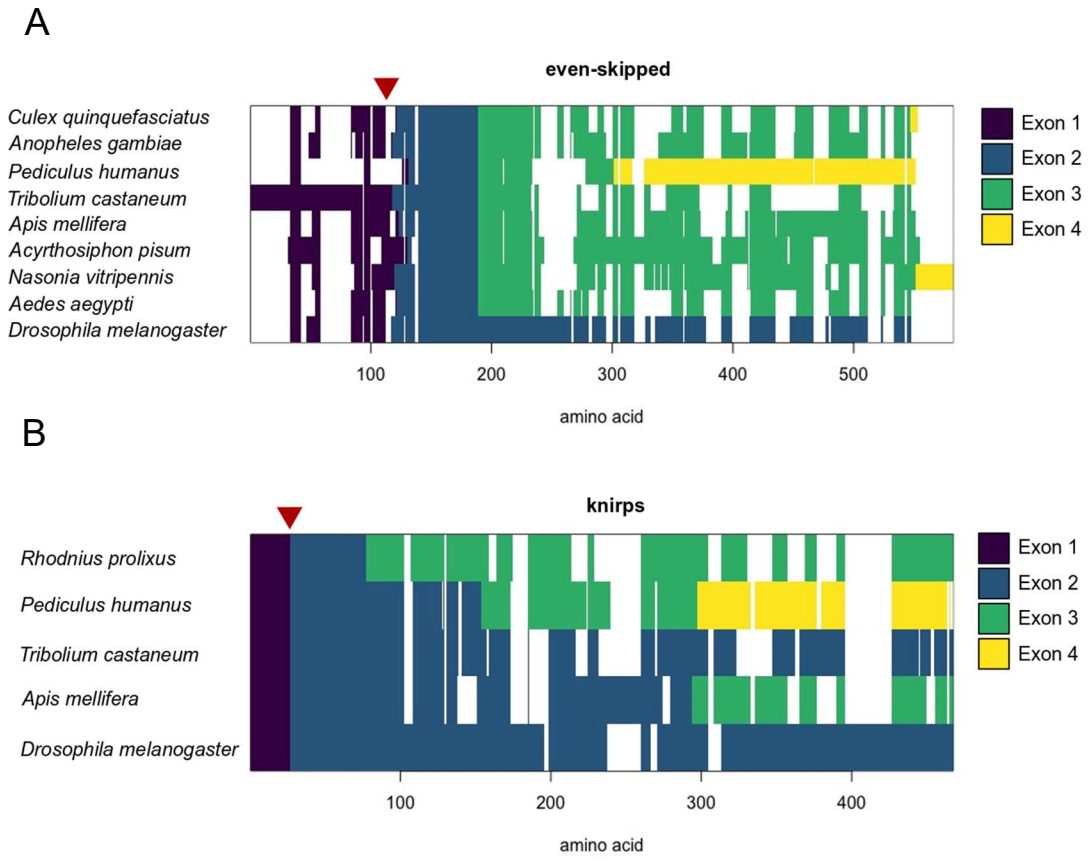



**Figure 3.7. knirps and even-skipped first introns are conserved among arthropods.** (A and B) Alignment between conserves amino acid sequences in different arthropods species, for

even-skipped (A) and knirps (B) genes. Correspondent protein sequences from each exon were colored in order to highlight common introns (red arrows).

### 3.2.3. even-skipped intron is not required for expression and function.

To investigate whether the first intron of eve or kni are functionally relevant for the correct expression of these genes, we attempted to generate transgenic flies carrying transgenes of kni and eve with or without their first introns. Because the correct regulation of eve and kni transcription requires several enhancer sequence elements, which are in some cases located several kilobases upstream and downstream of the gene transcription unit, we used bacterial artificial chromosomes (BACs) that were previously shown to rescue the correspondent mutants (-26kb to +8,5kb from TSS) for eve (Fujioka, 1999) and (-31,8kb to +7,1kb from TSS) for kni (Pankratz et al., 1992).

After successful deletion of both first introns by homologous recombination (see methods for more detail), both control and intron-depleted BACs were integrated in *Drosophila* using PhiC31 integrase-mediated transgenesis. While injection of 300 and 4500 embryos was suficient, respectively, to generate flies carrying a eve control and eve first-intron-deleted BAC transgenes, injection of 1500 embryos failed to generate flies carrying neither kni control or intron deleted BAC transgenes. This suggest some degree of toxicity associated to this specific BAC, which led us to abandon the generation of kni transgenes. To investigate if the first intron of eve is rate-limiting for the function of this gene, we compared the ability of both transgenes to complement the viability of the $eve^{3}$ null mutant allele. *Drosophila* embryos are sensitive to eve dosage, being two wild-type copies of this gene necessary to rescue the mutant viability (Ludwig et al., 2011). As expected null mutant $eve^{3}$ viability was fully rescued by 2 copies of the eve control BAC (BAC eve ctr), but not by only one copy. Surprisingly, 2 copies of the eve intron-depleted BAC (BAC eve Del-Int) equally rescue $eve^{3}$ viability (Figure 3.8A), indicating that the first intron of eve is apparently not rate-limiting for the function of this gene, at least in the tested conditions.

Since the presence of an intron was shown to facilitate transcription rate of a reporter gene during *Drosophila* early embryonic development (Fukaya et al., 2017), we investigated in first intron deletion somehow impaired the total levels of eve mRNA. To test this hypothesis, we decided to measure and compare eve mRNA levels of both control and first intron-deleted

transgenes expressed in embryo. Once more no differences were observed (Figure 3.8B), again suggesting no relevant function of this intron during early embryonic development.

A

| | | Control | | Deleted Intron | |
|---|---|---|---|---|---|
| Cross | virgin | *w/w ; eve $^{3}$/CyO ; BAC eve Ctr / BAC eve Ctr* | | *w/w ; eve $^{3}$/CyO ; BAC eve Del-Intron / BAC eve Del-Intron* | |
| | males | *w ; Df eve / CyO ; BAC eve Ctr /* MKRS | | *w ; Df eve / CyO ; BAC eve Del-Intron /* MKRS | |
| | | Avg. | St.Dev. | Avg. | St.Dev. |
| Offspring | CyO Sb *(w/w ; eve $^{3}$ or Df eve / CyO ; BAC eve /BAC eve)* | 45.49% | 4.85% | 44.01% | 8.60% |
| | CyO Sb+ *(w/w ; eve $^{3}$ or Df eve/ CyO ; BAC eve /* MKRS) | 45.17% | 3.04% | 47.05% | 6.11% |
| | CyO+ Sb *(w/w ; eve $^{3}$/Df eve ; BAC eve /BAC eve)* | 9.34% | 5.26% | 8.93% | 3.99% |
| | CyO+ Sb+ *(w/w ; eve $^{3}$/Df eve ; BAC eve /* MKRS) | 0.00% | 0.00% | 0.00% | 0.00% |
| | total flies counted | 182 | | 166 | |

B

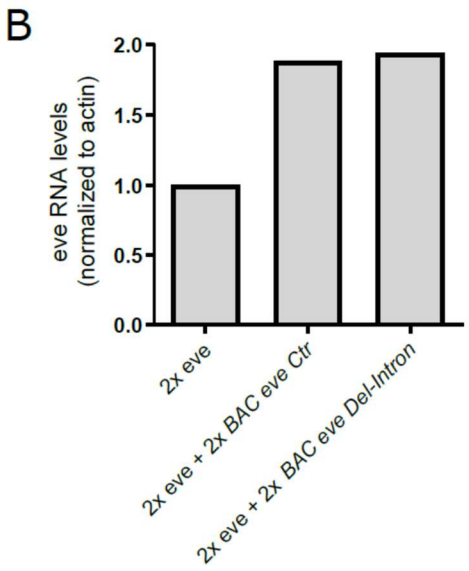



**Figure 3.8. eve-Deletion-intron transgene rescues *eve* mutant viability phenotype and eve intron do not affect gene expression levels**. (A) 2 copies of transgene are required to rescue *eve*$^{3}$ mutant viability. Both transgenes Control (Ctr) and Del-intron, equally rescue *eve* mutant viability, Paired t test, P value = 0.9416. (B) eve qRT-PCR shows that both Control and Del-Intron BAC transgenes are equally expressed in embryos.

## 3.3 Splicing takes place as RNA polymerase II transcribes past recursive and canonical splice sites in the developing *Drosophila* embryo.

The results presented below are, at the moment of this thesis submission, in manuscript format ready to be submitted. An article describing the methodology used to analysed NET-seq data was published in a peer-reviewed Methods journal in 2019 during my PhD entitled: Analysis of Mammalian Native Elongating Transcript sequencing (mNET-seq) high-throughput data. The publication format can be found in the Appendix of this thesis.

Pedro Prudêncio [1, 2], Kenny Rebelo [1], Rosina Savisaar [1], Rui Gonçalo Martinho[1, 2, 3] and Maria Carmo-Fonseca [1]

[1] Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal.

[2] Center for Biomedical Research, Universidade do Algarve. Faro, Portugal.

[3] iBiMED, Departamento de Ciências Médicas, Universidade de Aveiro. Aveiro, Portugal

**Author contribution**

Pedro Prudêncio, Rui Gonçalo Martinho and Maria Carmo-Fonseca designed the experiments. Pedro Prudêncio performed most of the experiments. Pedro Prudêncio, Kenny Rebelo and Rosina Savisar performed the data analysis.

### 3.3.1. Overview

Widespread co-transcriptional splicing has been demonstrated from yeast to human. However, measuring the kinetics of splicing relative to transcription has been hampered by technical challenges. Here, we took advantage of the native elongating transcript sequencing strategy in *Drosophila* embryos (*d*NET-seq) to identify the position of RNA polymerase II (Pol II) when exons become ligated in the newly synthesized transcript. We detected reads spanning recursive splicing and exon-exon junctions in transcripts connected to the active site of Pol II molecules positioned a few nucleotides downstream of recursive and canonical 3′ splice sites, indicating that splicing can occur shortly after a splice site emerges at the polymerase exit channel. Immediate splicing was observed in genes transcribed in early embryos, which are intronless or contain few short introns, as well as in genes expressed later in development that contain multiple long introns. The latter observation was quite unexpected as it argues that exon definition is not mandatory for metazoan splicing. We further found a higher density of polymerases associated with nascent spliced transcripts suggesting a splicing-coupled mechanism that slows down transcription elongation. Taken together our data reveal a tight temporal coordination between splicing and the elongating Pol II complex during embryonic development.

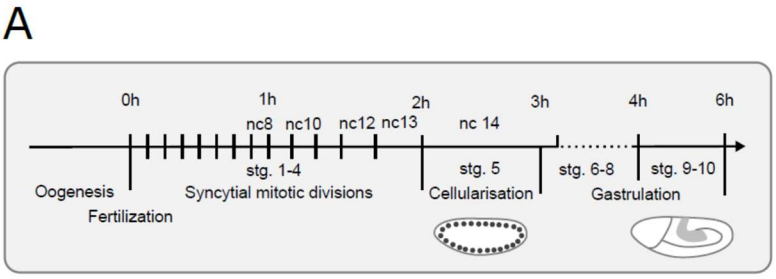
### 3.3.2. Native elongating transcript sequencing in *Drosophila* embryos (*d*NET-seq)

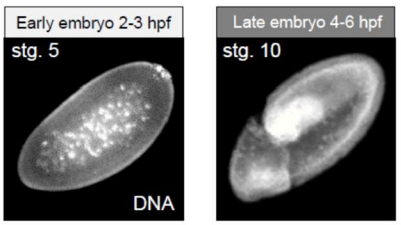
Our first task was to adapt the NET-seq method for use in *Drosophila* embryos. NET-seq relies on the intrinsic stability of ternary transcription complexes formed by RNA polymerase II (Pol II), the DNA template and nascent RNA to isolate Pol II elongation complexes by immunoprecipitation without crosslinking. NET-seq was initially established in yeast to visualize the genomic position of the active site of Pol II by identifying the 3′ ends of the nascent RNA (Churchman and Weissman, 2011a). The method was later applied to mammalian cells and termed mNET-seq (Nojima et al., 2015a, 2016) (see also (Mayer et al., 2015)). Because solubilisation of Pol II complexes under native conditions is typically incomplete in metazoan cells (Kimura et al., 1999), the mNET-seq protocol uses micrococcal nuclease (MNase) digestion of isolated chromatin (Nojima et al., 2016). In adapting mNET-seq to *Drosophila* embryos, we optimized buffers and washing conditions to purify the

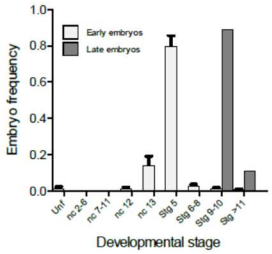
chromatin fraction from manually sorted embryos, solubilize the transcription complexes with MNase digestion and immunoprecipitate Pol II with antibodies. We used rabbit polyclonal antibodies raised against synthetic peptides of the YSPTSPS repeat of the CTD of the largest Pol II subunit in *Saccharomyces cerevisiae*, phosphorylated at either S5 (ab5131 Abcam) or S2 (ab5095 Abcam). Both antibodies have been extensively used for chromatin immunoprecipitation experiments and shown to react with *Drosophila melanogaster* (Dahlberg et al., 2015). Embryos were collected at 2-3 hours after fertilization (referred to as early embryos) and 4-6 hours after fertilization (referred to as late embryos) (Figure 3.1A). Analysis of embryos stained with a fluorescent dye to visualize DNA (Figure 3.1B) revealed that early embryos were predominantly in cycle 14 (stage 5), whereas the majority of late embryos were in the late stage of germ-band extension (stage 10; Figure 3.1C).

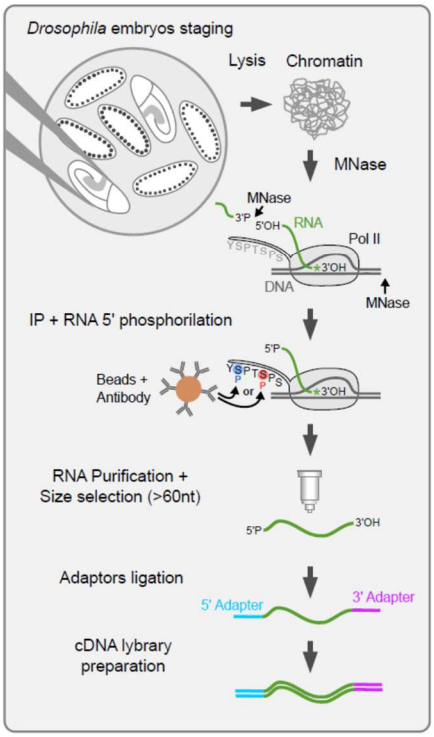
To enable directional sequencing, the 5′ hydroxyl (OH) generated by MNase digestion of RNA was first converted to a 5′ phosphate by T4 polynucleotide kinase (Figure 3.1D). RNA was then purified from the immunoprecipitated Pol II complexes and size selected using a column. RNAs with a size above 60 nucleotides (nt) were used for subsequent ligation of specific adapters to the 5′ P and 3′ OH ends of each RNA fragment followed by PCR-based preparation of a cDNA library for high-throughput Illumina sequencing (Figure 3.1D). After sequencing, adapter sequences were trimmed and paired-end reads with sequence overlaps were merged into a single read that spans the full length of the original RNA fragment (dark orange; Figure 3.1E). The resulting single reads were aligned to the *Drosophila* reference genome. The nucleotide at the 3′ end of each RNA fragment was identified and its genomic position recorded (asterisk; Figure 3.1E).

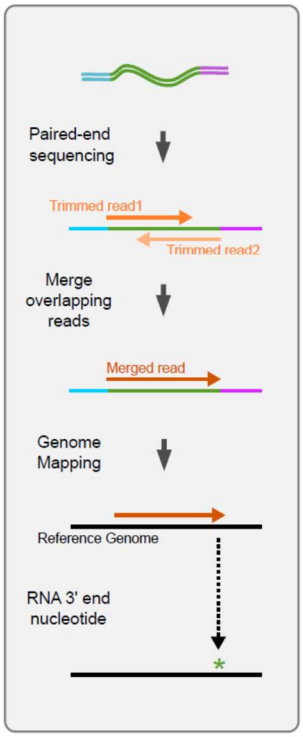
Two to three NET-seq libraries were independently prepared from early and late embryos using S5P antibody; three additional libraries were prepared from late embryos using S2P antibody. Each library was sequenced to high coverage with a read length of 150 bp (Figure 3.9 - figure supplement 1A). Experimental reproducibility was demonstrated by strong agreement of uniquely aligned read density between biological replicates (Figure 3.9F and Figure 3.9 - figure supplement 1 B, C). Very similar values of read density were also observed in late embryo samples prepared with S5P and S2P antibodies (Figure 3.9G), suggesting that in *Drosophila* embryos, the CTD of elongating Pol II is phosphorylated on both serine 5 and serine 2 positions.

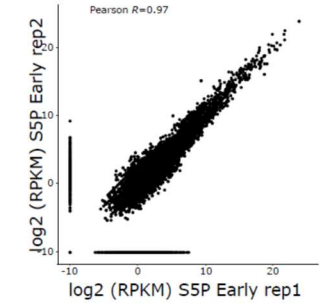

A
0h
1h
2h
3h
4h
6h
nc8 nc10 nc12 nc13
nc 14
stg. 1-4
stg. 5
stg. 6-8
stg. 9-10
Oogenesis
Fertilization
Syncytial mitotic divisions
Cellularisation
Gastrulation


B
Early embryo 2-3 hpf
stg. 5
DNA
Late embryo 4-6 hpf
stg. 10


C
Early embryos
Late embryos
Embryo frequency
Developmental stage


D
Drosophila embryos staging
Lysis
Chromatin
MNase
RNA
Pol II
DNA
IP + RNA 5' phosphorilation
Beads + Antibody
RNA Purification + Size selection (>60nt)
Adaptors ligation
5' Adapter
3' Adapter
cDNA lybrary preparation


E
Paired-end sequencing
Trimmed read1
Trimmed read2
Merge overlapping reads
Merged read
Genome Mapping
Reference Genome
RNA 3' end nucleotide


F
Pearson R=0.97
log2 (RPKM) S5P Early rep2
log2 (RPKM) S5P Early rep1

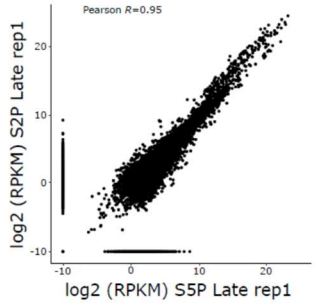

G
Pearson R=0.95
log2 (RPKM) S2P Late rep1
log2 (RPKM) S5P Late rep1

**Figure 3.9. Native elongating transcript sequencing in *Drosophila* embryos. (A)** Timeline of *Drosophila* early embryonic development. **(B)** Representative images (stained for DNA) of embryos in nuclear cycle 14 (stage 5) and late germ-band expansion (stage 10). **(C)** The graph depicts the developmental stage of embryos sorted into the "early" and "late" groups. **(D)** Outline of the *d*NET-seq experimental protocol. **(E)** Outline of *d*NET-seq data analysis. **(F)** Density of uniquely aligned reads per gene (RPKM in log2 scale) for two *d*NET-seq/S5P biological replicates from early embryos (Pearson's correlation, R = 0.97). **(G)** Density of uniquely aligned reads per gene (RPKM in log2 scale) from late embryos analyzed by *d*NET-seq/S5P and *d*NET-seq/S2P (Pearson's correlation, R = 0.95).

### 3.3.3. *d*NET-seq reveals co-transcriptional splicing associated with Pol II phosphorylated on CTD serine 5.

NET-seq captures not only the final (3′ OH end) nucleotide of nascent RNA but also the 3′ OH end of RNAs that associate with the Pol II elongation complex (Nojima et al., 2015a, 2018; Schlackow et al., 2017). Notably, in humans, mNET-seq of Pol II phosphorylated on CTD serine 5 detected splicing intermediates formed by cleavage at the 5′ splice site after the first splicing reaction. The presence of such intermediates manifests as an enrichment of reads whose 3′ ends map precisely to the last nucleotide of an exon. In both early and late *Drosophila* embryos, we indeed observed large peaks of *d*NET-seq/S5P reads mapping to the last nucleotide of spliced exons, as shown for the *kuk* gene (asterisk; Figure 3.10A. We also detected *d*NET-seq/S5P peaks at the last nucleotide of introns, as shown for the *eEF1alpha1* gene (asterisk; Figure 3.10B); enrichment for these reads results from co-immunoprecipitation of released intron lariats after completion of the splicing reaction (Figure 3.10B). In addition, we observed reads corresponding to mature snRNAs engaged in co-transcriptional spliceosome assembly, suggesting that *d*NET-seq was capturing the free 3′ OH ends of the snRNAs. (Figure 3.10C). We found prominent peaks at the end of spliceosomal U1, U2, U4 and U5 snRNAs. As expected, no peak was detected mapping to the end of the U3 snRNA, which is involved in the processing of pre-rRNA synthesized by Pol I. Noteworthy, we observed an accumulation of *d*NET-seq signal at the end of U6 snRNA (Figure 3.10 - figure supplement 1A), contrasting with a lack of peak observed in mammalian cells (Nojima et al., 2018). This is consistent with the finding that mammalian U6 snRNAs contain a 2',3′-cyclic phosphate terminal group at the 3′ end, whereas U6 3′ ends in *Drosophila* embryos consist of either P or OH (Lund and Dahlberg, 1992).

To quantify how many constitutively spliced exons have *d*NET-seq peaks at the end, we applied an algorithm that finds nucleotides where the NET-seq read density is at least three standard deviations above the transcript mean in a local region defined by 100 bp upstream and downstream (Churchman and Weissman, 2011a; Prudêncio et al., 2019). Upon analyzing replicates of *d*NET-seq/S5P and *d*NET-seq/S2P libraries, we identified over 10,000 exons showing splicing intermediate peaks (Figure 3.10D). As peaks were more frequently detected on exons of genes with higher read density (Figure 3.10 - figure supplement 1B), we classified the exons into four groups (quartiles) based on the *d*NET-seq read density of the corresponding gene and restricted the analysis to exons in the fourth quartile, i.e., from genes with the highest read density. The results show that splicing intermediate peaks are detected in approximately 80% of all constitutively spliced exons. The proportion is similar for pre-MBT genes and for genes expressed in late embryos (Figure 3.10E), as well as for both the S5P and the S2P datasets (Figure 3.10E). We then used the same methodology and the same set of genes to detect peaks at the last intronic nucleotide, corresponding to released intron lariats. Such peaks were detected in less than 10% of introns (Figure 3.10F).

Taken together, these results demonstrate that in *Drosophila* embryos, co-transcriptional splicing is associated with Pol II phosphorylated at the CTD serine 5 position, as previously reported in mammalian cells (Nojima et al., 2015a, 2018; Schlackow et al., 2017).

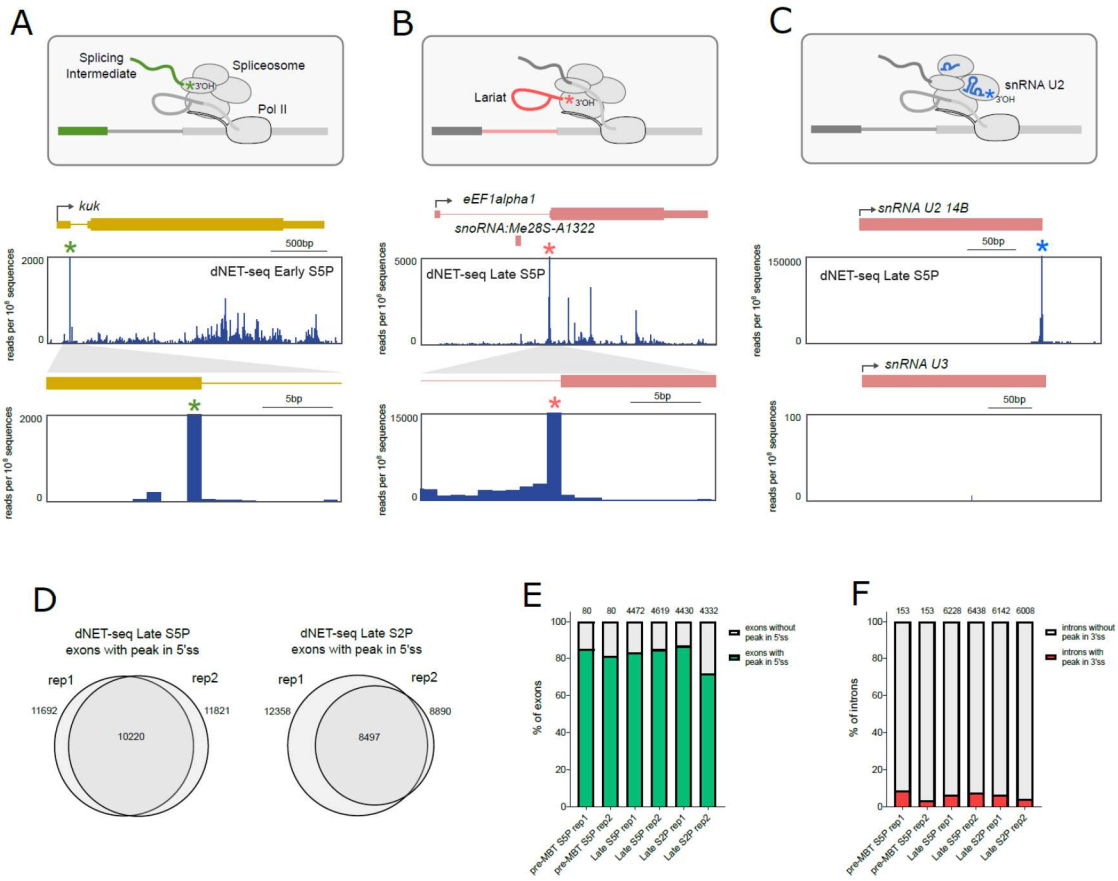



**Figure 3.10. Co-transcriptional splicing associated with Pol II phosphorylated on CTD serine 5. (A-C)** *d*NET-seq/S5P profiles over the indicated genes in late embryos. The diagrams outline the 3′ OH ends generated by co-transcriptional cleavage at the 5′ splice site (green, **A**), the 3′ splice site (red, **B**) and the free 3′ OH end of spliceosomal snRNAs (blue, **C**). Arrows indicate the direction of transcription. Exons are represented by boxes. Thinner boxes represent UTRs. Introns are represented by lines connecting the exons. Asterisks denote the peak at the end of the exon (**A**), the end of the intron (**B**), and the end of the U2 snRNA gene (**C**). (**D**) Comparison (Venn diagrams) of exons with a splicing intermediate peak detected in biological replicates of *d*NET-seq/S5P and *d*NET-seq/S2P libraries. (**E, F**) Frequency of peaks corresponding to splicing intermediates (**E**) and released intron lariats (**F**) in pre-MBT genes and genes expressed in late embryos. Only genes with the highest read density (fourth quartile) were considered.

### 3.3.4. *d*NET-seq specifically captures nascent RNA.

To validate that we were detecting nascent transcription in *Drosophila* embryos, we analyzed maternal genes that are transcribed during oogenesis and loaded into the egg (Figure 3.11A). As expected, maternal transcripts such as *bicoid* (Figure 3.11B), *nanos* (Figure 3.11 - figure supplement 1A), *gurken* (Figure 3.11 - figure supplement 1B) and *Rab32* (Figure 3.11 - figure supplement 1C) were detected by RNA-seq in embryos collected 2-3 hours after

fertilization, but *d*NET-seq signal over these genes was negligible. However, robust *d*NET-seq signal was found at the *pumilio* gene (Figure 3.11 - figure supplement 1D). This gene is expressed maternally, and its function is essential during early embryogenesis for the formation of abdominal segments (Barker et al., 1992). *pumilio* was considered purely maternal and not maternal-zygotic expressed based on RNA-seq data (Lott et al., 2011a) and Pol II Chip-seq data (Chen et al., 2013). However, *pumilio* expression is detected in 2-2.5 hours post-fertilisation embryos with GRO-seq (Saunders et al., 2013). This means that *d*NET-seq is equally sensitive to global run-on methodologies without requiring RNA labelling procedures.

Strong *d*NET-seq signal was detected in early embryos over the bodies of zygotic genes that become transcriptionally active before MBT, such as *snail* (Figure 3.11C), *fushi tarazu* (Figure 3.11 - figure supplement 1E) and *odd skipped* (Figure 3.11 - figure supplement 1F). Altogether, we examined 117 previously identified pre-MBT genes, i.e. genes that start to be transcribed before cycle 14 and remain active thereafter (Chen et al., 2013). A characteristic feature of pre-MBT genes is that they are short in length with few if any introns (Ali-Murthy et al., 2013b; Heyn et al., 2014a; Kwasnieski et al., 2019b; De Renzis et al., 2007b). In contrast, most genes expressed in late embryos contain multiple introns, as shown for *Akap200* (Figure 3.11D), *His3.3A* (Figure 3.11 - figure supplement 1G) and *twinstar* (Figure 3.11 - figure supplement 1H). Analysis of *d*NET-seq datasets from late embryos revealed widespread signal on introns and regions downstream of the polyadenylation site, further confirming that the technique is capturing nascent transcripts.

To conclude, robust *d*NET-seq signal was recovered from zygotically but not from maternally expressed transcripts. In addition, we detected reads mapping to introns and downstream of the poly-adenylation site. These findings suggest that *d*NET-seq is indeed specifically targeting the nascent transcriptome.

### 3.3.5. *d*NET-seq reveals transcriptional read-through associated with the presence of overlapping antisense genes

Having confirmed that *d*NET-seq was capturing nascent RNA, we next investigated the distribution of Pol II density over transcript regions. The *d*NET-seq profiles over individual genes in early and late embryos (Figure 3.11C, D and Figure 3.11 - figure supplement 1) do not show the characteristic higher read density near the promoter, as previously described in

mammalian cells (Mayer et al., 2015; Nojima et al., 2015a). This is most likely because Pol II typically pauses ~30-60 bp downstream of the transcription start site (TSS) (Kwak et al., 2013) and in our *d*NET-seq approach we enrich for RNAs longer than 60 nt (Figure 3.13B); thus, we only record the position of polymerases that have transcribed at least 60 bp past the TSS.

We then turned our attention to the *d*NET-seq signal around polyadenylation sites (also known as the transcription end sites or TES). Many genes, such as *Akap200* (Figure 3.11D), *His3.3A* (Figure 3.11 - figure supplement 1G) and *tsr* (Figure 3.11 - figure supplement 1H), have significant *d*NET-seq/S5P signal after the TES, indicative that in these cases, Pol II continues to transcribe after cleavage and polyadenylation. We noted that the 3′ UTR of these three genes is overlapped by another gene (*gurken, Nepl3,* and *IntS1,* respectively), which is transcribed in the opposite direction. We therefore asked whether transcriptional read through is related to the presence of an overlapping convergent gene. In order to identify all genes that are transcriptionally active in late embryos, we used a strategy adapted from GRO-seq analysis (Core et al., 2008b) that relies on read density in gene desert regions as background reference for absence of transcription. Very large intergenic regions (gene deserts) were divided into 500kb windows, and read densities were calculated by dividing read counts in each window by the window length in bp (Figure 3.11 - figure supplement 2A). Genes with *d*NET-seq signal over the gene body (in RPKM) above the 90th percentile of read density for all intergenic regions analyzed were considered to be transcriptionally active (Figure 3.11 - figure supplement 2B). We identified approximately 7 thousand active genes, and similar results were obtained from *d*NET-seq/S5P and *d*NET-seq/S2P datasets (Figure 3.11 - figure supplement 2C). This set of active genes includes over 85% of the 3500 active MBT genes previously annotated by ChIPseq (Chen et al., 2013). Next, we divided the genes transcribed in late embryos into two groups, depending on whether their 3′ UTR was or was not overlapped by another convergent gene. The metagene analysis shows that on all transcribed genes, the *d*NET-seq/S5P signal is abruptly reduced at the TES (Figure 3.11F). This is expected, as when Pol II reaches this site, the nascent transcript is cleaved and polyadenylated and therefore there is no RNA to be sequenced. However, the presence of nascent transcripts after the TES, indicating that Pol II fails to terminate transcription after cleavage and polyadenylation, is mainly observed on genes that have the 3′ UTR overlapped by an antisense gene (Figure 3.11F). Similar results were observed after metagene analysis of *d*NET-seq/S2P datasets (Figure 3.11G), confirming an association between transcriptional read-through and genomic architecture at the TES. We then asked whether transcriptional read-through depends on the

transcriptional activity of the overlapping antisense (convergent) gene. Using the methodology described above to identify transcribed genes, we found that the vast majority (>80%) of overlapping convergent genes were transcriptionally active and only 219 genes were silent. The metagene analysis shown in Figure 3.11H clearly indicates that transcription of the convergent overlapping gene is not required for inefficient termination at the overlapped TES.

The vast majority of pre-MBT genes do not show transcriptional read-through (Figure 3.11C and Figure 3.11 - figure supplement 1E, F). This is likely explained by the fact that for the majority of these genes (65), there is no other gene on either strand within a region of 500 bp downstream of the TES, as shown for *tailless (tll)* (Figure 3.11 - figure supplement 2D). Another 26 pre-MBT genes are embedded in larger genes, as shown for *nullo*, which is located within a long intron of the *CG12541* gene (Figure 3.11 - figure supplement 2E). A smaller group of pre-MBT genes (21) have neighboring genes located on either strand within a region of 500 bp downstream of the TES, as shown for *Elba2* (Figure 3.11 - figure supplement 2F). We further identified 5 pre-MBT genes that have the 3′ UTR overlapped by an antisense convergent gene, as shown for *spook (spo)* (Figure 3.11 - figure supplement 2G). Transcriptional read-through was mostly observed in this last group (Figure 3.11 - figure supplement 2G).

In conclusion, *d*NET-seq reveals a strong correlation between inefficient transcription termination at the cleavage and polyadenylation site, and presence of an antisense convergent gene overlapping the TES. The vast majority of pre-MBT genes transcribed in early embryos do not have the TES overlapped by other genes and are efficiently terminated, whereas many genes transcribed in late embryos show transcriptional read through.

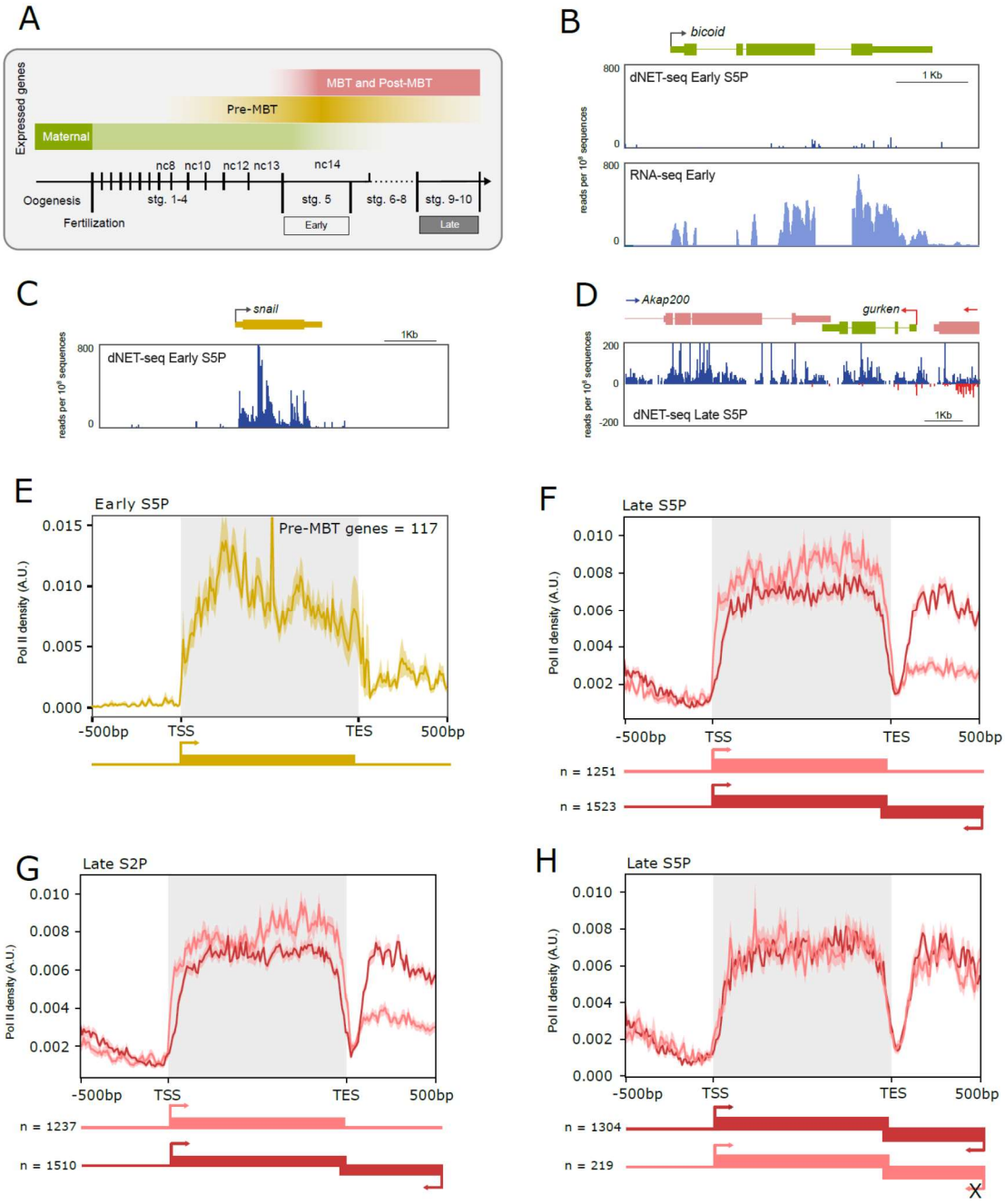



**Figure 3.11. Transcription over gene bodies and beyond the TES. (A)** The diagram illustrates the temporal expression of maternal, pre-MBT, MBT, and post-MBT genes during *Drosophila* embryonic development. Maternal genes are transcribed during oogenesis (green) and the resulting mRNAs persist during early embryogenesis (light green). Zygotic transcription starts before the midblastula transition (pre-MBT genes, yellow). Most genes become transcriptionally active during or after the midblastula transition (MBT and post-MBT, red). **(B-D)** *d*NET-seq/S5P and RNA-seq profiles over the maternal gene *bicoid* (**B**), the pre-MBT gene *snail* (**C**), and the post-MBT gene *Akap200* (**D**). Reads that aligned to the positive strand are in blue, and reads that aligned to the negative strand are in red. **(E-H)** Normalized metagene analysis in arbitrary units (A.U.). The *d*NET-seq signal is depicted along the normalized gene length (grey background), as well as 500 bp upstream of the transcription start site (TSS) and 500 bp downstream of the transcription end site (TES). **(E)** *d*NET-seq/S5P signal

on pre-MBT genes in early embryos. **(F)** *d*NET-seq/S5P signal on transcriptionally active genes in late embryos; the signal over genes that have the 3′ UTR overlapped by an antisense gene is depicted in dark red, while the signal over genes with a non-overlapped 3′ UTR is depicted in light red. **(G)** *d*NET-seq/S2P signal on transcriptionally active genes in late embryos. **(H)** *d*NET-seq/S5P signal on transcriptionally active genes in late embryos; the signal over genes that have the 3′ UTR overlapped by a transcriptionally active antisense gene is depicted in dark red, while the signal over genes with the 3′ UTR overlapped by a transcriptionally inactive antisense gene is depicted in light red.

### 3.3.6. Evidence for Pol II pausing after the 3′ splice site

We next asked whether there was evidence of Pol II pausing around splice sites, which could suggest interactions between the splicing and transcription machineries. The number of nascent RNA reads whose 3′ ends map at a particular genomic position is proportional to the number of Pol II molecules at that position. Thus, Pol II pause sites can be detected as local peaks in read density (Larson et al., 2014). For this analysis, we excluded signal from co-transcriptional splicing (i.e., reads that map to the very last nucleotide of introns and exons were discarded, and the corresponding genomic positions were not considered). In addition, we specifically focused on the S5P data, as co-transcriptional splicing has been linked to S5P in humans (Nojima et al., 2015a). Many strong peaks of *d*NET-seq/S5P signal were observed over both exons and introns (Figure 3.12A).

Our next task was to identify such pause sites more systematically. A difficulty of this analysis is that transcripts with higher initiation rates will contain more reads and thus peaks are more likely to be detected than in more lowly transcribed genes. To control for such confound, we developed a peak calling algorithm that detects regions where the local read density is significantly higher than expected by chance, given the over-all read density of the transcript (Figure 3.12A; see also Methods). We emphasize that whereas the peak calling method used above for detecting significant splicing intermediate peaks (Churchman and Weissman, 2011a) looks for significant single-nucleotide positions, this method instead detects putative pause regions of variable length.

We then aligned exons and introns on the splice sites and calculated the average peak density at each position. Exons have a higher over-all peak density than introns (mean proportion of nucleotides in peaks ~0.020/~0.015 for replicate 1/replicate 2 introns and ~0.045/~0.034 for replicate 1/replicate 2 exons; one-tailed $p = 9.999 * 10^{-5}$ for both replicates;

permutation test with 10,000 iterations), suggesting that the elongation rate is decreased over exons. This is consistent with previous reports in mammalian (Jonkers et al., 2014b; Mayer et al., 2015) and *Drosophila* cells (Kwak et al., 2013) (Figure 3.12C). In addition, peak density is sharply increased around the 5′ splice site (Figure 3.12B). This could indicate Pol II pausing associated with splice site recognition. However, we cannot exclude that the pattern could also be due to misalignment of splicing intermediate reads that should map to the final exonic nucleotide. The profile around the 3′ splice site is more complex, with first an isolated high peak density region just after the 3′ splice site, and then a drastic increase in average peak density starting roughly 60 nt after the 3′ splice site (Figure 3.12B). It thus appears that Pol II elongation behavior indeed covaries with the locations of exon-intron boundaries.

A potential caveat is that nucleotide composition varies systematically across exons and introns. For example, exons tend to have a higher GC content than introns (Amit et al., 2012b). This could be problematic as NET-seq relies on MNase digestion of DNA and RNA to solubilize chromatin. MNase digestion of DNA is known to be sequence-biased, with most notably a preference for cleaving just 5′ of an adenine (Dingwall et al., 1981; Gaffney et al., 2012; Hörz and Altenburger, 1981). An analysis of the 5′ ends of our reads revealed similar biases for MNase digestion of RNA (Figure 3.12D). This sequence preference could lead to artefactual variation in read density, with more reads being sampled from transcripts and transcript regions whose nucleotide composition is more similar to MNase digestion biases. To verify to what extent our results were affected by this confound, we performed a simulation to determine the expected distribution of reads based on the digestion bias alone (Supplementary methods). We concluded that MNase biases are unlikely to explain either the enrichment of peaks in exons or the high peak densities observed just around the splice sites. However, it is possible that the exact location and size of the large peak starting ~60 nt into the exon is affected by the bias (Supplementary methods).

In conclusion, we detect both a general decrease in elongation rates over exons as well as pausing in the vicinity of the splice sites.

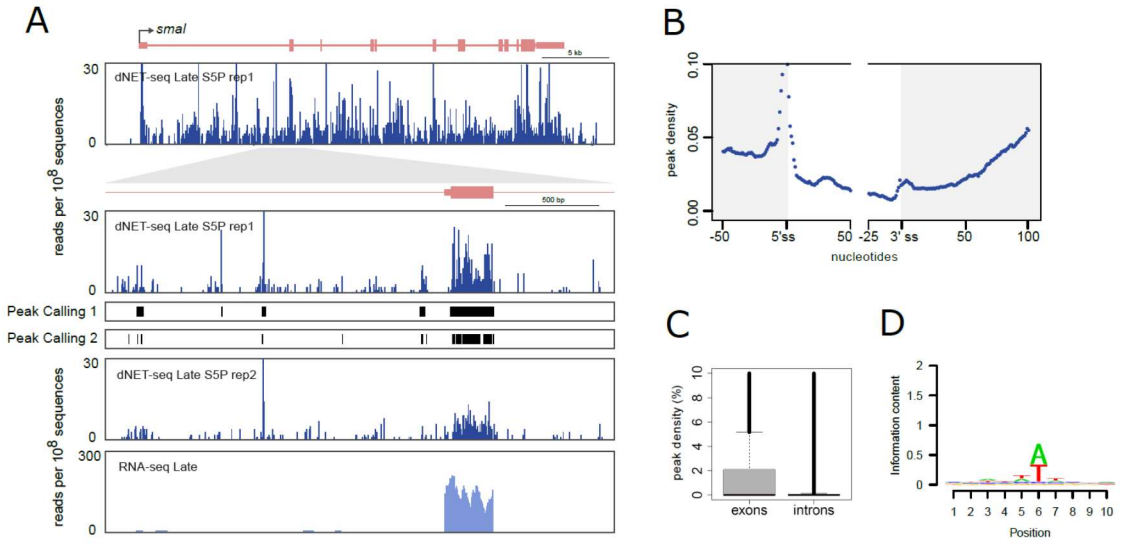



**Figure 3.12. Evidence for Pol II pausing after the 3′ splice site.** (**A**) *d*NET-seq/S5P and RNA-seq profiles over the post-MBT gene *smoke alarm* (*smal*). The number of NET-seq reads is depicted at two magnification levels in two biological replicates. For replicate 1, peaks called by our in-house peak caller are also shown. The line "Peak Caller 1" shows peaks called using the "large peaks" setting, which is appropriate for detecting larger regions of putative Pol II pausing. The line "Peak Caller 2" shows peaks called using the "small peaks" setting, which provides higher spatial resolution and has been used for the analyses presented in this paper. See Methods for further details on peak caller settings. RNA-seq data for the same regions is also shown. (**B**) Metagene analysis of Pol II peak density estimated from *d*NET-seq/S5P datasets from late embryos (replicate 1). Peak density has been calculated as the proportion of introns that overlap with a peak at any given position. The last 50 nucleotides of exons, the first 50 nucleotides of introns, the last 25 nucleotides of introns, and the first 100 nucleotides of exons are shown. Only internal and fully coding exons from transcriptionally active genes that are at least 100 nucleotides long are shown. Exons shorter than 150 nucleotides contribute to both the exon end and start. Only intorns that were at least 50 nt long were considered. **(4C)** Peak density in the exons and introns of transcriptionally active genes (*d*NET-seq/S5P, replicate 1). Peak density has been defined as the percentage of nucleotides within a given exon or intron that overlap with a significant peak. Exons have a higher over-all peak density than introns (mean percentage of nucleotides in peaks ~2%/~1.5% for replicate 1/replicate 2 introns and ~4.5%/~3.4% for replicate 1/replicate 2 exons; one-tailed $p$ = 9.999 * $10^{-5}$ for both replicates; permutation test with 10,000 iterations). See Methods for more details. **(4D)** Sequence logo of nucleotide frequencies within a 10-nt window around the 5′ ends of NET-seq reads. The combined height of the bases at each position is proportional to the information content at that position. Position 6 corresponds to the 5′-most nucleotide of the read. Putative internal priming reads, as well as reads mapping to the last nucleotide of exons or introns (possible splice intermediate and intron lariat reads, respectively) were ignored.

### 3.3.7. *d*NET-seq captures recursive splicing intermediates.

Having shown that the spliceosome forms a complex with the elongating Pol II in *Drosophila* embryos and that Pol II pausing is correlated with position relative to the splice sites, we asked when splicing takes place relative to transcription. We first looked at recursive splicing of long introns because this process involves the formation of inherently unstable intermediates that are more likely to be formed soon after the transcription of each intronic splice site (Pai et al., 2018b). In recursive splicing, long introns are removed by sequential excision of adjacent sections involving separate splicing reactions, each producing a distinct lariat (Hatton et al., 1998b). Recursively spliced intron segments are bounded at one or both ends by recursive sites or ratchet points (Burnette et al., 2005), which correspond to zero nucleotide exons consisting of juxtaposed 3′ and 5′ splice sites around a central AG|GT motif, where the vertical line represents the splice junction (Figure 3.13A).

To capture recursive splicing intermediates using *d*NET-seq, it is essential to have a good coverage of reads corresponding to nascent transcripts and spanning the splice junctions. The total number of reads resulting from nascent RNA in each *d*NET-seq dataset is depicted in Figure 3.13 - figure supplement 1A. By merging the sequencing information of overlapped paired-end reads (Figure 3.9E), we were able to sequence on average ~103 nucleotides per nascent RNA (Figure 3.13B and Figure 3.13 - figure supplement 1B). Focusing on previously identified *Drosophila* ratchet points (Duff et al., 2015; Joseph et al., 2018a), we found *d*NET-seq/S5P reads that span the junction between the canonical 5′ splice site at the end of the exon and the first ratchet point (RP1) internal to the downstream intron, as shown for the second intron of the *Megalin* gene (Figure 3.13C). Reads spanning the subsequent intronic RPs were also observed (Figure 3.13C). Overall, we detected *d*NET-seq/S5P and *d*NET-seq/S2P spliced reads supporting most of the previously identified recursive splicing events (Figure 3.13D and Figure 3.13 - figure supplement 1C).

Analysis of *d*NET-seq profiles around a RP reveals an enrichment of reads in a region located a few nucleotides downstream of the RP, as shown for RP2 in the first intron of the *Tenascin major* gene (Figure 3.13E). Noteworthy, most of these reads are already spliced to the previous RP (Figure 3.13E). A meta-analysis of *d*NET-seq/S5P reads around 137 RPs confirms that many spliced reads can be observed just downstream of RPs (Figure 3.13F, G). Taken together, these results indicate that recursive splicing can occur soon after the

transcription of intronic recursive sites and suggest the presence of paused Pol II downstream of these sites.

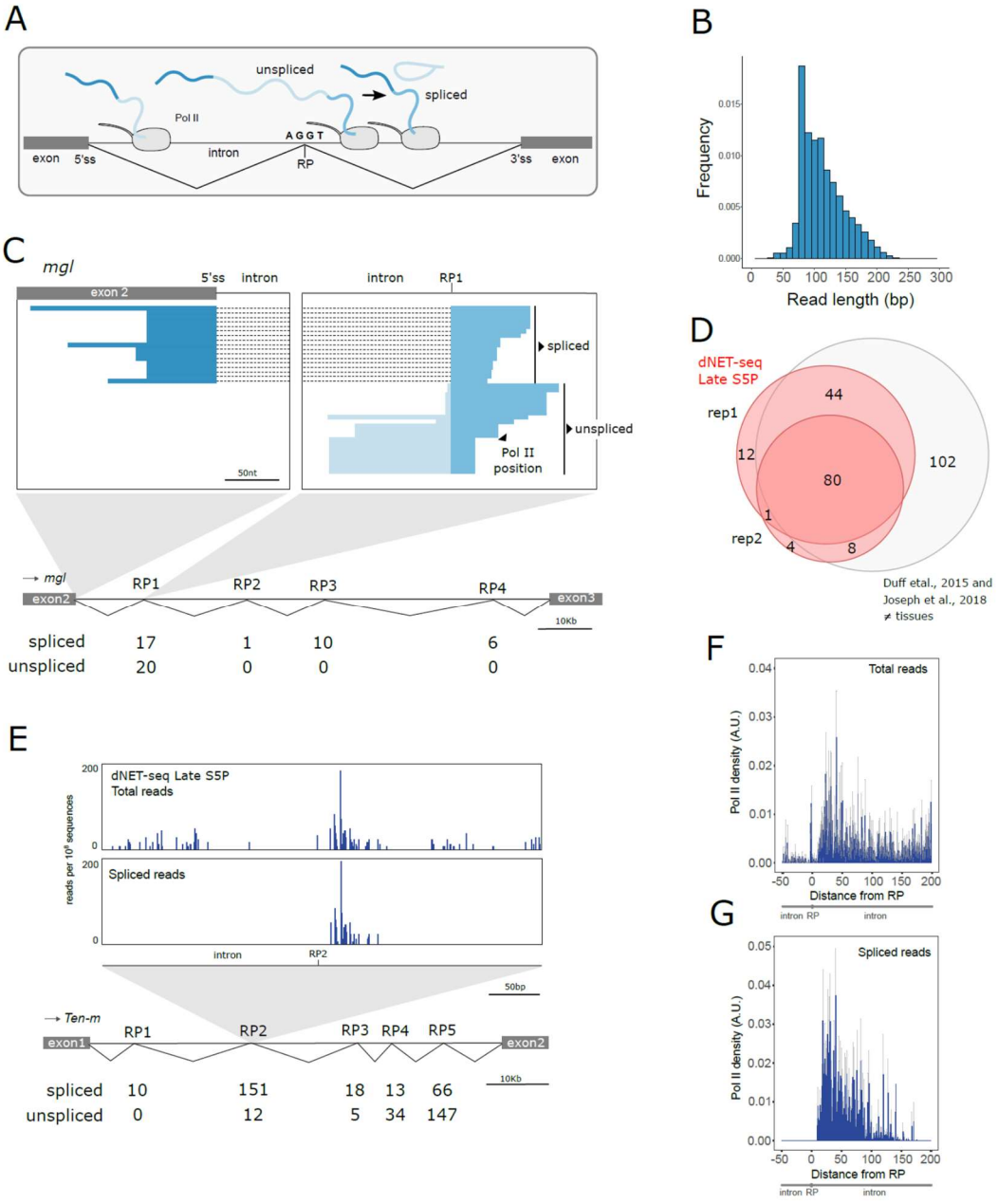



**Figure 3.13. *d*NET-seq captures recursive splicing intermediates. (A)** Schematic illustrating recursive splicing. A ratchet point (RP) with juxtaposed acceptor and donor splice site motifs is indicated. **(B)** Histogram of the read lengths in *d*NET-seq/S5P data from late embryos. **(C)** Visualization of *d*NET-seq/S5P reads that align to the second intron of the *Megalin* (*mgl*) gene. Recursively spliced reads align to exon 2 (green) and the intron after RP1 (dark red). The number of spliced and unspliced reads at each RP in the intron is indicated. **(D)** Venn diagram comparing RPs identified in two *d*NET-seq/S5P biological replicates and in previously reported studies (Duff et al., 2015; Joseph et al., 2018a). **(E)** Number of *d*NET-

seq/S5P reads that have the 3′ end mapped around RP2 in the first intron of *Tenascin major* (*Ten-m*) gene. The top panel depicts all reads, and the bottom panel depicts only reads that have been spliced to RP1. **(F-G)** Meta-analysis with single nucleotide resolution of normalized *d*NET-seq/S5P reads around RPs (n=137) using all reads **(E)** or only reads spliced to the previous RP or exon **(F).**

### 3.3.8. Splicing takes place as Pol II transcribes past the 3′ splice site, yet many nascent transcripts remain unspliced

Reads that span exon-exon junctions corresponding to spliced nascent transcripts were found in both *d*NET-seq/S5P and *d*NET-seq/S2P libraries prepared from early and late embryos (Figure 3.14A). To identify which splicing events were detected by *d*NET-seq, we considered all internal and fully-coding exons that are at least 100 nt long in actively transcribed genes. For each splice junction, we counted how many reads had the 3′ end mapped to the first 100 nt of the exon. Only exons with at least 10 reads mapping to this region were considered (see Methods for justification of the threshold). Then, the *d*NET-seq splicing ratio (SR) was calculated by dividing the number of spliced reads by the sum of the number of spliced and unspliced reads (Figure 3.14A). Reads could be counted as spliced or unspliced if their 3′ end mapped to within the first 100 nt of the exon and their 5′ end reached upstream of the 3′ splice site, allowing to check whether the intron was still present. A robust agreement of estimated SRs was observed between biological replicates for pre-MBT (Figure Figure 3.14 - figure supplement 1B) and late (Figure 3.14B) genes.

SRs showed a bimodal distribution, with peaks at both extremes (SR = 0 and SR = 1; Figure 3.14D). Hence, for the remainder of the analysis, we performed all comparisons between three discrete SR classes (SR = 0, 0 < SR < 1, and SR = 1). We observed that only ~5% of splice junctions in pre-MBT genes were devoid of reads spanning ligated exons and thus presented an SR of 0 (~5.10% replicate 1/~4.35% replicate 2), whereas in late genes this proportion was ~20% (~19.76% replicate 1/~18.56% replicate 2) (Figure 3.14C; two-tailed binomial test for difference between late and pre-MBT, $P \sim 6.145 * 10^{-5}/9.439 * 10^{-4}$ (replicate 1/replicate 2)). However, this difference between pre-MBT and late genes disappeared once the higher read density of pre-MBT genes had been controlled for (Figure 3.14C).

Thus, in most cases, *Drosophila* co-transcriptional splicing can occur when Pol II is still transcribing the downstream exon, implying an intron definition mechanism as previously

proposed for *S. cerevisiae* (Carrillo Oesterreich et al., 2016). Consistent with this view, many *Drosophila* transcripts have relatively long exons separated by short introns – a gene architecture suggested to be conducive to intron definition. Very long introns flanked by short exons have instead been associated with exon definition (under which the downstream exon needs to be fully transcribed before splicing can take place) (Keren et al., 2010). It is unclear, however, whether the choice between exon and intron definition is dependent on the absolute sizes of exons and introns (for instance, (Fox-Walsh et al., 2005) proposed a switch to exon definition once the size of the intron surpasses ~200 nt), or rather the ratio of intron to exon size.

We found no clear relationship between the *d*NET-seq splicing ratio and exon length in either pre-MBT or late genes (Figure 3.14 - figure supplement 1C, G for replicates 1 and 2). Regarding intron size, we found that introns with SR = 0 (and thus no evidence for intron definition) were, on average, indeed larger than other introns (Figure 3.14E and Figure 3.14 - figure supplement 1F). However, more careful examination revealed a more complex picture, with a lower proportion of introns with SR = 0 both for introns of intermediate size (~55-100 nt, which corresponds to ~55% of the introns studied) and for very large introns (>1000 nt) (Figure 3.14E). These intron sizes may thus be optimal for fast splicing. We also uncovered a significantly lower exon to intron length ratio for introns with SR = 0 than for others (Figure 3.14 - figure supplement 1D, H).

Taken together, these results suggest that although fast splicing (implying intron definition) is indeed skewed towards small introns flanked by large exons, there is also frequent and efficient intron definition for large introns. Our results are inconsistent with a threshold model, where splicing would systematically switch to exon definition after a given intron size is reached.

We also investigated the relationship between SR and several other gene architecture parameters. Firstly, as the GC content in exons and introns decreases, the proportion of introns with SR = 1 increases and the proportion with SR = 0 decreases, showing more efficient immediate splicing (Figure 3.14F, G and Figure 3.14 - figure supplement 1J, K). A similar effect is observed as the ratio of the downstream exon GC content to intron GC content increases (Figure 3.14H and Figure 3.14 - figure supplement 1L). Thus, the most efficient immediate splicing is observed when the GC content is low in both exons and introns but higher in exons than introns. Secondly, similarly to previous reports (Herzel et al., 2018; Khodor et

al., 2011a, 2012a), we uncovered an effect of exon rank, whereby exons that are more central appear to be spliced more efficiently (Figure 3.14J and Figure 3.14 - figure supplement 1N). Thirdly, as 5′ splice site strength increases, the proportion of introns with SR = 0 decreases (Figure 3.14I and Figure 3.14 - figure supplement 1E, I, M). although this effect is stronger for the 3’ splice site (Figures 3.14I and Figure 3.14 - figure supplement 1M) than for the 5’ splice site (Figure 3.14 - figure supplement 1E, I). Finally, as previously observed (Khodor et al., 2012a), transcripts with only a single intron seemed to be spliced less efficiently than multi-intron ones, with a higher proportion of SR = 0 introns (Figure 3.14K and Figure 3.14 - figure supplement 1O).

In conclusion, *d*NET-seq reveals that splicing can occur as Pol II transcribes past the 3′ splice site, yet many nascent transcripts remain unspliced.

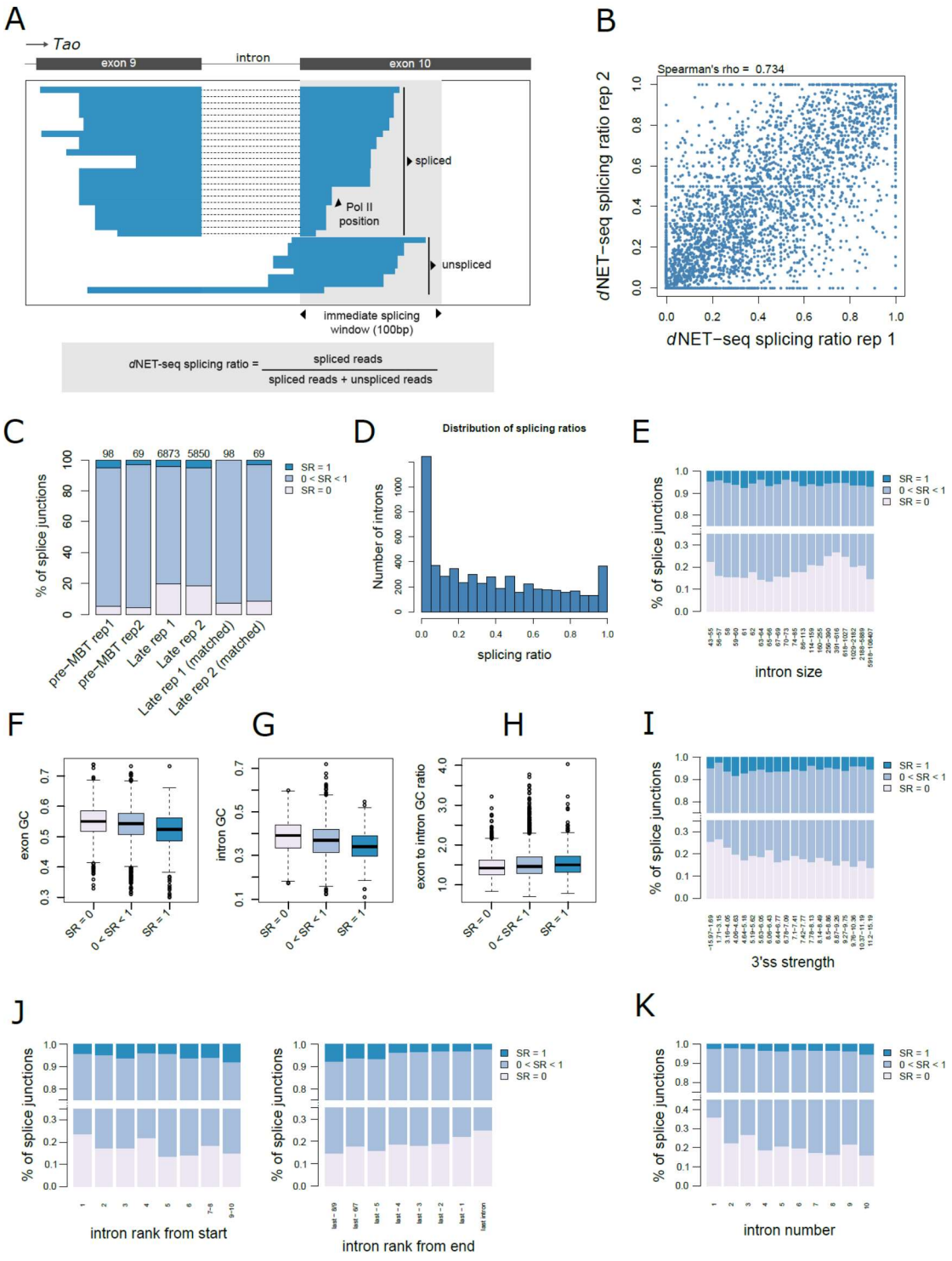



**Figure 3.14. Splicing occurs as Pol II transcribes past the 3′ splice site. (A)** Visualization of *d*NET-seq/S5P reads that align to exon 10 of the *Tao* gene. For **(B-J)**, unless otherwise specified, only introns from transcriptionally active genes where the downstream exon is a fully coding internal exon at least 100 nt long were included. In addition, enough spliced/unspliced reads had to end within the first 100 exonic nucleotides that obtaining a splicing ratio (SR) of 0 or 1 by chance alone was highly unlikely (see Methods for details). For late genes, this threshold was 10 reads for both replicates. For early genes, it was 14 for replicate 1 and 9 for replicate 2. **(B)** Values of *d*NET-seq SRs estimated in two biological replicates of *d*NET-

seq/S5P datasets from late embryos (Spearman correlation, $\rho = \sim 0.734$, $P < 2.2 * 10^{-16}$). **(C)** Splice junctions in pre-MBT genes and genes expressed in late embryos classified according to their *d*NET-seq splicing ratio. As many pre-MBT genes are single-intron, last introns were exceptionally included in this analysis. For **(D-I)**, N = 5637, with 1048 introns with SR = 0 and 306 introns with SR = 1. **(D)** Histogram of SRs for *d*NET-seq/S5P. **(E)** Introns with an SR of 0 are significantly larger than other introns (two-tailed Mann-Whitney *U*-test, W = 2561204, $P = 9.862*10^{-4}$). However, the figure shows a more complex pattern, with potentially several optimal size ranges. There is no significant size difference between introns with SR = 1 and other introns (two-tailed Mann-Whitney *U*-test, W = 783087, P = ~0.240). **(F-G)** Introns with higher SRs have significantly lower exonic **(F)** and intronic **(G)** GC content (one-way ANOVA; exonic GC: $F = 37.85$, $P < 2 * 10^{-16}$; intronic GC: $F = 50.23$, $P < 2 * 10^{-16}$). All pairwise comparisons between groups are significant with $P <= 4 * 10^{-6}$ (Tukey Honest Significant Differences). **(4H)** Introns with higher SRs have significantly higher ratios of exonic to intronic GC content (one-way ANOVA, $F = 18.98$, $P = 6.08 * 10^{-9}$). Pairwise comparisons were significant at $P <= 6.2 * 10^{-6}$, except for the comparison between SR = 1 and $0 < SR < 1$, which was non-significant at $P \sim 0.119$ (Tukey Honest Significant Differences). **(I)** Introns have been binned by 3′ splice site strength, as estimated by MaxEntScan. Introns with a SR of 0 have a significantly lower 3′ splice site score than other introns (two-tailed Mann-Whitney *U*-test, W = 2152620, $P = 1.147*10^{-7}$). There is no significant difference between introns with SR = 1 and other introns (Mann-Whitney *U*-test, W = 845776, P = ~0.276). However, the figure reveals a trend whereby introns with stronger 3′ splice sites tend to have a *lower* proportion of introns with SR = 1 than 3′ splice sites of intermediate strength. In **(J-K)**, last introns have been included in the analysis. **(4J)** Introns with SR = 0 are closer to the 3′ end of the transcript (two-tailed Mann-Whitney *U*-test, W = 2624472, $P = 1.886*10^{-5}$), whereas no significant effect is observed for introns with SR = 1 (two-tailed Mann-Whitney *U*-test, W = 971198, $P \sim 0.125$). Only the last 10 introns of transcripts were considered. As a result, N = 6362, with 1298 introns with SR = 0 and 235 introns with SR = 1. **(4K)** Introns with SR = 0 are unusually rare among single-intron genes. Only genes with 10 or fewer introns were considered. As a result, N = 5150, with 1111 introns with SR = 0 and 157 introns with SR = 1. In **(E, I-K)**, the bin ranges have been set so that intron numbers would be as equal possible between bins.

### 3.3.9. Pol II pausing associated with co-transcriptional splicing.

Analysis of individual *d*NET-seq profiles around constitutive splice junctions with a high splicing ratio frequently revealed an enrichment of reads downstream of the 3′ splice site, coincident with the appearance of spliced reads, as shown for the *cno* gene (Figure 3.15A). In contrast, profiles around a constitutive splice junction with SR = 0 had an accumulation of reads further along the exon, as shown for the *ND-51* gene (Figure 3.15B). A clearly distinct type of profile was observed on skipped exons, on which very few reads were observed, as shown for the *zip* gene (Figure 3.15C).

We next performed a meta-analysis of NET-seq peak densities over different exonic and intronic regions (Figure 3.15D-F). Introns with SR = 1 displayed a characteristic Pol II peak density profile, with a single region of high density ~10-20 nucleotides after the 3′ splice site (Figure 3.15F). This profile was strikingly different from introns with lower SR values, which showed two high density regions: a small and narrow peak immediately after the 3′ splice site and a broader and higher peak starting ~60 nucleotides into the exon (Figure 3.15D-E). We performed a separate meta-analysis of skipped exons. These exons showed no accumulation of Pol II. However, the sample size was too small for meaningful comparisons with included exons.

Taken together, these results indicate that, as observed for recursive splicing, Pol II pauses soon after the 3′ splice site when the splicing reaction is completed at that position. When no immediate splicing is detected by *d*NET-seq on constitutively spliced exons, Pol II pausing is still observed but it starts further away from the 3′ splice site, suggesting that splicing may occur as Pol II reaches that position. It is unclear how to interpret the small increase in peak density immediately after the 3′ splice site for introns with SR = 0.

Finally, we asked how prevalent co-transcriptional splicing is in the developing *Drosophila* embryo. As shown in Figure 3.10E, approximately 80% of exons in pre-MBT and late genes covered by a high density of NET-seq reads have a peak corresponding to the splicing intermediate formed after cleavage at the 5′ splice site but before exon-exon ligation. The majority of these exons (~85%) were covered by *d*NET-seq reads that span the junction to the downstream exon either directly on nascent transcripts or indirectly on splicing intermediates formed by cleavage at the 5′ splice site of the downstream exon (Figure 3.15G), confirming that they are co-transcriptionally spliced. We then focused on those exons for which no splicing intermediate spike was detected by the peak calling algorithm (Figure 3.15H). Over 71% of these exons were also covered by *d*NET-seq reads that span the junction to the downstream exon either directly on nascent transcripts or indirectly on splicing intermediates formed by cleavage at the 5′ splice site of the downstream exon (Figure 3.15H), arguing that even exons without a detectable splicing intermediate peak are co-transcriptionally spliced. Altogether, our *d*NET-seq results support co-transcriptional splicing for over 95% of the analysed exons.

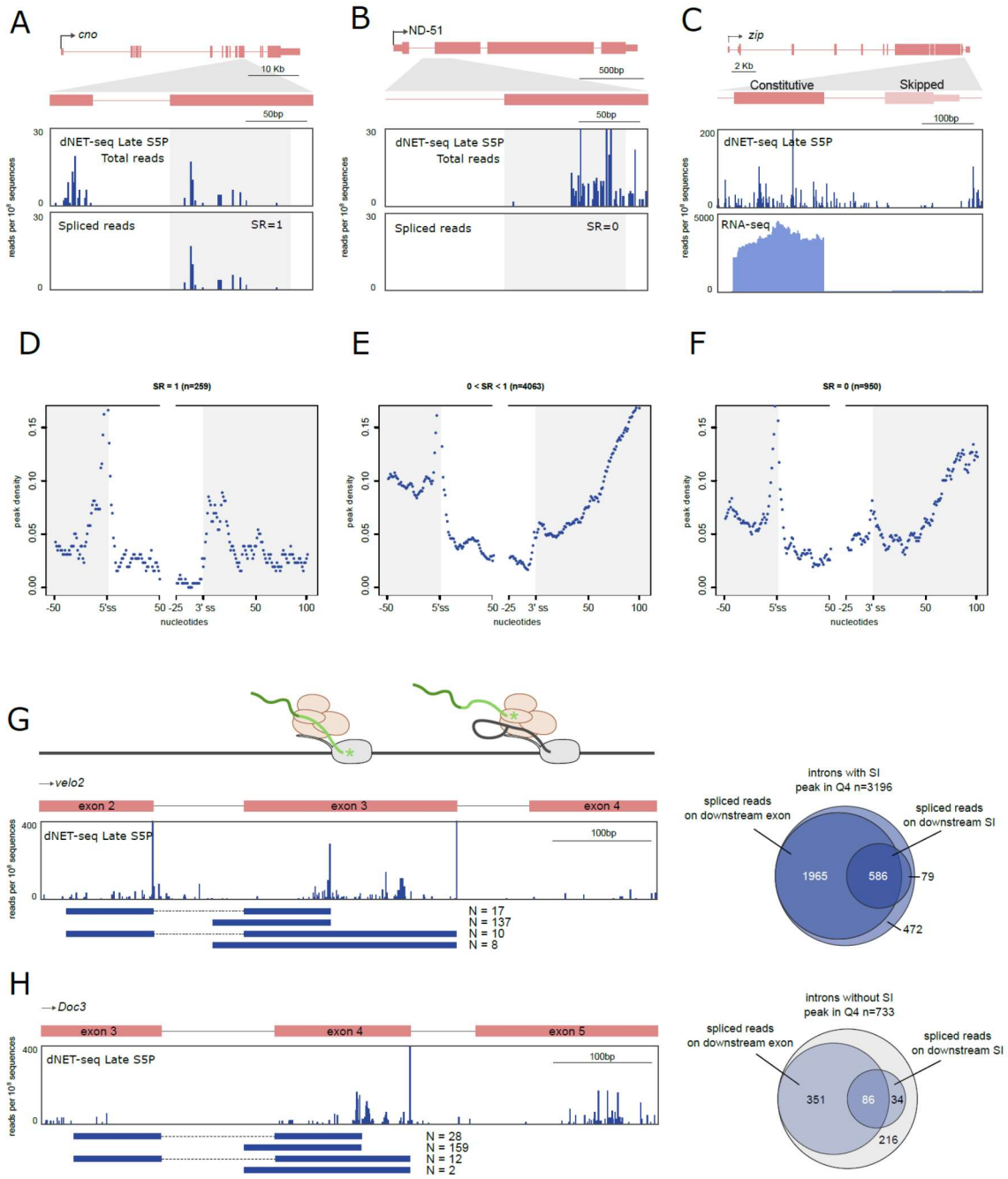



**Figure 3.15. Pol II pausing is associated with co-transcriptional splicing. (A-C)** *d*NET-seq/S5P profiles surrounding the indicated exons in the late genes *cno* (**A**), *ND-51* (**B**), and *zip* (**C**). The top panels depict all reads. The bottom panels depict either the 3′ end coordinate of reads that span the splice junction **(A, B)**, or the RNA-seq profile (**C**). **(D-F)** Metagene analysis of Pol II peak density estimated from *d*NET-seq/S5P datasets from late embryos (replicate 1). Peak density has been calculated as the proportion of introns that overlap with a peak at any given position. The last 50 nucleotides of exons, the first 50 nucleotides of introns, the last 25 nucleotides of introns, and the first 100 nucleotides of exons are shown. Only internal and fully coding exons from transcriptionally active genes that are at least 100 nucleotides long are shown. In addition, at least 10 spliced/unspliced reads had to end within the first 100 nucleotides of the exon. The introns have been split between SR = 0 **(D)**, 0 < SR < 1 **(E)** and SR = 1 **(F)**. **(G)** *d*NET-seq/S5P profiles on the indicated regions of the *velo2* and *Doc3* genes.

Below, spliced reads are depicted. Asterisks denote the splicing intermediate peaks at the end of exons. **(H)** Venn diagrams showing how many junctions with or without a splicing intermediate peak are covered by spliced reads or have a downstream splicing intermediate covered by spliced reads.

A

| Dataset | Raw reads | Uniquely Mapped after Merge |
|---|---|---|
| Early_2-3h_S5P_rep1 | 62.891.575 | 12.959.070 |
| Early_2-3h_S5P_rep2 | 56.917.697 | 3.610.607 |
| Late_4-6h_S5P_rep1 | 65.865.258 | 21.564.673 |
| Late_4-6h_S5P_rep2 | 75.909.131 | 27.375.998 |
| Late_4-6h_S5P_rep3 | 71.923.876 | 11.987.034 |
| Late_4-6h_S2P_rep1 | 77.838.463 | 38.595.780 |
| Late_4-6h_S2P_rep2 | 64.862.996 | 23.137.529 |
| Late_4-6h_S2P_rep3 | 76.719.886 | 17.912.504 |

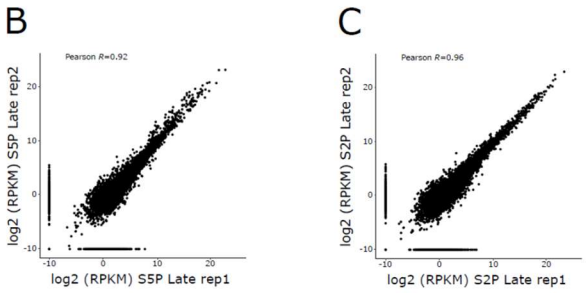



**Figure 3.9—figure supplement 1**. **(A)** For each *d*NET-seq library prepared, the number of total reads and uniquely mapped reads is indicated. **(B, C)** Density of uniquely aligned reads per gene (RPKM in log2 scale) for two *d*NET-seq/S5P and *d*NET-seq/S2P biological replicates from late embryos (Pearson's correlation, R = 0.92; R = 0.96).

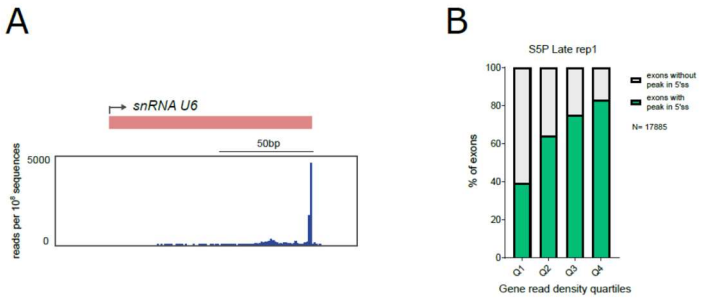



**Figure 3.10—figure supplement 1.** (**A**) *d*NET-seq/S5P profiles over the U6 snRNA gene in early and late embryos. (**B**) Frequency of peaks corresponding to splicing intermediates (green) detected by *d*NET-seq/S5P on exons of genes expressed in late embryos. Genes were grouped into quartiles (Q) based on their *d*NET-seq read density.

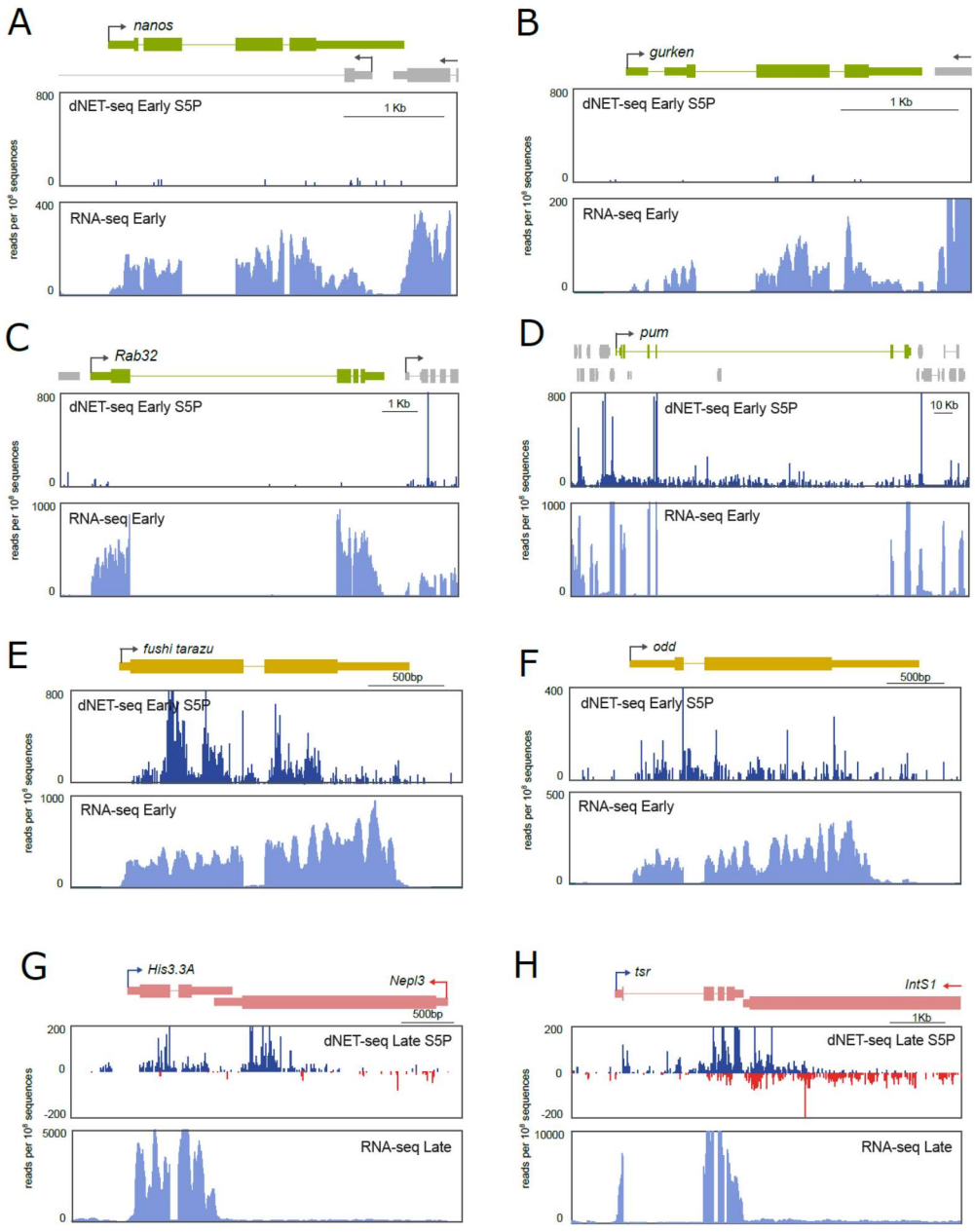



**Figure 3.11—figure supplement 1**. **(A-D)** *d*NET-seq/S5P and RNA-seq profiles over maternal genes in early embryos: *nanos* (**A**), gurken (**B**), *Rab32* (**C**) and *pumilio (pum)* (**D**). **(E-F)** *d*NET-seq/S5P and RNA-seq profiles over pre-MBT genes in early embryos: *fushi tarazu* (**E**) and *odd skipped (odd)* (**F**). **(G-H)** *d*NET-seq/S5P and RNA-seq profiles over MBT genes in late embryos: *His3.3A* (**G**) and *twinstar (tsr)* (**H**). Reads that aligned to the positive strand are in blue, and reads that aligned to the negative strand are in red. The direction of transcription is indicated by an arrow.

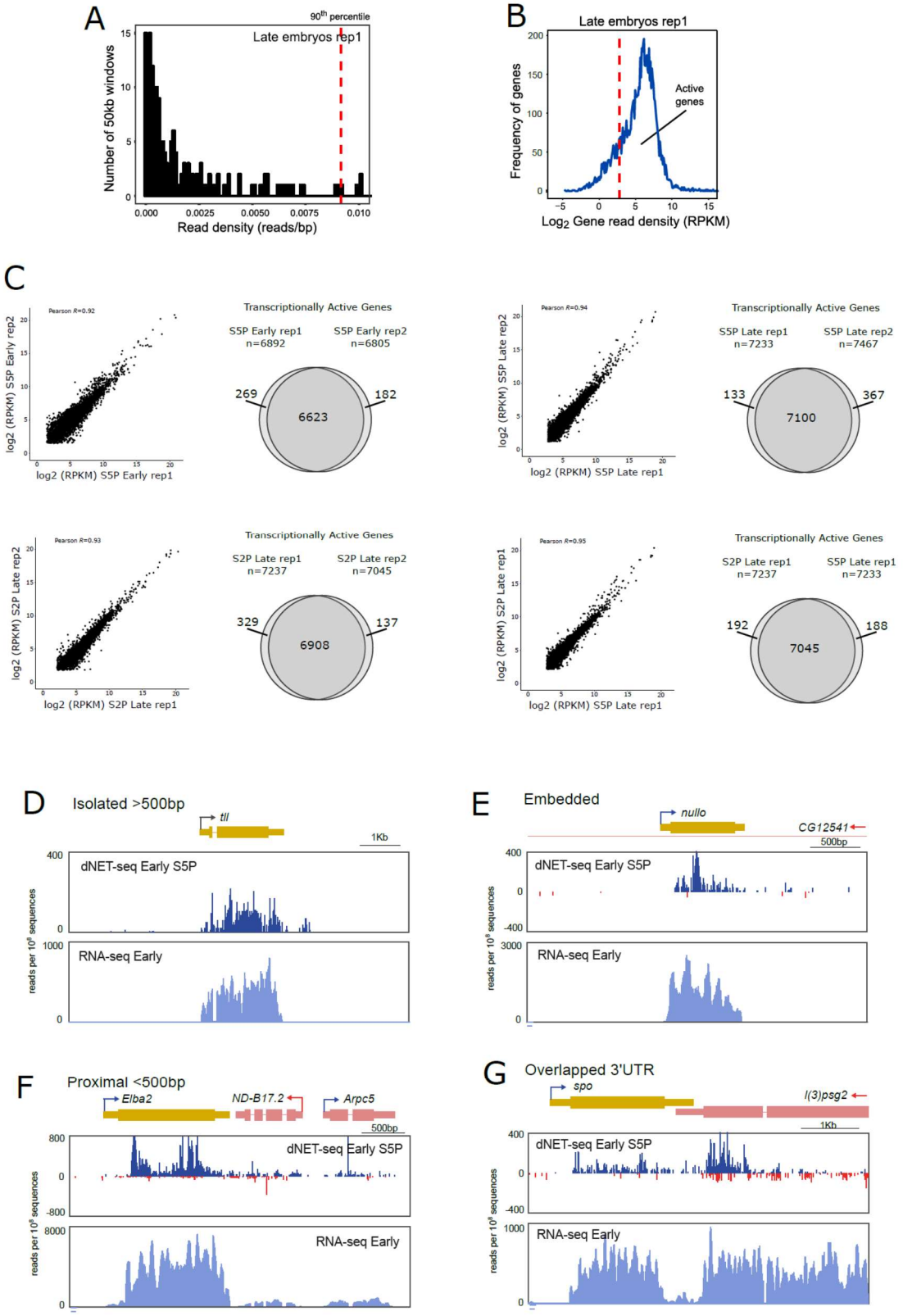



**Figure 3.11—figure supplement 2**. **(A-C)** Identification of transcriptionally active genes based on *d*NET-seq signal along the gene body in late embryos. **(A)** Read density distribution in intergenic regions (gene deserts). The red dashed line represents the 90th percentile of read density in all intergenic regions analyzed. **(B)** Read density distribution per gene (RPKM) represented in Log2 scale. The 90th percentile of read density over gene deserts is set as

threshold (red dashed line). **(C)** Comparison of read density per active gene (RPKM in log2 scale) between the indicated datasets. Total number of active genes identified in each dataset is indicated (n). Venn diagrams show common active genes between datasets. **(D-G)** *d*NET-seq/S5P and RNA-seq profiles over the indicated pre-MBT genes. Reads that aligned to the positive strand are in blue, and reads that aligned to the negative strand are in red. The direction of transcription is indicated by an arrow.

A

| Dataset | Number of nascent reads | Nascent reads / Uniquely mapped reads (%) |
|---|---|---|
| Early_2-3h_S5P_rep1 | 7.984.411 | 61,6 |
| Early_2-3h_S5P_rep2 | 2.430.876 | 67,3 |
| Late_4-6h_S5P_rep1 | 13.413.045 | 62,2 |
| Late_4-6h_S5P_rep2 | 13.777.880 | 50,3 |
| Late_4-6h_S5P_rep3 | 8.505.056 | 71,0 |
| Late_4-6h_S2P_rep1 | 23.153.659 | 60,0 |
| Late_4-6h_S2P_rep2 | 12.926.707 | 55,9 |
| Late_4-6h_S2P_rep3 | 10.356.743 | 57,8 |

B

| Dataset | min | mean | max | 25th percentile | 50th percentile | 75th percentile |
|---|---|---|---|---|---|---|
| Early_2-3h_S5P_rep1 | 35 | 99,6 | 293 | 74 | 81 | 119 |
| Early_2-3h_S5P_rep2 | 35 | 94,3 | 293 | 72 | 81 | 110 |
| Late_4-6h_S5P_rep1 | 35 | 114,2 | 293 | 86 | 108 | 135 |
| Late_4-6h_S5P_rep2 | 35 | 102,0 | 294 | 80 | 87 | 119 |
| Late_4-6h_S5P_rep3 | 35 | 88,4 | 294 | 76 | 81 | 89 |
| Late_4-6h_S2P_rep1 | 35 | 107,0 | 294 | 81 | 96 | 128 |
| Late_4-6h_S2P_rep2 | 35 | 111,1 | 294 | 81 | 102 | 132 |
| Late_4-6h_S2P_rep3 | 35 | 109,0 | 293 | 83 | 105 | 130 |

C

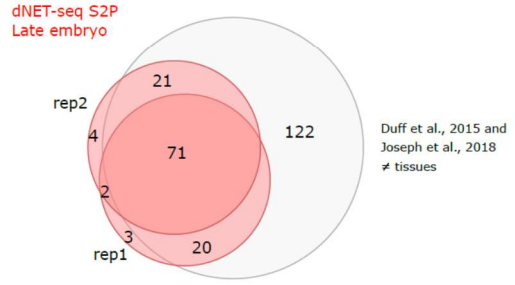



**Figure 13—figure supplement 1**. **(A)** For each *d*NET-seq library generated, the number of uniquely mapped reads corresponding to nascent transcripts is indicated. **(B)** Length of sequenced nascent RNA (in nucleotides). **(C)** Venn diagram comparing RPs identified in two *d*NET-seq/S2P biological replicates and in previously reported studies (Duff et al., 2015; Joseph et al., 2018b).

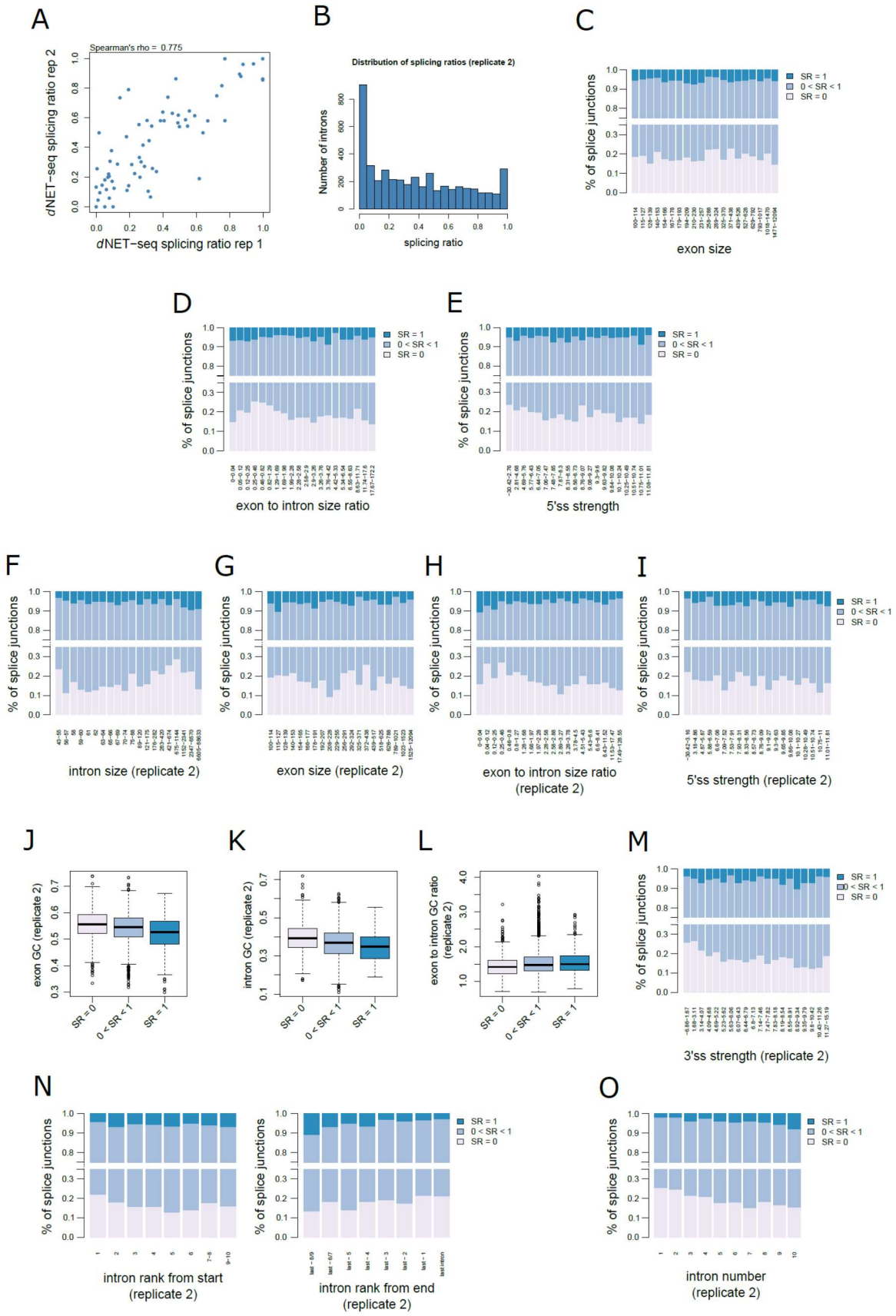



**Figure 14—figure supplement 1.** In all panels, unless otherwise specified, only introns from transcriptionally active genes where the downstream exon is a fully coding internal exon at least 100 nt long were included. In addition, enough spliced/unspliced reads had to end within

the first 100 exonic nucleotides that obtaining a splicing ratio (SR) of 0 or 1 by chance alone was highly unlikely (see Methods for details). For late genes, this threshold was 10 reads for both replicates. For early genes, it was 14 for replicate 1 and 9 for replicate 2. (**A**) Values of *d*NET-seq SRs estimated in two biological replicates of *d*NET-seq/S5P datasets for pre-MBT genes in the early data sets (Spearman correlation, $\rho = \sim 0.775$, $P \sim 1.348 * 10^{-14}$). Only introns where the downstream exon was at least 100 nt long and where at least 14 (replicate 1)/9 (replicate 2) spliced/unspliced reads mapped to the first 100 exonic nucleotides were included. (B-O) show *d*NET-seq/S5P data for the late datasets. **(B)** Histogram of SRs for *d*NET-seq/S5P (replicate 2). For **(C-E)** (replicate 1), N = 5637, with 1048 introns with SR = 0 and 306 introns with SR = 1. For **(F-M)** (replicate 2), N = 4511, with 797 introns with SR = 0 and 254 introns with SR = 1. (**C**, **G**) No relation between exon size and the proportion of introns with SR = 0 or SR = 1. **(D, H)** Introns with a lower ratio of the size of the downstream exon to the size of the intron have a higher proportion of SR = 0 (two-tailed Mann-Whitney *U*-test; replicate 1 W = 2251317, $P \sim 0.001$; replicate 2 W = 1332374, $P = 9.604*10^{-6}$). (**E, I**) Introns with an SR of 0 have (near-)significantly weaker 5′ splice sites than other introns (two-tailed Mann-Whitney *U*-test; replicate 1 W = 2278654, $P \sim 0.008$; replicate 2 W = 1415016, P = 0.051). (**F**) Introns with an SR of 0 are significantly larger than other introns (two-tailed Mann-Whitney *U*-test, W = 1625173, $P = 1.352*10^{-5}$). For replicate 2, introns with an SR of 1 are also significantly smaller than other introns (two-tailed Mann-Whitney *U*-test, W = 486616, P = ~0.007). **(J-K)** Introns with higher SRs have significantly lower exonic **(J)** and intronic **(K)** GC content (one-way ANOVA; exonic GC: F = 34.32, $P \sim 1.62 * 10^{-15}$; intronic GC: F = 44.37, $P < 2 * 10^{-16}$). All pairwise comparisons between groups are significant with P <= 0.001 (Tukey Honest Significant Differences). **(4L)** Introns with higher SRs have significantly higher ratios of exonic to intronic GC content (one-way ANOVA, F = 17.22, $P = 3.55 * 10^{-8}$). Pairwise comparisons were significant at $P <= 7.257 * 10^{-4}$, except for the comparison between SR = 1 and 0 < SR < 1, which was non-significant at $P \sim 0.820$ (Tukey Honest Significant Differences). **(M)** Introns have been binned by 3′ splice site strength, as estimated by MaxEntScan. Introns with a SR of 0 have a significantly lower 3′ splice site score than other introns (two-tailed Mann-Whitney *U*-test, W = 1315628, $P = 8.314*10^{-7}$). There is no significant difference between introns with SR = 1 and other introns (Mann-Whitney *U*-test, W = 521801, P = ~0.350). In **(N-O)**, last introns have been included in the analysis. **(4N)** Introns with SR = 0 are closer to the ends of the transcript (distance from start: two-tailed Mann-Whitney *U*-test, W = 876979, $P \sim 1.007*10^{-4}$; distance from end: two-tailed Mann-Whitney *U*-test, W = 2027956, $P \sim 0.028$), whereas no significant effect is observed for introns with SR = 1 (distance from start: two-tailed Mann-Whitney *U*-test, W = 512195, $P \sim 0.300$; distance from end: two-tailed Mann-Whitney *U*-test, W = 779548, $P \sim 0.654$). Only the first/last 10 introns of transcripts were considered. As a result, for distance from start, N = 4138, with 716 introns with SR = 0 and 234 introns with SR = 1. For distance from end, N = 5447, with 1039 introns with SR = 0 and 233 introns with SR = 1 (the N is different in the two cases because for distance from end, the 3′ most introns were included). **(4O)** Introns with SR = 0 are unusually common among single-intron genes. Only genes with 10 or fewer introns were considered. As a result, N = 4370, with 853 introns with SR = 0 and 169 introns with SR = 1. In **(C-I, M-O)**, the bin ranges have been set so that intron numbers would be as equal possible between bins.

### 3.3.10 Supplementary methods

**_d_NET-seq sequence biases**

Previous work has shown that MNase cleaves DNA more frequently at AT than at GC base pairs (Dingwall et al., 1981). More specifically, Hörz and Altenburger (1981) found a preference for sites where a few alternating AT-TA base pairs were surrounded by a more GC-rich region, rather than for sites within longer AT-rich stretches. More recent genome-wide work in humans has come to similar conclusions (Gaffney et al., 2012). *d*NET-seq relies on MNase cleavage of both DNA and RNA to solubilize chromatin. Any nucleotide biases in cleavage may be problematic, as they could mean that the solubilization is more efficient in regions where the nucleotide composition is more in line with such cleavage biases. This could lead to an artefactually increased read density in such regions.

To verify whether there was any evidence for RNA cleavage biases, we determined the nucleotide composition around the locations of the 5' ends of NET-seq reads (**Figure 4D** in main text). A clear preference for cleavage just 5' of an adenine was observed, with additional weaker biases at surrounding sites. Our results are therefore broadly in line with those obtained previously for DNA.

To what extent could these biases affect our results? We generated a simulated version of a *d*NET-seq dataset (S5P late, replicate 1) where the read positions were randomized in such a way as to preserve the nucleotide biases at the 5' ends of reads. Only reads that overlapped transcriptionally active genes were considered. More specifically, for each read, we defined the "starting hexamer" as the hexamer centred on the 5' end of the read (the three nucleotides just 5' of the read and the first three nucleotides of the read). We then picked a random position from among transcriptionally active genes where this hexamer also occurred and defined a "simulated read" at that position, preserving the length of the initial read. We then called peaks on the simulated reads similarly to true reads.

The mean proportion of nucleotides within simulated peaks was similar for both exons and introns (~0.017). However, this belies a difference in medians, with a higher peak density in exons than in introns (exon median: ~0.008; intron median: 0; $P < 2.2 * 10^{-16}$; two-tailed Mann-Whitney *U*-test; Figure supplementary 1). To know if the 5' end nucleotide bias effect could explain the bias towards exons observed in real data, we calculated a normalized peak

density for each exon and intron, as $\frac{true\ peak\ density - simulated\ peak\ density}{simulated\ peak\ density}$. Exons and introns with a simulated peak density of 0 were discarded. Normalized peak densities were still higher for exons than for introns ($P < 2.2 * 10^{-16}$; two-tailed Mann-Whitney *U*-test). Hence, 5' end nucleotide biases may exaggerate but do not explain the bias towards exons in real data.

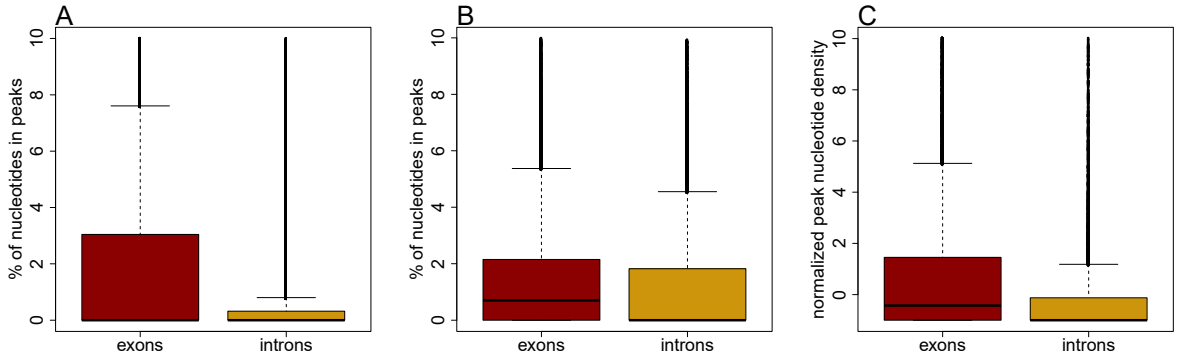



**Figure supplementary 1**: distribution of peak densities for exons and introns. A: true peaks. B: simulated peaks. C: normalized peaks, obtained as $\frac{true\ peak\ density - simulated\ peak\ density}{simulated\ peak\ density}$. For all three panels, the y axis has been limited to values below 10 for visualization reasons.

It may seem surprising that simulated read density would be higher in exons, given that *Drosophila* exons tend to have a higher GC content than introns (Zhu et al., 2009). Indeed, one would, at first sight, expect the bias to exaggerate intronic peak density, rather than exonic peak density. However, peaks are called based on the locations of the 3' ends of reads, whereas the bias affects the locations of the 5' ends of reads. Therefore, many reads whose 3' ends are in GC-rich exonic sequence are expected to have their 5' ends in AT-rich intronic sequence. This becomes even clearer when the peak density meta-profile is plotted out (Figure Supplementary 2). Peak density is high at the beginnings of exons, where most reads are expected to have their 5' ends in the intron, but low at the ends of exons, where read 5' ends are more likely to map to the exon itself.

Further examination of the meta-profiles reveals that the simulated peaks do not display the increased peak density at the very beginning of the exons, as observed for true data. It thus appears that this peak cannot be explained by read 5' end biases. Regarding the second peak

further downstream in the exon, Figure Supplementary 2 suggests that it may be the result of two separate phenomena: a first increase (~40-60 nucleotides into the exon) that is also observed for the simulants and is thus likely artefactual, and a second increase ~80 nucleotides into the exon, that is not observed for simulants and is thus likely genuine.

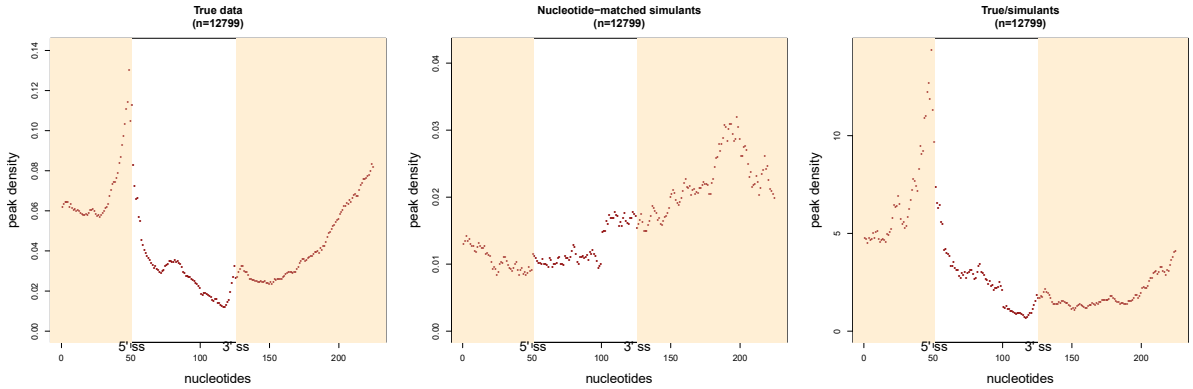



**Figure Supplementary 2**: average peak density for the last 50 nucleotides of exons, the first 50 nucleotides of introns, the last 25 nucleotides of introns, and the first 100 nucleotides of exons (concatenated). For the last panel, the meta-profile observed for true peaks has been divided by that obtained for nucleotide-matched controls. The orange boxes show the positions of exons. Only exons at least 100 nt long have been included.

# 4. Discussion

### 4.1. Gene architecture adaptation during development

We showed that in wild-type embryos pre-mRNA splicing imposes significant constraints on early zygotic expression, which is a likely explanation why most early zygotic genes are intronless (De Renzis et al., 2007a). Although a moderate decrease in the length of syncytial blastoderm interphases (seen in grapes mutant embryos (Sibon et al., 1997)) was not sufficient to induce splicing defects in otherwise wild-type embryos (data not shown), we hypothesize that avoidance of pre-mRNA splicing during early zygotic expression is a consequence of the extremely short interphases and frequent mitotic cycles. Consistent with our observation, substantial intron retention was observed in transcripts expressed during early nuclei divisions 9 and 10 (Kwasnieski et al., 2019b). Similarly to *Drosophila*, mosquito *Aedes aegypti* and the zebrafish *Danio rerio,* the early zygotic transcripts are frequently intronless when compared with the rest of the transcriptome (Biedler et al., 2012; Heyn et al., 2014b). This suggests that highly proliferative tissues need coordination between cell cycle and gene architecture for correct gene expression and avoidance of abnormally processed transcripts. Our results highlight the cell cycle constraints during early embryonic development as a force capable of driving changes in gene architecture of multicellular organisms. In unicellular organisms intron paucity correlates with a bias toward the 5′ ends, whereas introns from multicellular genomes are evenly distributed throughout the genes (Mourier, 2003). This suggests that similar constraints on gene architecture are also likely to exist in yeast and other fast-dividing single cell eukaryotes.

An alternative, non-mutually exclusive hypothesis is related to the immature state of the chromatin and lack of histone marks in the early embryo. At mitotic cycle 8, when few zygotic genes are being transcribed, chromatin is still immature, with few nucleosome free regions, and undetectable levels of the histone methylation and acetylation marks characteristic of mature chromatin. With the activation of zygotic transcription, and though to result from the action of transcription factors like Zelda that recruits histones acetyltransferases, these marks start to increase, becoming widespread at cycle 14. Curiously, marks such as H3K36me3 and H3K4me3 remain with low levels specially on pre-MBT genes during that stage (Chen et al., 2013; Li et al., 2014). While H3K4me3 normally localizes near the 5′ region of genes, H3K36me3 is more enriched over the gene body. Despite these differences, both were shown to be involved in splicing regulation, by the recruitment of splicing factors to the chromatin or delaying the rate of transcription (de Almeida et al., 2011; Kolasinska-Zwierz et al., 2009; Luco

et al., 2010; Sims et al., 2007). In this way, if on one hand the absence of these marks could be correlated with the high levels of expression (Pu et al., 2015), a typical feature of pre-MBT genes, on the other side this can compromise the efficiency of splicing in those genes.

### 4.2. Regulation of gene expression by spliceosome modulation during development

The way splicing efficiency might be regulated through changes in constitutive spliceosome factors and how it might influence differential gene expression is a new area of interest. In this study, we presented experimental evidences supporting the hypothesis of a requirement for highly efficient pre-mRNA splicing during early embryonic development. Since the NTC/Prp19 complexes are known to be important for efficient spliceosome activation, and our mutant alleles specifically impaired pre-mRNA splicing of early zygotic but not maternally encoded transcripts, we propose that overall requirements for splicing efficiency are likely to vary during development, being the NTC/Prp19 complexes a key modulator of spliceosome activation rates. In agreement with this hypothesis, Prp19 expression varies during neuronal differentiation (Urano et al., 2006). In plants it was recently shown that, removal of retained introns regulates translation in rapidly developing gametophytes (Boothby et al., 2013). Moreover, granulocytes differentiation is controlled by a set of genes that become downregulated by non-mediated decay that results from intron retention, which in your turn is a consequence of downregulation of several splicing factors (Wong et al., 2013). In *Drosophila* wild-type embryos, a sub-population of early zygotic transcripts similarly showed some degree of intron retention (Figure 3.2B,D, Figure 3.2—figure supplement 2B). Therefore, and although, our evidences suggest a pre-requisite for highly efficient splicing during early embryonic development, intron retention is paradoxically also likely to play an important regulatory role in the expression of early zygotic transcripts. This further supports the possibility that modulation of spliceosome activation *per se* is important for differential regulation of gene expression during development.

### 4.3. Intron function in early zygotic expressed genes

We showed that *Drosophila* early embryonic development impose constraints on splicing. This is consistent with the fact that most genes expressed at these stages of development have no introns. However, a small subset of pre-MBT genes (including several embryonic patterning genes) do contain introns. Since some of these introns are conserved and intronic sequences are known to play a variety of functions, we hypothesised that these introns have important roles in the regulation of gene expression. However, deletion of even-skipped (eve) intron had no obvious effect in the expression and embryonic patterning functions of this gene. It is nevertheless still possible that eve intron is required for the correct expression and development of nervous central system (Doe et al., 1988).

Since it was previously shown that the presence of an intron would significantly increase the transcription rate of an intronless reporter gene during *Drosophila* cellular blastoderm (nc14) (Rose, 2008), we hypothesized that splicing of eve intron would influence gene expression levels. Nonetheless, no differences in eve expression were observed between control and intronless eve, which suggests that this intron and its splicing is not rate-limiting for normal expression levels. Since co-transcriptional splicing is known to prevent the formation of RNA-DNA hybrid structures (R-loops) between the nascent transcript and the template DNA, it is possible that eve intron helps to avoid the accumulation of such hybrids. Considering that R-loop accumulation during the fast syncytial nuclear divisions is likely to led to significant levels of DNA damage due to DNA replication fork collapse, it would be interesting to investigate if deletion of eve intron is associated to a localized accumulation of R-loops, and if that was the case, if the presence of such intron helps to avoid the accumulation of R-loops because of splicing and/or Pol II density.

Finally, and although we failed to detect any obvious phenotype associated to eve intron deletion, it is still possible that other introns from pre-MBT genes are functionally relevant for *Drosophila* early embryonic development.

### 4.4. Efficient termination vs. readthrough

Taking advantage of *d*NET-seq it was possible not only to map with single nucleotide resolution the position Pol II over gene bodies of transcriptionally active genes, but also to analyse unstable nascent RNAs resulted from transcription readthrough. In fact, many genes were found to have transcriptionally active Pol II downstream of TES, in some cases, over several kilobases within intergenic regions or downstream convergent genes (Figure 3.11D). Although transcriptional termination occurred in many cases efficiently, which was observed by the lack of *d*NET-seq signal after the TES (Figure 3.11F), we also detected widespread pervasive transcription in the developing *Drosophila* embryo. This observation is consistent with similar patterns observed in human tissue-cultured cells, in which termination is found to occur over 3 kb downstream the TES (Nojima et al., 2015a), but contrasting to plants, where Pol II pauses in a narrower region (Zhu et al., 2018). A detailed look into gene architecture showed that transcription readthrough mostly occurred in the presence of convergent overlapping genes, when compared to isolated genes (Figure 3.11F), regardless of the transcription activity of the convergent overlapping gene (Figure 3.11H).

Since transcription readthrough could be observed even in cases where the overlapping convergent gene was transcriptionally active (Figure 3.11H), gene expression interference due to convergent Pol II collisions and/or other mechanisms is likely to occur in the developing *Drosophila* embryo (Hobson et al., 2012; Prescott and Proudfoot, 2002; Shearwin et al., 2005). Consistent with previous suggestions that transcription readthrough could regulate gene expression (Gullerova and Proudfoot, 2012; Muniz et al., 2017; Wery et al., 2018), we detected many cases where the convergent downstream gene was transcriptionally silent. For example, and suggesting some degree of functional significance, antisense transcription readthrough could be observed over purely maternal genes, which are transcriptionally silent during embryonic development (Figure 3.11D). Yet, no significant differences were observed in the levels of gene expression between genes with and without antisense transcription readthrough (data not shown). This suggests that although it might apply to some cases, it is unlikely to be a genome-wide mechanism of gene expression regulation. Transcriptionally active convergent genes could also be expressed in distinct cells, which would easily explain why these genes have expression levels comparable to genes without antisense transcription readthrough.

The fact that transcription termination is more efficient in isolated genes, compared to convergent overlapping genes, suggests distinct termination mechanisms. Both allosteric and torpedo termination models predict pol II pausing downstream to TES (Richard and Manley, 2009). However, in isolated genes Pol II is barely detected past the TES. This is consistent with the fact that *Drosophila polo* termination show reduced levels of Pol II on the 168bp intergenic region downstream of the gene. Interestingly, this termination process occurs independently of Xrn2 and Pcf11 (known to be involved in allosteric and torpedo termination processes), but dependent of TFIIB (Henriques et al., 2012). Thus, it was proposed that this process could be mediated by the interaction between the promoter and termination sites, which forms the structures known as gene loops (O'Sullivan et al., 2004), since TFIIF is known to facilitate the formation of these structures (Singh and Hampsey, 2007). Gene looping may facilitate reinitiation of Pol II, and in this way improving gene expression (Ansari, 2005). Interestingly, pre-MBT genes are found to be: highly transcribed, normally isolated from other genes, and to have reduced levels of Pol II downstream of TES (Figure 3.11E). Thus, suggesting that this class of genes may form gene loops to enhance their gene expression.

### 4.5. Co-transcriptional splicing is associated to both, phosphorylated CTD Ser2 and Ser5.

Besides the native elongating RNA bound to Pol II, *d*NET-seq was also able to capture spliceosome components, such as the snRNAs and the splicing intermediates resultant from the first catalytic step of splicing. This is consistent with the previous mammalian NET-seq (mNET-seq) results (Nojima et al., 2015a, 2018). However, while mNET-seq showed reduced levels of these components associated to phosphorylated Ser 2 (pSer2) of CTD compared with pSer5, in *d*NET-seq both phosphorylated serines were associated with spliceosome components and splicing intermediates, which suggests that both CTD pSer2 and pSer5 are associated to the splicing process in *Drosophila*. The *d*NET-seq results are consistent with previous suggestions that phosphorylation of Ser2 and Ser5 are both associated to splicing in yeast and human cells (Alexander et al., 2010; Gu et al., 2013; Koga et al., 2015). Since CTD phosphorylated Ser2 was previously suggested to be involved in human splicing, the non-consistent mNET-seq result, can thus result from a lack of specificity of the phosphor Ser2 antibody used in the mNET-seq analysis, which differs from the one used in our *d*NET-seq.

Other possibility can be explained by a difference on conformation between the two CTD species that hides the phosphorylated Ser 2 epitope from the antibody in mammals.

### 4.6. Recursive splicing occurs a few nucleotides past the RS

Recursive splicing (RS) is an evolutionarily conserved process of long intron removal via multiple steps of splicing. RS is presumed to occur co-transcriptionally, meanwhile Pol II is transcribing the intron. *Drosophila* recursive splice sites were originally identified through a RNAseq sawtooth pattern of nascent transcripts (Duff et al., 2015). Since *d*NET-seq allows the correlation of RNA Pol II position to co-transcriptional splicing, it is an ideal technique to study RS. Consistently, *d*NETseq identified most RS sites previously described in *Drosophila* (Duff et al., 2015; Joseph et al., 2018b), in addition to some novel, previously unknown, RS sites. Furthermore, we also observed that RS splicing typically occurs few nucleotides past transcription of the recursive splicing site, and that spliced reads detected in *d*NET-seq are nascent transcript derived, since these introns are not detected in mature RNAs. The RS sites undetected by *d*NET-seq can potentially be explained by low levels expression in the embryo and/or the fact splicing might occur in more downstream positions during transcription, thus escaping the ability of *d*NETseq to detect splicing.

### 4.7. Splicing takes place as Pol II transcribes past the 3′ splice site, yet many nascent transcripts remain unspliced

Run-on Nascent RNA sequencing approaches quantify splicing kinetics based on the time that each reaction takes, with no information about Pol II position relative to the transcribed intron. *d*NET-seq allows to directly examine if splicing occurs when Pol II is positioned within 100 nucleotides (nt) past the 3′ splice site (3′ss), thus contributing to better understanding of splicing kinetics relative to transcription. Although limited by read size, *d*NET-seq has a significantly higher read coverage per splice junction, compared with the most recent approaches using long reads (Drexler et al., 2020). By defining the *d*NETseq splicing ratio of more than 6800 introns in developing *Drosophila* embryos (stage 9-10), it was possible to show that splicing can occur 100nt past transcription of the 3′ss, showing that similarly to

yeast cells (Carrillo Oesterreich et al., 2016), immediate splicing also occurs in *Drosophila* embryos. These results contrast with a recent work in tissue culture *Drosophila* S2 cells, that took advantage of long-read sequencing (nano-COP) (Drexler et al., 2020), where they failed to detect immediate splicing in *Drosophila* and human tissue culture cells. This difference can be explained by the low coverage of reads obtained with the nano-COP and/ or because this sequencing technique might have a bias for unspliced reads. Interestingly, while our *d*NETseq analysis identified many introns with both spliced and unspliced reads, two discrete classes of introns, showing all-reads (100% immediate splicing) or none-of-the reads spliced (0% immediate splicing), could nevertheless be easily detected. This clearly suggested that distinct mechanisms regulating the moment of splicing relative to Pol II position during transcription are likely to exist.

### 4.8 Evidence of pausing for splicing

We observed that Pol II pausing frequency is more pronounced in exons compared with introns (Figure 3.12C). A closer look, to intron-exon boundaries, clearly shows that Pol II has a tendency to pause near the 3'ss and also along the first 100nt of the exon (Figure 3.12B). Although this remains to be tested, is possible that these pauses can be associated to increase nucleosome occupancy on those regions.

Interestingly, the classes of introns with different *d*NET-seq splicing ratio, display very distinct Pol II pausing profiles over the exon. While SR=1 show a clear pause immediately after the 3'ss, probably due to splicing, thus suggesting that in these cases splicing is very efficient. In cases where SR<1, Pol II pausing only becomes more pronounced in more downstream regions of the exon. This is probably because splicing only occurs when Pol II is located on more downstream regions. Moreover, and although the number of cases is very reduced, we observed that most alternative skipped exons show low densities of Pol II, compared with constitutive exons, which is in agreement with the hypothesis that Pol II pausing is required for splicing to occur.

### 4.9 Other observed significant pauses

We also noticed two additional pauses of Pol II near the 5'ss, potentially associated to splicing. One near the 5'ss can potentially resulting from splicing intermediate reads misalignment. And a pause detected 30nt downstream to 5'ss, possibly resulting, although very speculative, from a Pol II pause associated to snRNP U1 recruitment to recently transcribed 5'ss.

### 4.10 Sequence related features influencing splicing efficiency

Supporting the hypothesis that splicing kinetics relative to transcription is potentially regulated, we observed that weaker 3´ss correlate with a higher frequency of introns without immediate splicing (SR=0) and reduced number of introns with high levels of immediate splicing (SR=1), which suggests that weaker 3'ss causes a delay of splicing kinetics, only occurring when the elongating Pol II complex is already in a significantly more downstream position relative to 3'ss.

Moreover, high GC content in downstream exons compared with upstream intron also correlated with an increase number of introns with SR=1 and a decrease number of SR=0 introns, which is consistent with the fact that high GC content helps to define exon inclusion (Amit et al., 2012a). This is probably because GC content is associated with high nucleosome occupancy, thus contributing for a slowdown of Pol II elongation rate and consequently changing the timing of splicing during transcription. Accordingly, it was previously shown that histones are enriched in exons over introns (Spies et al., 2009) and that increase in nucleosome density at intron–exon boundaries could promote Pol II pausing, which has been shown to occur in yeast (Churchman and Weissman, 2011b), and human cells (Nojima et al., 2015b).

### 4. 11 Gene architecture features influencing splicing efficiency – intron and exon size

Both intron and exon sizes are known to influences the splicing. Accordingly, short introns show reduced relative number of SR=0 cases compared with longer introns in the range

of 500 nt (Figure 3.14E). Suggesting that longer introns are spliced in downstream position during transcription compared with short introns. This consistent with the fact that longer introns are less efficiently spliced compared with short introns (Khodor et al., 2011b, 2012b), and that expansion of a small intron (62 nt to 373 nt) produces aberrant splicing phenotypes in *Drosophila* (Talerico and Berget, 1994). In the same work was also shown that short introns do not required strong 3'ss to be efficiently spliced, compared with long introns that do. One possible explanation is the fact that short distance between the two splice sites allow the efficient splicing, thus following the intron definition proposed model (Berget, 1995). In opposite way, splicing of long introns were found to be dependent on flanking exons definition by the spliceosome prior to exon juxtaposition and intron removal, since the 5'ss of downstream intron is required for it (Talerico and Berget, 1994). From this perspective was proposed that the efficiency of splicing of both modes splicing will depend on short distances between splice sites, between exons in exon definition and between intron in intron definition. Accordingly, long intron followed by exons with similar size are known to impair splicing. (Pai et al., 2017; Sterner et al., 1996). Hence, cases in which the relative size between introns and downstream exon are similar (exon to intron ratio = 1) the relative number of introns with SR=0 increase (Figure 3.14E). Surprisingly, and since exon definition predicts the requirement of downstream intron 5'ss for the splicing of long intron, we found an unexpected decrease in SR=0 number of very long introns (Figure 3.14E). This suggesting, that in the most of these cases splicing is performed by intron rather than exon definition, since Pol II is found associated to spliced reads before the transcription of downstream 5'ss. For such distant spliced sites to be recognized so efficiently, models has been proposed in which the upstream exon is tethered by the Pol II CTD to the downstream exon, (Dye et al., 2006; Hollander et al., 2016; Zeng and Berget, 2000). Although this observation does not fit with higher levels of intron retentions observed in very long intron in (Khodor et al., 2011a, 2012a) is consistent with the fact that very long introns have shorter splicing half-life, meaning more efficiently splicing reported in a different work (Pai et al., 2017)

### 4.12 Gene architecture features influencing splicing efficiency – intron positioning

Moreover, we also observed how intron position influences the moment of splicing during transcription. Interestingly both first and last introns show significant more cases with

SR=0, meaning that in these splicing is prone to occur later during transcription or possibly post-transcriptionally. These results are consistent with the observations done specifically in *Drosophila,* in which splicing efficiency was calculated based on intron retention in nascent transcripts (Khodor et al., 2011a, 2012a). One hypothesis that could explain this observation, could be related with the fact that spliceosome components have to compete with capping machinery by the CTD, thus affecting the efficiency and delay during transcription of splicing reaction. Concerning the last introns, there are several evidences showing that both splicing of the terminal intron and terminal polyadenylation and cleavage are processes that depend on each other (Dye and Proudfoot, 1999; Niwa and Berget, 1991; Rigo and Martinson, 2008). Therefore, is expected that in those cases splicing of the last intron occurs when Pol II past the TES, which could explain the significant increase in cases with SR=0. In agreement, and since single introns satisfies both the first and the last position conditions, almost 40% of these cases were showed to have only unspliced reads. Interestingly, most pre-MBT genes harbouring introns have only one intron, which suggests that this could be an extra feature contributing for the intron retention observed in this class of genes (Guilgur et al., 2014; Kwasnieski et al., 2019a).

Overall, the effect on splicing ratio induced by each feature tested was not very in pronounced. This is expected since in each case, the splicing ratio depends on many different features like sequence, gene architecture and epigenetic marks, thus requiring more complex analysis. However, was curious to observed that only features related with sequences like 3'ss strength and GC content influences significantly the numbers of SR=1 cases. Thus, suggesting that these efficiently spliced cases are dependent, at some extend, on the sequences from the spliced intron or downstream exon. In the other side, SR=0 cases were show to be more responsive to gene architecture features, like intron size and intro-exon ratio. Why SR=1 cases are not perturbed by theses last features? This can happen because sequence related features, or by recruiting splicing enhancer factors or contributing for nucleosome positioning, can have a much more determinant effect in splicing decisions rather than gene architecture.

### 4.13 Impact of Development on splicing efficiency.

Since the early embryo was show to impose constrains for splicing (Guilgur et al., 2014), we hypothesized whether splicing efficiency would be lower in genes expressed during

these early stages of development. Due to a technical constrains to obtain enough biological material for *d*NET-seq approach, MBT stage during embryo development was the earliest stage obtained. To test our hypothesis, we compared the splicing ratio of introns from a set of early zygotic (pre-MBT) genes when expressed during MBT with introns from Later expressed genes (stage 9-10). Although, differences observed may indicate that pre-MBT were more efficiently spliced compared with later genes (Figure 3.14C), a closer look by comparing the Late expressed genes with similar read density to Pre-MBT, show no significant difference between this two groups. This does not mean that the earliest stages could not show reduced levels of splicing efficiency or in this case less immediate splicing. Actually, our observation is very consistent with the fact MBT expressed genes have a tendency to retain less the intron compared with earliest stages (Kwasnieski et al., 2019a). This observation suggests that if splicing constraints exist, they are likely to occur in stages before MBT.

### 4.14 Genome-wide Co-transcriptional splicing

Besides unspliced and spliced reads, *d*NET-seq allow to the detect one of the splicing intermediate products resulted from the first splicing reaction. This product is captured in *d*NET-seq because the active spliceosome is immunoprecipitated together with Pol II in this technique (Nojima et al., 2015a, 2018), becoming an indicator of co-transcriptional splicing. Quantification of these cases show that around 80% of introns are co-transcriptionally spliced. In agreement, the large majority (82%) of these cases show spliced reads mapping the spliced junction, reinforcing the idea that these cases are spliced during transcription. The remaining 18% of cases does not showed splicing occurring when Pol II is located at least 100 nt downstream the 3'ss or spliced reads associate to downstream intermediates. This can result from cases in which the splicing occurs in even more downstream regions, where the technique is not able to reach to the Pol II position. Surprisingly, cases where no splicing intermediate is detected, also show evidences of splicing after Pol II transcribes the 3'ss. Thus, suggesting that in these particular cases splicing reaction must be extremely fast between the 2 reaction, since no product from the first reaction is detected. Overall, *d*NET-seq show that 95% of the introns analysed show evidences of being co-transcriptionally spliced in genes expressed in *Drosophila* embryos. These values are consistent with the high percentages (87%) of the

introns that show more than 50% co-transcriptional splicing in *Drosophila* S2 cells (Khodor et al., 2011a).

### 4.15 Future perspectives

Adapted to *Drosophila* embryos for the first time in this work, *d*NET-seq proved to be a promisor technique able to link the Pol II position to the splicing status of the nascent RNA in a developing organism. By taking advantage of *Drosophila* genomic architecture particularities, was possible to stablish important links between Pol II pausing, splicing, gene architecture and other factors, thus raising many interesting questions.

Since the early embryo is known to impose constrains for slicing specially during pre-MBT (Guilgur et al., 2014; Kwasnieski et al., 2019a), and no significant differences were observed in splicing efficiency with *d*NET-seq between genes expressed during embryo cycle 14 (stg5) and stg. 9-10 (data not shown). This leaved us with a question whether earlier stages of *Drosophila* embryogenesis could influence the expression of the first zygotic transcribed genes. We hypothesized that splicing in those early stages should occur in downstream position in the gene compared with later stages. If that is the case, that would be an evidence that the moment of splicing during transcription can change between different developmental stages or even cells types. To answer this question, we intend to optimize the *d*NET-seq technique in order specifically sequence the nascent transcriptome from early stage.

To explore the differences between cell types, we can go further and adapt this technique in order to explore the nascent transcriptome in specific tissues. Taking advantage of UAS-Gal4 driver system in *Drosophila*, we intend to induce the expression of a tagged Pol II in a specific tissue and perform *d*NET-seq. This way we can analyse nascent transcription and co-transcriptional processes, as well the transcription of unstable ncRNAs, between different tissues, for instance different central nervous system sections, without the need to dissect each section.

The variation on splicing kinetics associated to very distinct Pol II pause profiles, thus suggest the existence of different mechanisms of co-transcriptional splicing. We hypothesized that there are conserved sequences within the exon or intron that regulate the variation of splicing kinetics. To test this hypothesis, our primary approach will be to determine where

relevant sequence information is located by searching for regions of increased evolutionary conservation. Next, and taking advantage of these sequences we want to predict and validate the relevant proteins that influences this splicing kinetic. Finally, we want to test if this variation is functional. For instance, if a normally immediately spliced intron was spliced slowly instead, would this negatively affect the organism? We will address this problem by performing *d*NET-seq on a large number of *Drosophila* species and checking whether splicing kinetics is more conserved for genes that have greater functional relevance, as expected if splicing kinetics is functionally important.

Non less interesting, was the pause observed around 30nt downstream of 5'ss. We hypothesized that this may be caused by the recruitment of U1 snRNP to the nascent RNA, associated to the mechanism of tethering of the upstream exon to Pol II CTD. Since only part of cases show this pause, we intend to test whether this is dependent on some specific sequence and if it is functional relevant, by using the same approach that we proposed before for the variation of splicing kinetics.

We also notice that in many cases while many spliced reads are detected close to 3'ss, unexpected unspliced reads emerge downstream to them. Whether these unspliced reads represent paused Pol II that fail the splicing on the first position and is waiting for splicing to occurs, eventually later, or will fail completely co-transcriptionally, is not understood. If splicing does not occur always in the same Pol II pause position for each intron, is this stochastically originated in the cell or is caused by the differences between each cell within the population analysed? To answer these questions new approaches, have to be developed. As for example: new live imaging techniques that allowed to track the Pol II and each intron moment of splicing; long read sequencing with more accuracy and coverage; and also, or even a NET-seq technique coupled to single cell sequencing.

# Materials and Methods

# Materials and Methods related to chapter 3.1

## Fly work and genetics

Flies were raised using standard techniques. The *fandango* alleles were isolated in a previously reported maternal screen (Pimenta-Marques et al., 2008). Maternal mutant embryos and germ-line mutant clones were generated using the FLP/FRT ovoD system (Chou and Perrimon, 1992). Germ-line clones of *fand*1 and *fand*2 were established by crossing FRT42B *fand*1/CyO or FRT42B *fand*2/CyO virgins to hs-FLP; FRT42B ovoD/CyO males and the progeny was heat shocked once at 37°C for 1 hr during second and third larval instar stages. As control we generated germ-line clones with FRT42B by crossing FRT42B/CyO virgins to hs-FLP; FRT42B ovoD/CyO males and followed by the heat shock procedure described before. To generate homozygous mutant clones in ovaries for *fand*1 (negative for nuclear GFP label, nGFPminus) we used FLP/FRT to induce mitotic recombination. Females y, w, hs-FLP; FRT42B nGFP/ CyO hs-hid flies were crossed with w; FRT42B, *fand*1/CyO hs-hid males. Recombination was induced by 1-hr heat shock at 37°C at second and third instar larval stage. Adult ovaries were harvested from 4–5-day-old females and subsequently processed for immunofluorescence.Viability and phenotypes of *fandango* alleles were complemented by crossing a transgenic fly carrying a genomic fragment construct that contained a wild-type copy of CG6197 (wt-*fandango*). w; FRT42B, *fand*1/CyO virgins were crossed to wt-*fandango*; FRT42B, *fand*2/CyO males; reciprocal crosses were also performed. Offspring were counted to determine viability. Rescue of maternal phenotypes (cellularized blastoderm defects and splicing defects in early zygotic transcripts) was also analyzed in embryos laid by F1 wt-*fandango*/ +; FRT42B, *fand*1/FRT42B *fand*2 females. Germ-line clones of *fand*1 and *fand*2 were also rescued (cellularized blastoderm defects and splicing defects in early zygotic transcripts) by a copy of wt-*fandango* in the third chromosome. FRT42B *fand*1/CyO; wt-*fandango* or FRT42B *fand*2/CyO; wt-*fandango* virgins were crossed with hs-FLP; FRT42B ovoD/CyO males and heat shock performed as described above. To induce maternal and zygotic expression of the UAS-kuk-LacZ construct in control and *fandango* maternal mutant embryos, we performed the following crosses: Maternal expression in control genetic background: virgin females +/+; Nanos-Gal4, UAS-kuk- LacZ/TM6B crossed with wild-type males. Zygotic expression in control genetic background: virgin females +/+; actin-Gal4/TM6B crossed with +/+; UAS-kuk-LacZ males.

Maternal expression in *fandango* maternal mutant genetic background: firstly, virgin females FRT42B *fand*1/CyO; Nanos-Gal4, UAS-kuk-LacZ/TM6B crossed with hs-FLP; FRT42B ovoD/CyO males, and heat shocked as described above. After eclosion, Cy+ and Tb+ females were selected from the progeny and crossed to wild-type males. Zygotic expression in *fandango* maternal mutant genetic background: firstly, virgin females FRT42B *fand*1/CyO; actin-Gal4/TM6B were crossed with hs-FLP; FRT42B ovoD/CyO males, and heat shocked as described. After eclosion, Cy+ and Tb+ virgin females were selected from the progeny and crossed to +/+; UAS-kuk-LacZ males.

To induce maternal and zygotic expression of the 4x intron kuk-LacZ and no intron kuk-LacZ constructs we performed following crosses: Maternal expression: firstly, virgin females +/+; actin-Gal4/TM6B crossed with +/+; UAS-4xintronkuk- LacZ/TM6B or UAS-nointron-kuk-LacZ/TM6B males. After eclosion, females Tb+ (+/+; actin-Gal4/ UAS-4xintron-kuk-LacZ or +/+; actin-Gal4/UAS-nointron-kuk-LacZ) were selected and crossed with wild-type males. Zygotic expression: virgin females +/+; actin-Gal4/TM6B were crossed with +/+; UAS-4xintron-kuk- LacZ/TM6B or UAS-nointron-kuk-LacZ/TM6B males. To analyze zygotic expression of the 4x intron kuk-LacZ and no intron kuk-LacZ constructs under the control of the minimal promoter of the gene nullo, females carrying the corresponding construct were selected and crossed with wild-type males. To drive embryonic and ovarian expression of Myc-tagged Fandango and Myc-tagged Prp19 proteins, Nanos-Gal4 homozygous virgins were crossed with UAS-Fandango-6xMyc males or UAS-Prp19-6xMyc/TM6B males, respectively. After eclosion females (in case of Myc-Fandango) or Tb+ females (in case of Myc-Prp19) were selected, dissected ovaries from 4–5-day-old females, or laid embryos after a cross with wild-type males were used for protein extraction.

**Cloning of *fand*ango alleles**

To identify the gene responsible for lethality in *fandango* alleles, we performed a complementation analysis using the Bloomington 2R Deficiency kit. Deficiency Df(2R)CX1 (covering an interval from cytological band 49C1 to 50D2, Bloomington stock number 442) failed to complement zygotic viability of both *fandango* alleles (complementation group 7). All additional 22 overlapping deficiencies complemented both *fandango* alleles. The cytological interval between bands 50B4-B6 (comprising 6 genes) was not covered by the 22 deficiencies. We cloned and sequenced these 6 genes from genomic DNA of both control and

*fandango* alleles and identified mutations in gene CG6197 in both *fandango* alleles. To confirm the identity of our mutants, we digested DNA from genomic clone (BACR14P04, Flybase) with restriction enzymes XbaI and EcoRI to generate a genomic fragment comprising the wild-type gene sequence of CG6197 (wt-*fandango*). Then we cloned the fragment into pCasper vector and used it to generate transgenic stocks (Bestgene, Chino Hills, CA, USA). A genomic wild-type copy of CG6197 under the control of its endogenous promoter fully complemented all known phenotypes in both *fandango* alleles.

**Immunohistochemistry**

0–3 hr (after egg laying) embryos, both maternally mutant for *fandango* and control, were fixed and stained using standard procedures (Pimenta-Marques et al., 2008). For Neurotactin and Slam immunostaining, the fixation procedure was modified: embryos were added to boiling heat fix solution (68 mM NaCl +0.1% Triton) and stirred for 1 min, then cooled by adding an equal volume of cooled fix solution. Immunostaining for oogenesis phenotypic analysis was performed as described in (Guilgur et al., 2012). Following primary antibodies used were: mouse anti-Neurotactin clone BP106 at 1:133 (DSHB, Iowa City, Iowa, USA); mouse anti-pTyr at 1:1000 (9411; Cell Signaling, Danvers, MA, USA), and rabbit anti-slam at 1:1000 (Ruth Lehman Lab). For F-actin staining, a 5-min incubation with phalloidin-Rhodamine at 1:200 dilution (Sigma, St Louis, MO, USA; stock concentration 1 mg/ml) was employed at room temperature. For DNA staining, we used SYTOX Green (Invitrogen, Grand Island, NY, USA) at 1:5000 dilution with 5 mg/ml RNase A in PBT (PBS+0.1% Tween-20) for 30 min at room temperature. Cy3- or Cy5-conjugated secondary antibodies were used at 1:1000 dilution (Jackson ImmunoResearch, West Grove, PA, USA) and anti-rabbit Alexa Fluor 488 at 1:1000 dilution (Molecular Probes, Grand Island, NY, USA).

**Generation of constructs and cloning**

The kuk-LacZ construct was synthesized using the 5′UTR and intron of the kuk small transcript (kuk-RB, Flybase). The kuk ORF was replaced by the LacZ coding sequence and was followed by the 3′ UTR of the original transcript (Figure 3.5A). The kuk-LacZ construct was fused to a UASg promoter (Gateway system, Invitrogen, Grand Island, NY, USA). The 4x intron kuk-LacZ construct was synthesized using 4 repeats of the fragment of 5′UTR and intron

of the kuk small transcript, separated by 201 nucleotides of in-frame LacZ sequence. The stop codon is followed by the 3′UTR kuk small transcript sequence and 300 bp of the 3′-located genomic region to promote transcriptional termination (Figure 3.6A,D). In the case of the no intron kuk-LacZ, all splice sites (meaning 5′ splice site, branch point, and 3′ splice site) were mutated to thymidines (Figure 3.6A,D). To induce expression of these constructs, they were fused to UAS promoter or nullo minimal promoter. The 4x intron kuk-LacZ and no intron kuk-LacZ constructs were cloned into pWALIUM22. Fandango open reading frame, kuk-LacZ, 4x intron kuk-LacZ, and no intron kuk-LacZ constructs were synthetized (GenScript, Piscataway, NJ, USA). Prp19 open reading frame was cloned into pDONR221 from DGC gold BDGP clone LD09231. Prp19 and Fandango ORFs were subcloned into a vector containing the UASp promoter and 6x C-terminal Myc-tag (Gateway, Invitrogen, Grand Island, NY, USA). All constructs were then used to generate transgenic flies stocks (BestGene, Chino Hills, CA, USA).

**RT-PCR**

Total RNA was extracted from 0–3 hr (after egg laying) embryos, unfertilized embryos, and 4-day-old female ovaries mutant for *fandango* and control (FRT42B) following standard procedures (PureLink RNA Mini Kit, Ambion, Grand Island, NY, USA). 1 µg of RNA was then used for reverse transcription with Oligo(dT)12–18 and/or random hexamers primers following the manufacturer's protocol (Transcriptor First Strand cDNA Synthesis Kit, ROCHE, Germany). Primer combinations used were designed with PrimerSelect (Lasergene, Madison, WI, USA) and PCR was performed using GoTaq DNA polymerase (Promega, Fitchburg, WI, USA). Sequences of all primers used are listed in (Methods Table 1).

**Real-time qPCR**

To measure transcription levels, embryos were staged in three different groups based on the embryonic morphology: stage A (embryos from cycle 1 to 8, no pole cells, and no cortical nuclei are observed); stage B (embryos from cycle 8/9 to 13, pole cells present, and cortical nuclei are observed); and stage C (embryos at interphase 14, blastoderm cellularized). Three independent replicas for each stage, containing each 10 manually selected embryos were generated. Three different genetic backgrounds were analyzed (control (FRT42B), FRT42B

*fandango*, and grapes as positive control). To measure *fandango* mRNA levels, unfertilized eggs were analyzed (three replicas). To measure transcription level of the 4x intron kuk-LacZ and no intron kuk-LacZ constructs, 0–3 hr (after egg laying) embryo collections were used to analyze both maternal and zygotic induced expression. Total RNA was extracted from samples and then used for reverse transcription with Oligo(dT)12–18 as described above. Real-time mRNA quantification was performed following the manufacturer's protocol (QuantiFast SYBR Green RT-PCR Kit, Qiagen, Germany). For analysis of transcription levels of early zygotic genes (nullo, snail, scute, even-skipped, and tailless) the *Drosophila* QuantiTect Primer Assay (Qiagen, Germany) was used. For mRNA level measurements of *fandango*, 4x intron kuk-LacZ and no intron kuk-LacZ constructs primers were designed with Primer3 (Methods Table 1).

**Antibodies generated**

Anti-Fandango and anti-Prp19 rabbit polyclonal antibodies were raised against recombinant proteins corresponding to amino acids 551–750 of Fandango/CG6197-PA, and to amino acids 20–219 of Prp19-PA, respectively (Metabion international AG, Germany). In both cases it was used His-tagged recombinant proteins as antigen and the antibodies were affinity purified.

**Biochemistry**

Protein extracts were obtained from 0–3 hr (after egg laying) embryos or 4-day-old female ovaries. Embryos were dechorionated with 50% commercial bleach solution and ovaries dissected in PBS, samples then homogenized in NB buffer (150 mM NaCl, 50 mM Tris–HCl pH 7,5, 2 mM EDTA, 0,1% NP-40, 1 mM DTT, 10 mM NaF, and EDTA-free protease inhibitor cocktail, Roche, Germany), and centrifuge at 20000×g for 3 min. Supernatant was recovered and centrifuged twice. To analyze NTC/Prp19 complex composition (Table 1), co-immunoprecipitation was done using protein extracts from embryo or ovary tissues expressing Myc-tagged Fandango or Prp19. Briefly, 1 mg of protein was diluted in 1 ml NB buffer and incubated with 1 µg/ml of mouse c-Myc antiboby (9E10) (Santa Cruz Biotechnology, Dallas, Texas, USA) for 1 hr at 4°C. Subsequently, 0.9 mg of Dynabeads Protein G (Invitrogen, Grand Island, NY, USA) were added to the immune complex and

incubated 1 hr at 4°C. Beads were washed three times with NB buffer and protein elution performed with 50 µl of 100 mM Glycine pH 2.5 during 2 min at RT and stopped with 5 µl of 1M Tris–HCl pH 10.85. Eluted proteins were then precipitated in five times the volume of acetone at −20°C and samples analyzed by liquid chromatography coupled to tandem mass spectrometry (Mass Spectrometry Laboratory, Institute of Biochemistry and Biophysics, Poland). To analyze NTC/Prp19 complex composition (showed in Figure 3.3A), protein co-immunoprecipitation was performed using nuclear protein extracts (adapted from (Kamakaka and Kadonaga, 1994)Kamakaka and Kadonaga, 1994) from a collection of 0–3 hr (after egg laying) wild-type embryos (Oregon-R). 1 mg of protein extract was incubated with rabbit anti-Prp19 (1:1000 dilution) or the pre-immune (1:10,000 dilution) as control, in HNEB2 buffer (100 mM NaCl, 2,5 mM MgCl2, 10 mM Tris–HCl pH 7,5, 0,5% Triton X-100, and EDTA-free protease inhibitor cocktail, Roche, Germany) during 1 hr at 4°C. The procedure was carried out as described in previous co-immunoprecipitation and eluted complexes were boiled in Laemmli sample buffer and analyzed by Western Blot.

Size-exclusion chromatography was performed in protein extracts of 0–3 hr (after egg laying) embryo collections from FRT42B *fand*1 mutants or control (FRT42B). Extracts were prepared as described before in NB2 buffer (150 mM NaCl, 50 mM Tris–HCl pH 7,5, 2 mM EDTA, 0,01% NP-40, 1 mM DTT, and EDTA-free protease inhibitor cocktail, Roche, Germany). Subsequently, 2 mg of protein extract were fractionated using Superose 6 10/300 GL column (GE Healthcare, United Kingdom) in NB2 buffer and fractions collected and analyzed by Western blot. To analyze protein amount in ovaries and embryos (showed in Figures 3.1I and 3.3C), embryos were dechorionated and ovaries dissected as described above. Samples were homogenized in PBS supplemented with EDTA-free protease inhibitor cocktail (Roche, Germany) and centrifuged at 20000×g for 3 min at 4°C. Supernatant was collected and protein concentration determined using the Bradford method (BioRad, Hercules, CA, USA). Samples were immediately boiled in Laemmli sample buffer and 10 µg of protein was run in SDS-PAGE gel and analyzed by immunoblot. Levels of Pol II CTD Ser2 phosphorylation were analyzed in embryos dechorionated and manually selected at specific developmental stages based on the embryonic morphology (as described above). 15 embryos were selected for each stage and protein sample was obtained by lysing the embryos with a needle in Laemmli sample buffer and heating for 5 min at 100°C. Protein amounts corresponding to ∼7 embryos were running in SDS-PAGE and analyzed by immunoblot. Five independent replicas were analyzed. Antibodies used were: polyclonal rabbit anti-Prp19 at 1:8000 dilution; polyclonal rabbit anti-

Fandango at 1:1000 dilution; mouse anti-alpha-Tubulin Dm1A at 1:50,000 dilution (Sigma, St Louis, MO, USA); mouse anti-RNA Polymerase II H5 at 1:500 dilution (MMS-129R, Covance, Princeton, NJ, USA); rabbit anti-ISY1 at 1:500 dilution (ab121250; Abcam, United Kingdom); and mouse anti-CDC5L [2136C1a] at 1:200 dilution (ab51320; Abcam, United Kingdom).

**Mass spectrometry**

Peptides mixtures were analyzed by LC-MS-MS/MS (liquid chromatography coupled to tandem mass spectrometry) using Nano-Acquity (Waters, Milford, MA, USA) LC system and Orbitrap Velos mass spectrometer (Thermo Electron Corp., San Jose, CA, USA). Prior to analysis, proteins were subjected to standard 'in-solution digestion' procedure, during which proteins were reduced with 100 mM DTT (for 30 min at 56°C), alkylated with 0,5 M iodoacetamide (45 min in darkroom at room temperature), and digested overnight with trypsin (sequencing Grade Modified Trypsin—Promega V5111). The peptide mixture was applied to an RP-18 precolumn (nanoACQUITY Symmetry C18—Waters 186003514) using water containing 0,1% TFA as mobile phase, then transferred to nano-HPLC RP-18 column (nanoACQUITY BEH C18–Waters 186003545) using an acetonitrile gradient (0%–35% AcN in 180 min) in the presence of 0.05% formic acid with a flow rate of 250 nl/min. The column outlet was directly coupled to the ion source of the spectrometer, operating in the regime of data dependent MS to MS/MS switch. A blank run ensuring no cross contamination from previous samples preceded each analysis. Raw data were processed by Mascot Distiller followed by Mascot Search (Matrix Science, London, UK, on-site license) against Flybase database. Search parameters for precursor and product ions mass tolerance were 100 ppm and 0.6 Da, respectively, enzyme specificity: trypsin, missed cleavage sites allowed: 0, fixed modification of cysteine by carbamidomethylation, and variable modification of methionine oxidation. Peptides with Mascot Score exceeding the threshold value corresponding to <5% False Positive Rate, calculated by Mascot procedure, and with the Mascot score above 30 were considered to be positively identified. Human orthologs were determined using DSRC Integrative Ortholog Prediction Tool (DIOPT) (http://www.flyrnai.org/cgi-bin/DRSC_orthologs.pl). Only scores above two were considered such as the best matches when there was more than one match per input.

**High-throughput transcriptome sequencing (RNA-seq)**

Total RNA was isolated from 0–3 hr collections of *fandango* maternal mutant and control (FRT42B) embryos using TRIzol Reagent (Invitrogen, Grand Island, NY, USA), following standard protocol. DNAse I (Promega, Fitchburg, WI, USA) treatment was performed during 30 min at 37°C. DNAse was extracted by Phenol-Chloroform extraction; the RNA was precipitated with ethanol, and dissolved it in 25 µl of DEPC water. Bioanalyzer testing was used to analyze quality and concentration of the samples and made up to the volume to 100 µl with water, 50 µl of 7.5 M NH4OAc added, 0.5 µl of glycogen, and 250 µl of absolute ethanol. cDNA library was generated applying the standard Illumina protocol for RNA-Seq (polyA RNAs) and sequenced with an Illumina HiSeq (Oklahoma Medical Research Foundation, Oklahoma City, OK, USA). These generated RNA-Seq data for two biological replicates each of wild-type and *fandango* mutant (*fand*1), consisting of about 150 million illumina paired-end 100 bp reads per sample.

Paired-end reads were mapped with tophat (Trapnell et al., 2009) version 2.0.3 against the *Drosophila* melanogaster BDGP5 reference genome, using Flybase gene annotations downloaded from Ensembl e66 as guide. To analyze splicing defects, we first extracted exon–intron boundaries from gene annotations. To avoid potential confounding effects, we removed all boundaries that had overlapping exon sequence (from different genes or transcripts). Subsequent analysis used this set of 'safe' exon–intron boundaries. For coverage plots in Figure 3.2D, we also excluded boundaries where the intron or the exon were less than 50 b long. At each base within 50 bp either side of a splice site we count the number of reads that overlap that base, then divide by the total number of reads within the 100 bp centered around the splice site. To minimize noise, we require that at least 50 reads fall within the −50:50 window around the exon–intron boundary (reads that only partially overlap the window are also counted). To determine the frequency of splicing defects for each boundary, we extracted all reads that overlap the 5′ splice site by at least 10 bp to either side of the boundary. We classified each read as correctly spliced (if the read is split from the 5′ to the 3′ splice site), unspliced (if the read is not split) or mis-spliced (if the read is split but not matching the expected 5′ or/and 3′ sites). To reduce noise, we only include an exon–intron boundary if at least 10 fragments overlap that boundary. To determine the exon–intron gene structure (Figure 3.2—figure supplement 2D), each gene was divided in 1000 equal segments. For each segment of each gene, we checked for the presence (or absence) of an exon in that segment. For each segment, we then plotted the frequency of exon presence in all genes. If an exon randomly appears in a

given segment, it appears in ~50% of genes. For example, a set of intronless genes would produce a plot that would be always at 100%. To determine the splice site motif (Figure 3.2—figure supplement 2C), sequences around exon– intron boundaries were extracted and motifs drawn using Weblogo.

Early zygotic and maternal genes were defined using RNA-Seq developmental gene expression data from Flybase (Graveley et al., 2011). A gene was defined as an early zygotic gene when its expression at 2–4 hr is at least moderate (more than 10 expression units, as defined in the Flybase dataset) and at least 5× greater than its expression at 0–2 hr (irrespective of the 0–2 hr value). Maternal genes are those genes that are not early zygotic and have high expression (more than 50 expression units) at 0–2 hr. To avoid potential artifacts, genes that have an extremely high expression (more than 1000 expression units) were not considered. Applying this definition we obtained 270 early zygotic genes (including 43 genes from (De Renzis et al., 2007a)) and 2048 maternal genes. All scripts used for this analysis are available upon request. RNA-Seq data are available in the ArrayExpress database (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-2321.

**In situ hybridization**

The procedure has been described in (Stein et al., 2002). Antisense digoxigenin-labeled RNA probes were synthesized using the DIG RNA labeling Kit (Roche, Germany). eve and nos probes were made from pBluescript plasmids containing the respective cDNAs.

**Sequence alignment**

Sequences were aligned using ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/) and BoxShade 3.21 (http://www.ch.embnet.org/software/BOX_form.html) for printing and shading of multiple alignment file.

**Statistical analysis**

Unpaired t test and two-way ANOVA were performed using Prism 5.00 for Windows (GraphPad Software, San Diego, CA, USA).

**Methods Table 1 - Complete list of primers.** Sequences of all the primers used in the RT-PCR and real-time qPCR assays.

| Primer name | Primer sequence |
|---|---|
| **kuk-lt1** Forward | 5'-GCGTCGGTTGAATTGAAATCAG-3' |
| **kuk-lt2** Forward | 5'-GCAGGTAGATTGTCCATAC-3' |
| **kuk-lt3** Reverse | 5'-TGCCTGTTGAGCTGCCATTCTGTG-3' |
| **kuk-st1** Forward | 5'-TGTGCGGACGTGCTAGAAAATC-3' |
| **kuk-st2** Forward | 5'-CATAGCCGGTTCTCGCAAG-3' |
| **kuk-st3** Reverse | 5'-TGCCTGTTGAGCTGCCATTCTGTG-3' |
| **kuk-st4** Reverse | 5'-TGTACTTTATTTATGAAATGTG-3' |
| **grk ex** Forward | 5'-GCCACCCCAAGCGTTTTC-3' |
| **grk int** Forward | 5'-GCTTTCTGTTCGTGGATGATG-3' |
| **grk ex** Reverse | 5'-TGGGAGTCGTGGAGTCAGG-3' |
| **nos ex** Forward | 5'-GTCGACGCACTTCCCTTTTC-3' |
| **nos int** Forward | 5'-TGCAGAGGTCAGAGGATTTTC-3' |
| **nos ex** Reverse | 5'-ACGCCGAGATTGGTGGAC-3' |
| **osk ex** Forward | 5'-TTCCTCCAGCTGCCTCAAC-3' |
| **osk int** Forward | 5'-CCAGAAGTTTCGACTTCATCA-3' |
| **osk ex** Reverse | 5'-GTTCGACTCCGCTGCCTCAT-3' |
| **kr ex** Forward | 5'-GCGAGCACAAAATTAGGAGCAC-3' |
| **kr int** Forward | 5'-GAGAGCTGCGAAAACACTAAGAG-3' |
| **kr ex** Reverse | 5'-CATGCTGATCTCGGTCTGAAAC-3' |
| **eve ex** Forward | 5'-ATACCAAACATGCACGGATACC-3' |
| **eve int** Forward | 5'-TGACGAGTTACTTACACCCAATC-3' |
| **eve ex** Reverse | 5'-CAGTCCGGTATAGCAGGTG-3' |
| **ftz ex** Forward | 5'-TCCCAGCTACGACCAAGAGT-3' |
| **ftz int** Forward | 5'-GGCATCACACACGATTAACAACC-3' |
| **ftz ex** Reverse | 5'-CTGGTAGCTGCACTGATGTTGG-3' |
| **stg ex** Forward | 5'-CGCCAGCGAAATCAAACATC-3' |
| **stg int** Forward | 5'-TAGTTCAGCTCCCTCTTCACC-3' |
| **stg ex** Reverse | 5'-GCTGAGGCACCTTCTGACC-3' |
| **LacZ** Reverse | 5'-TCAAATTCAGACGGCAAACGACT-3' |
| **fand qRT** Forward | 5'-AGGAGTCATTCGGCACATTC-3' |
| **fand qRT** Reverse | 5'-GGGCCACTTGAACAGAGAGA-3' |
| **4xintron kuk qRT** Forward | 5'-TGGAGTGACGGCAGTTATCT-3' |
| **4xintron kuk qRT** Reverse | 5'-TTAAACTTCAGGGCGCGC-3' |
| **int1 ex** Forward | 5'-TGTGCGGACGTGCTAGAAAATC-3' |
| **int1 ex** Reverse | 5'-TCCAGACCAATGCCTCCCAGAC-3' |
| **int2 ex** Forward | 5'-GTCTGGGAGGCATTGGTCTGGA-3' |
| **int2 ex** Reverse | 5'-AGTTTGAGGGGACGACGACAGTA-3' |
| **int3 ex** Forward | 5'-TCTTCCTGAGGCCGATACTGTC-3' |
| **int3 ex** Reverse | 5'-ACGCCGAGTTAACGCCATCA-3' |
| **int4 ex** Forward | 5'- ACTCGGCGTTTCATCTGTGG-3' |
| **int4 ex** Reverse | 5'-GTCACTCCAACGCAGCACCAT-3' |

# Materials and Methods related to chapter 3.2

## Generation of transgenes

Plasmids containing a BAC (bacterial artificial chromosome) correspondent to the genomic region of knirps (-31,8kb to +7,1kb from TSS) (FlyFos029432) and eve (-26kb to +8,5kb from TSS) (FlyFos017066), were used to obtain the intron deleted transgenes of the respective genes. Introns were removed by *in vivo* Red/ET homologous recombination following the procedure (Counter-Selection BAC Modification Kit, Gene Bridges). Sequences of all primers used in this procedure are listed in (Methods Table 2). Obtained deleted intron plasmids, as well the original plasmid used as control, were injected in embryos carrying attP40 or attP-3B (#9750 BDSC Stock) integration sites in the case of knirps or eve respectively, to generate transgenic flies stocks (BestGene, Chino Hills, CA, USA).

## Fly work and genetics

Rescue of $eve^{3}$ null mutant allele viability (Fujioka, 1999), was complemented by crossing a transgenic fly carrying BAC construct that contained a wild-type copy of eve (BAC eve Ctr) or eve intron deleted version (BAC eve Del-Intron). w/w ; $eve^{3}$/CyO ; BAC eve Ctr / BAC eve Ctr or w/w ; *eve3*/CyO ; BAC eve Del-Intron / BAC eve Del-Intron virgins, were crossed to w ; Df(2R)eve /CyO ; BAC eve Ctr / MKRS or w ; Df(2R)eve /CyO ; BAC eve Del-Intron /MKRS males respectively. Offspring were counted to determine viability.

## RT-PCR

Total RNA was extracted from 0–3 hr (after egg laying) embryos, following standard procedures (PureLink RNA Mini Kit, Ambion, Grand Island, NY, USA). 1 µg of RNA was then used for reverse transcription with Oligo(dT)12–18 following the manufacturer's protocol (Transcriptor First Strand cDNA Synthesis Kit, ROCHE, Germany). Primer combinations used were designed with PrimerSelect (Lasergene, Madison, WI, USA) and PCR was performed using GoTaq DNA polymerase (Promega, Fitchburg, WI, USA). Sequences of all primers used are listed in (Methods table 1).

**Methods Table 2 – List of primers used in Red/ET BAC recombination.** Sequences of all the primers used in the Red/ET BAC recombination.

| primers to generate rpsL-neo product containing the homology arms |
|---|
| **knirps R1 FW** 5'- TTGCACTCGCTGATGGTGCTGATGTTGTTGTAAGAGCGGCCAAAGAAGGAGGCCTGGTGATGATGGCGGGATCG - 3' |
| **knirps R1 RV** 5' - AGCCGGCGGCGGGCTTCCATTTTGGCGCCTTCACCTGCGAGGGCTGCAAGTCAGAAGAACTCGTCAAGAAGGCG - 3' |
| **eve R1 FW** 5' - CTTGGCCACCCAGTACGGCAAGCCCCAGACACCGCCTCCCTCGCCAAATGGGCCTGGTGATGATGGCGGGATCG – 3' |
| **eve R1 RV** 5' - CTCCGAGCCGCGGCTGCCGTTCAAGGAGTTATCCGGACTGGATAGGCATTTCAGAAGAACTCGTCAAGAAGGCG – 3' |
| primers to generate non-selectable DNA lacking the intron |
| **knirps R2 FW** 5'- TTGCACTCGCTGATGGTGCTGATGTTGTTGTAAGAGCGGCCAAAGAAGGACTTGCAGCCCTCGCA – 3' |
| **knirps R2 RV** 5' – AGCCGGCGGCGGGCTTCCATTTTGGCGCCTTCACCTGCGAGGGCTGCAAGTCCTTCTTTGGCCGC – 3' |
| **eve R2 FW** 5' – CTTGGCCACCCAGTACGGCAAGCCCCAGACACCGCCTCCCTCGCCAAATGAATGCCTATCCAGTC – 3' |
| **eve R2 RV** 5' – CTCCGAGCCGCGGCTGCCGTTCAAGGAGTTATCCGGACTGGATAGGCATTCATTTGGCGAGGGAG – 3' |

## Materials and Methods related to chapter 3.3

### Embryo collection

*Drosophila melanogaster* flies (Oregon R (OrR) strain) were raised at 25°C, in polypropylene vials containing standard enriched culture medium (cornmeal, molasses, yeast, soya flour, and beetroot syrup). Three-day-old flies (counting from pupae eclosion) were fattened in culture medium supplemented with fresh yeast for 2 days. Embryos were collected into apple juice-agar plates supplemented with fresh yeast using appropriate cages containing approximately 200 flies each. To avoid female retention of older embryos, three pre-collections of 30min each were made before the first collection of embryos. To maximize egg laying, and avoid larvae contamination, adult flies were transferred to clean embryo collection cages every day over five days. For early stage embryos (2-3 hours after egg-laying), adult females were allowed to lay eggs for 1 hour in apple juice-agar plates. Plates were subsequently collected and embryos were aged at 25°C for 90 min. During the following 30 min, embryos were harvested from the plates, dechorionated in 50% bleach solution for 2 min, and washed once in Phosphate Buffered Saline supplemented with 0.1% Tween-20 (PBT) and twice in deionized water. In order to discard older embryos (stage 6 and older), manual staging of collected embryos was performed with the help of forceps and under a magnifier scope. Embryos were then resuspended in a solution containing 120mM NaCl and 0.04% Triton, and washed twice with 120mM NaCl solution. At the end of the 3 hours collection, the solution was removed and embryos were frozen in liquid nitrogen and stored at -80ºC. For the late stage (4-6 hours), eggs

were laid for 2 hours and aged at 25°C for 3.5 hours. Embryo collection and processing was similar to the early stage embryos, but in this case, no manual staging was performed.

**Embryo DNA staining**

For each embryo collection, and after dechorionation, a representative embryo sample was collected from the total pool and fixed in a scintillation flask, using a solution containing 1 volume of 4% formaldehyde in PBT and 4 volumes of heptane, for 20 min at 100rpm. The lower aqueous phase solution was subsequently removed, 4ml of methanol was added and embryos were shaken vigorously during 1 min. Embryos were then collected from the bottom of the scintillation flask, washed twice with methanol, and frozen at -20ºC in methanol. To rehydrate the embryos, they were washed for 5 min each, with 3:1, 1:1, and 1:3 mix solutions of methanol:PBT. Embryos were subsequently washed twice in PBT and incubated with 1:5000 Sytox green (Invitrogen), supplemented with 5µg/ml RNase A (Sigma-Aldrich) in PBT for 15 min. After washing with PBT, embryos were mounted in fluorescence mounting medium (Dako) and examined in a Zeiss AxioZoom V16 Fluorescence Stereo Microscope for image acquisition and embryo staging. Images were processed using ImageJ software (NIH).

**dNET-seq and library preparation**

The *d*NET-seq protocol was adapted from mNET-seq (Nojima et al., 2016). Briefly, 300 ul of frozen embryos was resuspended in 3.5ml of Buffer B1 (15mM HEPES-KOH, pH 7.6; 10mM KCl; 5mM MgCl2; 1mM DTT ; 0.1mM EDTA; 0.35M Sucrose; 4µg/ml Pepstatin; 10mM Sodium Metabisulfite; 0.5mM EGTA supplemented with complete EDTA free protease inhibitor (Roche) and PhoSTOP (Roche)). Embryos were homogenised in a Dounce homogenizer with 11x strokes using a tight pestle on ice. The suspension was centrifuged at 7700g for 15 min at 4ºC, the supernatant was discarded and the white pellet containing the nuclei was resuspended in 500 µl of Buffer B1. The suspension was again homogenised in the Dounce with 4x strokes and loaded without mixing on the top of buffer B2 (15mM HEPES-KOH, pH 7.6; 10mM KCl; 5mM MgCl2; 1mM DTT; 0.1mM EDTA; 0.8M Sucrose; 4µg/ml Pepstatin; 10mM Sodium Metabisulfite; 0.5mM EGTA supplemented with complete EDTA free protease inhibitor (Roche) and PhosSTOP (Roche)). The suspension was centrifuged at 1310g during 30 min at 4ºC, and the pellet was resuspended with 125 µl of NUN1 buffer

(20mM Tris-HCl (pH 7.9); 75 mM NaCl; 0.5mM EDTA and 50% Glycerol). 1.2ml of Buffer NUN2 (300mM NaCl, 7.5mM MgCl2, 1% NP-40, 1M Urea supplemented with complete EDTA free protease inhibitor (Roche) and PhoSTOP (Roche) was mixed with the nuclei and incubated on ice during 15 min preforming short vortex every 3 min. Chromatin was then centrifuged at 10000g for 10 min at 4ºC, washed with 100 µl of 1x MNase Buffer and incubated in 100µl of MNase reaction mix (1x MNase buffer and 30 gel unit/µl MNase (New England Biolabs)) during 3 min at 37ºC with mix at 1400rpm. The reaction was stopped with 10 µl of 250mM EGTA, centrifuged at 10000g for 5 min at 4ºC, and the supernatant containing the solubilized chromatin was recovered. For the early embryos sample, 2x 300µl of embryos were prepared in parallel and pooled together after the chromatin solubilization. Immunoprecipitation of Pol II-RNA complexes was performed using 50µl of Protein G Dynabeads (Thermo Fisher Scientific), pre-incubated over night with 5µg of the correspondent antibody: anti-Pol II CTD S5P (ab5131 Abcam) or anti-Pol II CTD S2P (ab5095 Abcam) in 100µl NET2 (50mM Tris-HCl pH 7.4; 150mM NaCl and 0.05% NP-40) and washed 3 times with NET2. Beads were incubated with the solubilized chromatin in 1ml total volume of NET2 during 1 hour at 4ºC, washed 7 times with 500µl of NET2 and once with 100µl of PNKT (1x PNK buffer and 0.1% tween) before incubation during 6 min in 50µl of PNK reaction mix (1x PNKT, 1 mM ATP and 0.05 U/ml T4 PNK 3′phosphatase minus (NEB) in a thermomixer at 37ºC and 1400rpm. After washing the beads with NET2, long RNA fragments were isolated using Quick-RNA MicroPrep (Zymo research): 300µl of RNA Lysis Buffer in 33% EtOH was mixed to the beads by pipetting. Beads were discarded and the suspension was loaded into a Zymo spin column that was centrifuged at 10000g during 30sec. The column was washed once with 400µl RNA prep buffer and twice with 700µl and 400µl RNA wash buffer respectively. RNA was then eluted in 15µl of DNase/RNase-Free water (zymo) and stored at -80ºC. 100ng of RNA was used to prepare each library, following the standard protocol of the Truseq small RNA library prep kit (Illumina). After adapter ligation and reverse transcription, the libraries were PCR amplified using 16 PCR cycles and cDNA libraries were fractionated in the gel between 130 to 300bp. The libraries were sequenced using PE-150 on the Illumina HiSeq X platform by Novogene Co., Ltd.

### *d*NET-seq data processing

Adapter sequences were removed from all *d*NET-seq paired-end samples using Cutadapt (version 1.18) (Martin, 2011) with the following parameters: -a TGGAATTCTCGGGTGCCAAGG -A GATCGTCGGACTGTAGAACTCTGAAC -m 10 -e 0.05 --match-read-wildcards -n 1. Paired-end read merging was performed using bbmerge.sh from BBMap (Bushnell et al., 2017) with the 'xloose' parameter. Merged reads were then aligned to the *Drosophila* reference genome (*dm6*; Ensembl release 95) (Cunningham et al., 2019) using STAR (version 2.6.0b) (Dobin et al., 2013) with –chimSegmentMin set to 20. Only uniquely mapped reads were considered, extracted using SAMtools (version 1.7) (Li et al., 2009) with -q set to 255. Exceptionally, in Figure 3.10 - figure supplement 1A, HiSat2 was used, using the same dm6 genome annotation file for genome indexing and using default parameters, with the --no-discordant –no-mixed flags set.

PCR internal priming events generated during library preparation were removed using a custom Python script (Prudêncio et al., 2019) with the following parameters: -a TGG.. -s paired. To obtain single-nucleotide resolution, a custom Python script (Prudêncio et al., 2019) was used to extract the 5′ end nucleotide of read 2 (after trimming) in each sequencing pair, with the directionality indicated by read 1 (Figure 3.9E).

### Publicly available RNA-seq datasets used

Publicly available *Drosophila* embryonic transcriptome sequencing data (Poly(A) RNA-seq), performed in developmental stages similar to the *d*NET-seq early and late samples, were used in this study. RNA-seq datasets corresponding to the 14B cycles (Lott et al., 2011b) GSM618409, GSM618410, GSM618421 and GSM618422 available at GEO: GSE25180 And RNA-seq datasets from 4-6h old embryos - were obtained from modENCODE project PRJNA75285 (The modENCODE Consortium et al., 2010): SRR023696, SRR023746, SRR023836, SRR035220, SRR023669, SRR035405, SRR035406, SRR024014 and SRR023539.

**RNA-seq data processing**

Adapters were removed from all datasets using Trim Galore (version 0.4.4) (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/; last accessed 26 April 2020). Datasets from modENCODE and GEO were aligned to the *dm6 Drosophila* reference genome (Ensembl release 95) (Cunningham et al., 2019) using STAR (version 2.6.0b) (Dobin et al., 2013) with --chimSegmentMin set to 20. Stringtie (version 1.3.3b) (Pertea et al., 2015) was used to quantify normalized gene expression as Transcripts Per Kilobase Million (TPM) values with the following parameters: -a 5 -e. In addition, the isoform list was provided together with the -G parameter corresponding to the *dm6* (Ensembl release 95) GTF file. Genes with TPM values above 2 were considered to be expressed.

**Selection of genes and isoforms for analysis**

Similar to a previous GRO-seq analysis (Core et al., 2008b), we used the read density of very large intergenic regions (gene desert regions) to define the reference for absence of transcription (Prudêncio et al., 2019). Gene deserts were divided into 50kb windows, and *d*NET-seq read densities were calculated by dividing the read counts in each window by the window length (in bp). Read counts per window were obtained with bedtools genome coverage (version 2.27.1-1-gb87c465) (Quinlan and Hall, 2010), and an arbitrary density threshold was defined as the 90th percentile of the read density distribution (Figure 3.11 - figure supplement 2A). Transcripts whose gene body *d*NETseq read density exceeded this threshold were considered to be transcriptionally active. (Figure 3.11 - figure supplement 2B). For all of the analyses performed on late genes, only transcriptionally active genes were considered. In addition, only one isoform per gene was considered for all analysis, selected as the isoform with the highest RPKM value in the RNA-seq dataset for the corresponding developmental stage.

The coordinates of previously identified pre-MBT genes were obtained from Chen et al. (2013) and converted to *dm6* coordinates. The most representative isoform for each gene was manually selected through visualization of individual profiles.

**Splicing intermediate and lariat detection**

Exons containing splicing intermediates or introns containing lariats were identified using a peak finder algorithm (NET_snrPeakFinder) (Churchman and Weissman, 2011a; Prudêncio et al., 2019) that detects the presence of a peak in the last nucleotide of an exon (splicing intermediate) or an intron (lariat), by comparing the accumulation of 3′ end reads mapping at that position with the mean read density of the flanking 200 nucleotides. A peak is called when the read density at the peak is superior to the mean of this surrounding region plus 3 standard deviations (Churchman and Weissman, 2011a). Since gene read density influences peak detection, the exons were divided into quartiles based on the *d*NET-seq read density of the corresponding gene. Only exons from the highest quartile (i.e. from genes with the highest read density) were considered in Figure 3.10 for *d*NET-seq Late analysis.

**Analysis of read density and peak calling**

RPKM values for the merged *d*NET-seq datasets were calculated in the following manner:

$$RPKM\ (transcript) = \frac{\text{reads} * \ 10^4 * \ 10^6}{(\text{total uniquely mapped reads}) * (gene\ length\ in\ bp)}$$

Where $10^4$ normalizes for gene length and $10^6$ normalizes for sequencing depth. Queries of gene 3′UTR overlaps (Figure 3.11) between genes were performed with bedtools intersect (version 2.27.1-1-gb87c465) (Quinlan and Hall, 2010).

To detect splicing intermediate and intron lariat peaks, the algorithm from Churchman and Weissman (2011a) was adapted as described in Prudêncio et al. (2019). To be able to call larger regions as peaks, a custom algorithm was developed and implemented in Python 3.7. Our peak caller detects regions where the local read density is significantly higher than expected by chance given the overall read density of the transcript. Almost all of the numerical parameters used by the peak caller can be adjusted. In Figure 3.12A, results obtained with two particular parameterizations are shown. Peak Caller 1 (“large peaks”) is adapted to detecting larger peaks and provides results that are more intuitive to a human observer. Peak Caller 2 (“small peaks”) provides a finer spatial resolution, and corresponds to the settings used in all of the analyses in this study. The peak caller takes as input a BED file with the 3′ ends of reads

(single-nucleotide resolution), as well as a GTF file with transcript and exon annotations and a list of transcripts to analyse.
It functions by calculating a sliding average of read density within each transcript (window size 5/21 for small/large; only reads mapping to the same strand as the annotated transcript are considered). It then randomly shuffles the positions of the reads within the transcript and recalculates the sliding averages to determine the random expectation. This can be repeated several times (5 in this study) for more robustness. Windows obtained with the true read distribution are called as significant if their read density is higher than the 99$^{th}$ percentile of the simulated windows. Note that in this study, we used a setting whereby this threshold is calculated separately for each exon (and its upstream intron, for exons other than the first), by excluding the intron-exon pair of interest and reads overlapping it during the simulation step. This is necessary so that when calling peaks within a given exon (and its upstream intron), the threshold set would not be affected by the reads within that particular exon and its upstream intron. This way, for instance, the calling of a peak in the beginning of the exon is not affected by the calling of a peak in the middle of the same exon (except through potential merging, see below). After the initial peaks are called, they are filtered to remove peaks where more than 90% of the reads come from a single nucleotide (probable PCR duplicates), that are shorter than 5 nucleotides, or that overlap with fewer reads than a specified threshold (10/5 for large/small). Finally, peaks that are within a specified distance of each-other (21/5 nucleotides for large/small) are merged together.

**Individual gene profiles**

Individual *d*NET-seq gene profiles were generated by separating reads by strand using SAMtools (version 1.7) (Li et al., 2009). Strand-separated read data was converted to bedGraph format using bedtools genomecov with the -bg flag (version 2.27.1-1-gb87c465) (Quinlan and Hall, 2010). Coverage values were normalized per nucleotide accounting for the total number of uniquely aligned reads and with the scale set to reads per $10^8$ sequences. The outcome was converted to bigwig files through the bedGraphToBigWig tool (Kent et al., 2010) and uploaded to the UCSC genome browser (James Kent et al., 2002).

**Metagene analysis**

The read density metagene plots in Figures 3.11 and 3.13 were created with deepTools (version 3.0.2) (Ramírez et al., 2016). Metagenes with normalized gene size (Figure 3.11) have bins of 10 bp while all other metagene plots in this study have single nucleotide resolution. Normalized gene and intron lengths (Figures 3.11, 3.12 and Figure 3.13 - figure supplement 1) were obtained through the scale-regions option. Exon-intron junctions without normalized lengths were obtained using the reference-option set to the 3′SS or the 5′SS. For normalization, we divided the number of reads at each nucleotide (or bin) by the total number of reads in the entire genomic region under analysis. These values were then used to calculate the mean for each nucleotide, and the results were plotted in an arbitrary units (A.U.) ranging from 0 to 1.

The peak density metagene plots in Figures 3.12 and 3.15 were prepared using custom Python and R scripts. The peak density value represents the proportion of introns/exons that overlap with a peak at that position. Only internal fully coding exons that were at least 100 nucleotides long were included. For Figure 3.15D-F, further filtering based on read coverage was performed (see below). Note that exons shorter than 150 nucleotides contribute both to the upstream and downstream exonic proportion of the plot.

**Immediate splicing analysis**

The immediate splicing analysis was performed solely on the S5P data sets. Only the 117 previously annotated pre-MBT genes (Chen et al., 2013) were analysed for early data sets. Reads were considered as spliced if they contained ‘N’s in the CIGAR string, in a position corresponding to an annotated intron. Reads that overlapped both the (unspliced) intron and the downstream exon were considered as unspliced. In both cases, only reads where the 3′ end was located at least 5 nt downstream of the 3′ ss were included, to avoid analysing misaligned reads whose 3′ end should have mapped to the end of the upstream exon instead. Spliced reads that had the 5′ end mapped to the upstream exon and the 3′ end mapped to the intron were considered indicative of recursive splicing if the first nucleotide of the downstream end (indicative of the ratchet point position) matched the second G in the AGGT canonical splicing motif. If the last nucleotide of a read matched the last nucleotide of an exon, it was considered a splicing intermediate read and not representative of nascent RNA. A splicing ratio was

calculated by dividing the number of nascent RNA spliced reads by the sum of the number of spliced and unspliced nascent RNA reads, only including reads whose 3′ ends mapped to the first 100 nt of the downstream exon. Only fully coding internal exons at least 100 nt long were considered (exceptionally, in Figure 3.14 C, J and K, the 3′ most coding exon was also analysed). Finally, we performed filtering to remove exons where the read coverage was too low to allow for robust estimation of the splicing ratio. The relevant threshold was calculated for each dataset separately. We calculated the total proportion of spliced reads out of all spliced/unspliced reads for the dataset to obtain the expected splicing ratio. We then performed a binomial test to know the probability of sampling only spliced/unspliced reads by chance under the null that the true splicing ratio equalled this expectation. We set the threshold as the lowest number of reads that had to be sampled for the probability to be below 0.01. Through his procedure, the threshold was set at >=10 reads for replicate 1 and 2 of the late data set (no terminal coding exons), at >=11/10 reads for replicate 1/2 of the late data set including terminal coding exons, and at >=14/9 reads for replicate 1/2 of the early data set.

**Gene architecture and nucleotide composition analysis**

Gene architecture and nucleotide composition parameters were calculated using custom Python and R scripts based on Ensembl annotations for dm6.18 (Cunningham et al., 2019). Splice site strength scores were calculated using MaxEntScan (Yeo and Burge, 2004) with default parameters.

**Data availability**

Data generated in this study is not yet publicly available. The Python code used is available at https://github.com/kennyrebelo.

# References

Alexander, R.D., Innocente, S.A., Barrass, J.D., and Beggs, J.D. (2010). Splicing-Dependent RNA Polymerase Pausing in Yeast. Mol. Cell *40*, 582–593.

Ali-Murthy, Z., Lott, S.E., Eisen, M.B., and Kornberg, T.B. (2013a). An Essential Role for Zygotic Expression in the Pre-Cellular Drosophila Embryo. PLoS Genet. *9*, e1003428.

Ali-Murthy, Z., Lott, S.E., Eisen, M.B., and Kornberg, T.B. (2013b). An Essential Role for Zygotic Expression in the Pre-Cellular Drosophila Embryo. PLoS Genet. *9*.

Allen, M.A., Hillier, L.W., Waterston, R.H., and Blumenthal, T. (2011). A global analysis of C. elegans trans-splicing. Genome Res. *21*, 255–264.

de Almeida, S.F., and Carmo-Fonseca, M. (2012). Design principles of interconnections between chromatin and pre-mRNA splicing. Trends Biochem. Sci. *37*, 248–253.

de Almeida, S.F., Grosso, A.R., Koch, F., Fenouil, R., Carvalho, S., Andrade, J., Levezinho, H., Gut, M., Eick, D., Gut, I., et al. (2011). Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36. Nat. Struct. Mol. Biol. *18*, 977–983.

Alwine, J.C., Kemp, D.J., and Stark, G.R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc. Natl. Acad. Sci. *74*, 5350–5354.

Ameur, A., Zaghlool, A., Halvardson, J., Wetterbom, A., Gyllensten, U., Cavelier, L., and Feuk, L. (2011). Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. Nat. Struct. Mol. Biol. *18*, 1435–1440.

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., et al. (2012a). Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition. Cell Rep. *1*, 543–556.

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., et al. (2012b). Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition. Cell Rep. *1*, 543–556.

Andersen, P.R., Tirian, L., Vunjak, M., and Brennecke, J. (2017). A heterochromatin-dependent transcription machinery drives piRNA expression. Nature *549*, 54–59.

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

Artieri, C.G., and Fraser, H.B. (2014). Transcript Length Mediates Developmental Timing of Gene Expression Across Drosophila. Mol. Biol. Evol. *31*, 2879–2889.

Ashwal-Fluss, R., Meyer, M., Pamudurti, N.R., Ivanov, A., Bartok, O., Hanan, M., Evantal, N., Memczak, S., Rajewsky, N., and Kadener, S. (2014). circRNA Biogenesis Competes with Pre-mRNA Splicing. Mol. Cell *56*, 55–66.

Audibert, A., Weil, D., and Dautry, F. (2002). In Vivo Kinetics of mRNA Splicing and Transport in Mammalian Cells. Mol. Cell. Biol. *22*, 6706–6718.

Baillat, D., and Wagner, E.J. (2015). Integrator: surprisingly diverse functions in gene expression. Trends Biochem. Sci. *40*, 257–264.

Barker, D.D., Wang, C., Moore, J., Dickinson, L.K., and Lehmann, R. (1992). Pumilio is essential for function but not for distribution of the Drosophila abdominal determinant Nanos. Genes Dev. *6*, 2312–2326.

Barne, K.A., Bown, J.A., Busby, S.J., and Minchin, S.D. (1997). Region 2.5 of the Escherichia coli RNA polymerase sigma70 subunit is responsible for the recognition of the "extended-10" motif at promoters. EMBO J. *16*, 4034–4040.

Barrett, S.P., and Salzman, J. (2016). Circular RNAs: analysis, expression and potential functions. Development *143*, 1838–1847.

Bauman, J.G.J., Wiegant, J., Van Duijn, P., Lubsen, N.H., Sondermeijer, P.J.A., Hennig, W., and Kubli, E. (1981). Rapid and high resolution detection of in situ hybridisation to polytene chromosomes using fluorochrome-labeled RNA. Chromosoma *84*, 1–18.

Berg, M.G., Singh, L.N., Younis, I., Liu, Q., Pinto, A.M., Kaida, D., Zhang, Z., Cho, S., Sherrill-Mix, S., Wan, L., et al. (2012). U1 snRNP Determines mRNA Length and Regulates Isoform Expression. Cell *150*, 53–64.

Berget, S.M. (1995). Exon Recognition in Vertebrate Splicing. J. Biol. Chem. *270*, 2411–2414.

Berget, S.M., Moore, C., and Sharp, P.A. (1977). Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. Proc. Natl. Acad. Sci. *74*, 3171–3175.

Bertrand, E., Chartrand, P., Schaefer, M., Shenoy, S.M., Singer, R.H., and Long, R.M. (1998). Localization of ASH1 mRNA Particles in Living Yeast. Mol. Cell *2*, 437–445.

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

Bieberstein, N.I., Carrillo Oesterreich, F., Straube, K., and Neugebauer, K.M. (2012). First Exon Length Controls Active Chromatin Signatures and Transcription. Cell Rep. *2*, 62–68.

Biedler, J.K., Hu, W., Tae, H., and Tu, Z. (2012). Identification of Early Zygotic Genes in the Yellow Fever Mosquito Aedes aegypti and Discovery of a Motif Involved in Early Zygotic Genome Activation. PLoS ONE *7*, e33933.

Bird, G., Zorio, D.A.R., and Bentley, D.L. (2004). RNA Polymerase II Carboxy-Terminal Domain Phosphorylation Is Required for Cotranscriptional Pre-mRNA Splicing and 3′-End Formation. Mol. Cell. Biol. *24*, 8963–8969.

Birse, C.E. (1997). Transcriptional termination signals for RNA polymerase II in fission yeast. EMBO J. *16*, 3633–3643.

Blythe, S.A., and Wieschaus, E.F. (2015). Zygotic Genome Activation Triggers the DNA Replication Checkpoint at the Midblastula Transition. Cell *160*, 1169–1181.

Boehning, M., Dugast-Darzacq, C., Rankovic, M., Hansen, A.S., Yu, T., Marie-Nelly, H., McSwiggen, D.T., Kokic, G., Dailey, G.M., Cramer, P., et al. (2018). RNA polymerase II clustering through carboxy-terminal domain phase separation. Nat. Struct. Mol. Biol. *25*, 833–840.

Bonnet, A., Grosso, A.R., Elkaoutari, A., Coleno, E., Presle, A., Sridhara, S.C., Janbon, G., Géli, V., de Almeida, S.F., and Palancade, B. (2017). Introns Protect Eukaryotic Genomes from Transcription-Associated Genetic Instability. Mol. Cell *67*, 608-621.e6.

Boothby, T.C., Zipper, R.S., van der Weele, C.M., and Wolniak, S.M. (2013). Removal of Retained Introns Regulates Translation in the Rapidly Developing Gametophyte of Marsilea vestita. Dev. Cell *24*, 517–529.

Borok, M.J., Tran, D.A., Ho, M.C.W., and Drewell, R.A. (2010). Dissecting the regulatory switches of development: lessons from enhancer evolution in Drosophila. Development *137*, 5–13.

ten Bosch, J.R., Benavides, J.A., and Cline, T.W. (2006). The TAGteam DNA motif controls the timing of Drosophila pre-blastoderm transcription. Dev. Camb. Engl. *133*, 1967–1977.

Bothma, J.P., Garcia, H.G., Esposito, E., Schlissel, G., Gregor, T., and Levine, M. (2014). Dynamic regulation of eve stripe 2 expression reveals transcriptional bursts in living Drosophila embryos. Proc. Natl. Acad. Sci. *111*, 10598–10603.

Boutz, P.L., Bhutkar, A., and Sharp, P.A. (2015). Detained introns are a novel, widespread class of post-transcriptionally spliced introns. Genes Dev. *29*, 63–80.

Brandt, A., Papagiannouli, F., Wagner, N., Wilsch-Bräuninger, M., Braun, M., Furlong, E.E., Loserth, S., Wenzl, C., Pilot, F., Vogt, N., et al. (2006). Developmental Control of Nuclear Size and Shape by kugelkern and kurzkern. Curr. Biol. *16*, 543–552.

Brody, Y., Neufeld, N., Bieberstein, N., Causse, S.Z., Böhnlein, E.-M., Neugebauer, K.M., Darzacq, X., and Shav-Tal, Y. (2011). The In Vivo Kinetics of RNA Polymerase II Elongation during Co-Transcriptional Splicing. PLoS Biol. *9*, e1000573.

Browning, D.F., and Busby, S.J. (2004). The regulation of bacterial transcription initiation. Nat. Rev. Microbiol. *2*, 57–65.

Brueckner, F., Hennecke, U., Carell, T., and Cramer, P. (2007). CPD Damage Recognition by Transcribing RNA Polymerase II. Science *315*, 859–862.

Brugiolo, M., Herzel, L., and Neugebauer, K.M. (2013). Counting on co-transcriptional splicing. F1000Prime Rep. *5*.

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.

Bushnell, B., Rood, J., and Singer, E. (2017). BBMerge – Accurate paired shotgun read merging via overlap. PLoS ONE *12*, 1–15.

Butler, J.E.F., and Kadonaga (2002). The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev. *16*, 2583–2592.

Campos-Ortega, J.A., and Hartenstein, V. (1985). The Embryonic Development of Drosophila melanogaster (Berlin, Heidelberg: Springer Berlin Heidelberg).

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.

Carrillo Oesterreich, F., Herzel, L., Straube, K., Hujer, K., Howard, J., and Neugebauer, K.M. (2016). Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. Cell *165*, 372–381.

Carrillo Oesterreich, F., Herzel, L., Straube, K., Hujer, K., Howard, J., and Neugebauer, K.M. (2016). Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. Cell *165*, 372–381.

Chan, S.-P. (2003). The Prp19p-Associated Complex in Spliceosome Activation. Science *302*, 279–282.

Chan, Y.A., Hieter, P., and Stirling, P.C. (2014). Mechanisms of genome instability induced by RNA-processing defects. Trends Genet. *30*, 245–253.

Chanarat, S., Seizl, M., and Strasser, K. (2011). The Prp19 complex is a novel transcription elongation factor required for TREX occupancy at transcribed genes. Genes Dev. *25*, 1147–1158.

Chang, H., Lim, J., Ha, M., and Kim, V.N. (2014). TAIL-seq: Genome-wide Determination of Poly(A) Tail Length and 3′ End Modifications. Mol. Cell *53*, 1044–1052.

Chang, K.-J., Chen, H.-C., and Cheng, S.-C. (2009). Ntc90 is required for recruiting first step factor Yju2 but not for spliceosome activation. RNA *15*, 1729–1739.

Chari, S., Wilky, H., Govindan, J., and Amodeo, A.A. (2019). Histone concentration regulates the cell cycle and transcription in early development. Development *146*, dev177402.

Chen, L.-L. (2016). The biogenesis and emerging roles of circular RNAs. Nat. Rev. Mol. Cell Biol. *17*, 205–211.

Chen, I.T., and Chasin, L.A. (1994). Large exon size does not limit splicing in vivo. Mol. Cell. Biol. *14*, 2140–2146.

Chen, K., Johnston, J., Shao, W., Meier, S., Staber, C., and Zeitlinger, J. (2013). A global change in RNA polymerase II pausing during the Drosophila midblastula transition. ELife *2013*, 1–19.

Cho, W.-K., Spille, J.-H., Hecht, M., Lee, C., Li, C., Grube, V., and Cisse, I.I. (2018). Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. Science *361*, 412–415.

Chou, T.B., and Perrimon, N. (1992). Use of a yeast site-specific recombinase to produce female germline chimeras in Drosophila. Genetics *131*, 643–653.

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

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

Clamer, M., Höfler, L., Mikhailova, E., Viero, G., and Bayley, H. (2014). Detection of 3′-End RNA Uridylation with a Protein Nanopore. ACS Nano *8*, 1364–1374.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008a). Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. Science *322*, 1845–1848.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008b). Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation. Science *322*, 1845–1849.

Cramer, P. (2019). Organization and regulation of gene transcription. Nature *573*, 45–54.

Csuros, M., Rogozin, I.B., and Koonin, E.V. (2008). Extremely Intron-Rich Genes in the Alveolate Ancestors Inferred with a Flexible Maximum-Likelihood Approach. Mol. Biol. Evol. *25*, 903–911.

Csuros, M., Rogozin, I.B., and Koonin, E.V. (2011). A Detailed History of Intron-rich Eukaryotic Ancestors Inferred from a Global Survey of 100 Complete Genomes. PLoS Comput. Biol. *7*, e1002150.

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M.R., Armean, I.M., Bennett, R., Bhai, J., Billis, K., Boddu, S., et al. (2019). Ensembl 2019. Nucleic Acids Res. *47*, D745–D751.

Custódio, N., Vivo, M., Antoniou, M., and Carmo-Fonseca, M. (2007). Splicing- and cleavage-independent requirement of RNA polymerase II CTD for mRNA release from the transcription site. J. Cell Biol. *179*, 199–207.

Dahlberg, O., Shilkova, O., Tang, M., Holmqvist, P.H., and Mannervik, M. (2015). P-TEFb, the Super Elongation Complex and Mediator Regulate a Subset of Non-paused Genes during Early Drosophila Embryo Development. PLoS Genet. *11*, 1–25.

David, C.J., Boyne, A.R., Millhouse, S.R., and Manley, J.L. (2011). The RNA polymerase II C-terminal domain promotes splicing activation through recruitment of a U2AF65-Prp19 complex. Genes Dev. *25*, 972–983.

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

De, I., Bessonov, S., Hofele, R., dos Santos, K., Will, C.L., Urlaub, H., Lührmann, R., and Pena, V. (2015). The RNA helicase Aquarius exhibits structural adaptations mediating its recruitment to spliceosomes. Nat. Struct. Mol. Biol. *22*, 138–144.

De Conti, L., Baralle, M., and Buratti, E. (2013). Exon and intron definition in pre-mRNA splicing: Exon and intron definition in pre-mRNA splicing. Wiley Interdiscip. Rev. RNA *4*, 49–60.

De Renzis, S., Elemento, O., Tavazoie, S., and Wieschaus, E.F. (2007a). Unmasking Activation of the Zygotic Genome Using Chromosomal Deletions in the Drosophila Embryo. PLoS Biol. *5*, e117.

Dedrick, R.L., Kane, C.M., and Chamberlin, M.J. (1987). Purified RNA polymerase II recognizes specific termination sites during transcription in vitro. J. Biol. Chem. *262*, 9098–9108.

Derti, A., Garrett-Engele, P., MacIsaac, K.D., Stevens, R.C., Sriram, S., Chen, R., Rohl, C.A., Johnson, J.M., and Babak, T. (2012). A quantitative atlas of polyadenylation in five mammals. Genome Res. *22*, 1173–1183.

Desterro, J., Bak-Gordon, P., and Carmo-Fonseca, M. (2020). Targeting mRNA processing as an anticancer strategy. Nat. Rev. Drug Discov. *19*, 112–129.

Dingwall, C., Lomonossoff, G.P., and Laskey, R.A. (1981). High sequence specificity of micrococcal nuclease. Nucleic Acids Res. *9*, 2659–2674.

Di Talia, S., She, R., Blythe, S.A., Lu, X., Zhang, Q.F., and Wieschaus, E.F. (2013). Posttranslational Control of Cdc25 Degradation Terminates Drosophila's Early Cell-Cycle Program. Curr. Biol. *23*, 127–132.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Doe, C.Q., Smouse, D., and Goodman, C.S. (1988). Control of neuronal fate by the Drosophila segmentation gene even-skipped. Nature *333*, 376–378.

Drexler, H.L., Choquet, K., and Churchman, L.S. (2020). Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. Mol. Cell *77*, 985-998.e8.

Duff, M.O., Olson, S., Wei, X., Garrett, S.C., Osman, A., Bolisetty, M., Plocik, A., Celniker, S.E., and Graveley, B.R. (2015). Genome-wide identification of zero nucleotide recursive splicing in Drosophila. Nature *521*, 376–379.

Dye, M.J., and Proudfoot, N.J. (1999). Terminal Exon Definition Occurs Cotranscriptionally and Promotes Termination of RNA Polymerase II. Mol. Cell *3*, 371–378.

Dye, M.J., and Proudfoot, N.J. (2001). Multiple Transcript Cleavage Precedes Polymerase Release in Termination by RNA Polymerase II. Cell *105*, 669–681.

Dye, M.J., Gromak, N., and Proudfoot, N.J. (2006). Exon Tethering in Transcription by RNA Polymerase II. Mol. Cell *21*, 849–859.

Edgar, B. (1986). Parameters controlling transcriptional activation during early drosophila development. Cell *44*, 871–877.

Edgar, B.A., and Datar, S.A. (1996). Zygotic degradation of two maternal Cdc25 mRNAs terminates Drosophila's early cell cycle program. Genes Dev. *10*, 1966–1977.

Edgar, B.A., Lehman, D.A., and O'Farrell, P.H. (1994). Transcriptional regulation of string (cdc25): a link between developmental programming and the cell cycle. Dev. Camb. Engl. *120*, 3131–3143.

Emili, A., Shales, M., McCRACKEN, S., Xie, W., Tucker, P.W., Kobayashi, R., Blencowe, B.J., and Ingles, C.J. (2002). Splicing and transcription-associated proteins PSF and p54nrb/NonO bind to the RNA polymerase II CTD. RNA *8*, 1102–1111.

Erickson, J.W., and Cline, T.W. (1993). A bZIP protein, sisterless-a, collaborates with bHLH transcription factors early in Drosophila development to determine sex. Genes Dev. *7*, 1688–1702.

Fabrizio, P., Dannenberg, J., Dube, P., Kastner, B., Stark, H., Urlaub, H., and Lührmann, R. (2009). The Evolutionarily Conserved Core Design of the Catalytic Activation Step of the Yeast Spliceosome. Mol. Cell *36*, 593–608.

Farrell, J.A., and O'Farrell, P.H. (2013). Mechanism and Regulation of Cdc25/Twine Protein Destruction in Embryonic Cell-Cycle Remodeling. Curr. Biol. *23*, 118–126.

Farrell, J.A., Shermoen, A.W., Yuan, K., and O'Farrell, P.H. (2012). Embryonic onset of late replication requires Cdc25 down-regulation. Genes Dev. *26*, 714–725.

Femino, A.M. (1998). Visualization of Single RNA Transcripts in Situ. Science *280*, 585–590.

Fey, M.F., Kulozik, A.E., Hansen-Hagge, T.E., and Tobler, A. (1991). The polymerase chain reaction: A new tool for the detection of minimal residual disease in haematological malignancies. Eur. J. Cancer Clin. Oncol. *27*, 89–94.

Fischer, T. (2002). The mRNA export machinery requires the novel Sac3p-Thp1p complex to dock at the nucleoplasmic entrance of the nuclear pores. EMBO J. *21*, 5843–5852.

Foe, V.E. (1989). Mitotic domains reveal early commitment of cells in Drosophila embryos. Dev. Camb. Engl. *107*, 1–22.

Foe, V.E., and Alberts, B.M. (1983). Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles that precede gastrulation in Drosophila embryogenesis. J. Cell Sci. *61*, 31–70.

Fourmann, J.-B., Schmitzova, J., Christian, H., Urlaub, H., Ficner, R., Boon, K.-L., Fabrizio, P., and Luhrmann, R. (2013). Dissection of the factor requirements for spliceosome disassembly and the elucidation of its dissociation products using a purified splicing system. Genes Dev. *27*, 413–428.

Fox-Walsh, K.L., Dou, Y., Lam, B.J., Hung, S. -p., Baldi, P.F., and Hertel, K.J. (2005). The architecture of pre-mRNAs affects mechanisms of splice-site pairing. Proc. Natl. Acad. Sci. *102*, 16176–16181.

Fuchs, G., Voichek, Y., Benjamin, S., Gilad, S., Amit, I., and Oren, M. (2014). 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. Genome Biol. *15*, R69.

Fuchs, G., Voichek, Y., Rabani, M., Benjamin, S., Gilad, S., Amit, I., and Oren, M. (2015). Simultaneous measurement of genome-wide transcription elongation speeds and rates of RNA polymerase II transition into active elongation with 4sUDRB-seq. Nat. Protoc. *10*, 605–618.

Fujioka, M. (1999). Regulation of novel eve pattern elements. 12.

Fukaya, T., Lim, B., and Levine, M. (2017). Rapid Rates of Pol II Elongation in the Drosophila Embryo. Curr. Biol. *27*, 1387–1391.

Furger, A. (2002). Promoter proximal splice sites enhance transcription. Genes Dev. *16*, 2792–2799.

Gaertner, B., and Zeitlinger, J. (2014). RNA polymerase II pausing during development. Development *141*, 1179–1183.

Gaffney, D.J., McVicker, G., Pai, A.A., Fondufe-Mittendorf, Y.N., Lewellen, N., Michelini, K., Widom, J., Gilad, Y., and Pritchard, J.K. (2012). Controls of Nucleosome Positioning in the Human Genome. PLoS Genet. *8*, 1–13.

Gaillard, H., Wellinger, R.E., and Aguilera, A. (2007). A new connection of mRNP biogenesis and export with transcription-coupled repair. Nucleic Acids Res. *35*, 3893–3906.

Gall, J.G., and Pardue, M.L. (1969). FORMATION AND DETECTION OF RNA-DNA HYBRID MOLECULES IN CYTOLOGICAL PREPARATIONS. Proc. Natl. Acad. Sci. *63*, 378–383.

Garalde, D.R., Snell, E.A., Jachimowicz, D., Sipos, B., Lloyd, J.H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., et al. (2018). Highly parallel direct RNA sequencing on an array of nanopores. Nat. Methods *15*, 201–206.

Gariglio, P., Bellard, M., and Chambon, P. (1981). Clustering of RNA polymerase B molecules in the 5′ moiety of the adult β-globin gene of hen erythrocytes. Nucleic Acids Res. *9*, 2589–2598.

Gelfman, S., Cohen, N., Yearim, A., and Ast, G. (2013). DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure. Genome Res. *23*, 789–799.

Gilmour, D.S., and Lis, J.T. (1985). In vivo interactions of RNA polymerase II with genes of Drosophila melanogaster. Mol. Cell. Biol. *5*, 2009–2018.

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.

Gonatopoulos-Pournatzis, T., and Cowling, V.H. (2014). Cap-binding complex (CBC). Biochem. J. *457*, 231–242.

Gornemann, J., Barrandon, C., Hujer, K., Rutz, B., Rigaut, G., Kotovic, K.M., Faux, C., Neugebauer, K.M., and Seraphin, B. (2011). Cotranscriptional spliceosome assembly and splicing are independent of the Prp40p WW domain. RNA *17*, 2119–2129.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W., et al. (2011). The developmental transcriptome of Drosophila melanogaster. Nature *471*, 473–479.

Greger, I.H. (1998). Poly(A) signals control both transcriptional termination and initiation between the tandem GAL10 and GAL7 genes of Saccharomyces cerevisiae. EMBO J. *17*, 4771–4779.

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., Leite, A.P., Carvalho, S., Matos, M.R., Martins, F.B., Vítor, A.C., Desterro, J.M., Carmo-Fonseca, M., and de Almeida, S.F. (2015). Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. ELife *4*.

Gu, B., Eick, D., and Bensaude, O. (2013). CTD serine-2 plays a critical role in splicing and termination factor recruitment to RNA polymerase II in vivo. Nucleic Acids Res. *41*, 1591–1603.

Guilgur, L.G., Prudencio, P., Ferreira, T., Pimenta-Marques, A.R., and Martinho, R.G. (2012). Drosophila aPKC is required for mitotic spindle orientation during symmetric division of epithelial cells. Development *139*, 503–513.

Guilgur, L.G., Prudêncio, P., Sobral, D., Liszekova, D., Rosa, A., and Martinho, R.G. (2014). Requirement for highly efficient pre-mRNA splicing during Drosophila early embryonic development.

Gullerova, M., and Proudfoot, N.J. (2012). Convergent transcription induces transcriptional gene silencing in fission yeast and mammalian cells. Nat. Struct. Mol. Biol. *19*, 1193–1201.

Guo, R., Zheng, L., Park, J.W., Lv, R., Chen, H., Jiao, F., Xu, W., Mu, S., Wen, H., Qiu, J., et al. (2014). BS69/ZMYND11 Reads and Connects Histone H3.3 Lysine 36 Trimethylation-Decorated Chromatin to Regulated Pre-mRNA Processing. Mol. Cell *56*, 298–310.

Guo, Y.E., Manteiga, J.C., Henninger, J.E., Sabari, B.R., Dall'Agnese, A., Hannett, N.M., Spille, J.-H., Afeyan, L.K., Zamudio, A.V., Shrinivas, K., et al. (2019). Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. Nature *572*, 543–548.

Hansen, T.B., Jensen, T.I., Clausen, B.H., Bramsen, J.B., Finsen, B., Damgaard, C.K., and Kjems, J. (2013). Natural RNA circles function as efficient microRNA sponges. Nature *495*, 384–388.

Hardenbol, P., and Van Dyke, M.W. (1992). In vitro inhibition of c-myc transcription by mithramycin. Biochem. Biophys. Res. Commun. *185*, 553–558.

Hare, M.P. (2003). High Intron Sequence Conservation Across Three Mammalian Orders Suggests Functional Constraints. Mol. Biol. Evol. *20*, 969–978.

Harlen, K.M., and Churchman, L.S. (2017). The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain. Nat. Rev. Mol. Cell Biol. *18*, 263–273.

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

Harrison, M.M., Li, X.-Y., Kaplan, T., Botchan, M.R., and Eisen, M.B. (2011). Zelda Binding in the Early Drosophila melanogaster Embryo Marks Regions Subsequently Activated at the Maternal-to-Zygotic Transition. PLoS Genet. *7*, e1002266.

Hatton, A.R., Subramaniam, V., and Lopez, A.J. (1998a). Generation of Alternative Ultrabithorax Isoforms and Stepwise Removal of a Large Intron by Resplicing at Exon–Exon Junctions. Mol. Cell *2*, 787–796.

Hatton, A.R., Subramaniam, V., and Lopez, A.J. (1998b). Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions. Mol. Cell *2*, 787–796.

Haugen, S.P., Berkmen, M.B., Ross, W., Gaal, T., Ward, C., and Gourse, R.L. (2006). rRNA promoter regulation by nonoptimal binding of sigma region 1.2: an additional recognition element for RNA polymerase. Cell *125*, 1069–1082.

Helmann, J.D. (1999). Anti-sigma factors. Curr. Opin. Microbiol. *2*, 135–141.

Helmann, J.D., and deHaseth, P.L. (1999). Protein-nucleic acid interactions during open complex formation investigated by systematic alteration of the protein and DNA binding partners. Biochemistry *38*, 5959–5967.

Henriques, T., Ji, Z., Tan-Wong, S.M., Carmo, A.M., Tian, B., Proudfoot, N.J., and Moreira, A. (2012). Transcription termination between polo and snap, two closely spaced tandem genes of D. melanogaster. Transcription *3*, 198–212.

Hernandez, N. (1985). Formation of the 3’ end of U1 snRNA is directed by a conserved sequence located downstream of the coding region. EMBO J. *4*, 1827–1837.

Herold, N., Will, C.L., Wolf, E., Kastner, B., Urlaub, H., and Luhrmann, R. (2009). Conservation of the Protein Composition and Electron Microscopy Structure of Drosophila melanogaster and Human Spliceosomal Complexes. Mol. Cell. Biol. *29*, 281–301.

Herzel, L., Ottoz, D.S.M., Alpert, T., and Neugebauer, K.M. (2017). Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. Nat. Rev. Mol. Cell Biol. *18*, 637–650.

Herzel, L., Straube, K., and Neugebauer, K.M. (2018). Long-read sequencing of nascent RNA reveals coupling among RNA processing events. 1008–1019.

Heyn, P., Kircher, M., Dahl, A., Kelso, J., Tomancak, P., Kalinka, A.T., and Neugebauer, K.M. (2014a). The earliest transcribed zygotic genes are short, newly evolved, and different across species. Cell Rep. *6*, 285–292.

Heyn, P., Kircher, M., Dahl, A., Kelso, J., Tomancak, P., Kalinka, A.T., and Neugebauer, K.M. (2014b). The earliest transcribed zygotic genes are short, newly evolved, and different across species. Cell Rep. *6*, 285–292.

Higuchi, R., Fockler, C., Dollinger, G., and Watson, R. (1993). Kinetic PCR Analysis: Real-time Monitoring of DNA Amplification Reactions. Nat. Biotechnol. *11*, 1026–1030.

Hilgers, V., Lemke, S.B., and Levine, M. (2012). ELAV mediates 3' UTR extension in the Drosophila nervous system. Genes Dev. *26*, 2259–2264.

Hirose, Y., and Manley, J.L. (2000). RNA polymerase II and the integration of nuclear events. Genes Dev. *14*, 1415–1429.

Hobson, D.J., Wei, W., Steinmetz, L.M., and Svejstrup, J.Q. (2012). RNA Polymerase II Collision Interrupts Convergent Transcription. Mol. Cell *48*, 365–374.

Hogg, R., McGrail, J.C., and O'Keefe, R.T. (2010). The function of the NineTeen Complex (NTC) in regulating spliceosome conformations and fidelity during pre-mRNA splicing. Biochem. Soc. Trans. *38*, 1110–1115.

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

Hörz, W., and Altenburger, W. (1981). Sequence specific cleavage of DNA by micrococcal nuclease. Nucleic Acids Res. *9*, 2643–2658.

Hoskins, A.A., and Moore, M.J. (2012). The spliceosome: a flexible, reversible macromolecular machine. Trends Biochem. Sci. *37*, 179–188.

Hoskins, R.A., Landolin, J.M., Brown, J.B., Sandler, J.E., Takahashi, H., Lassmann, T., Yu, C., Booth, B.W., Zhang, D., Wan, K.H., et al. (2011). Genome-wide analysis of promoter architecture in Drosophila melanogaster. Genome Res. *21*, 182–192.

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

Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L.E., Sibley, C.R., Sugimoto, Y., Tajnik, M., König, J., and Ule, J. (2014). iCLIP: Protein–RNA interactions at nucleotide resolution. Methods *65*, 274–287.

Huranová, M., Ivani, I., Benda, A., Poser, I., Brody, Y., Hof, M., Shav-Tal, Y., Neugebauer, K.M., and Staněk, D. (2010). The differential interaction of snRNPs with pre-mRNA reveals splicing kinetics in living cells. J. Cell Biol. *191*, 75–86.

Iannone, C., and Valcárcel, J. (2013). Chromatin's thread to alternative splicing regulation. Chromosoma *122*, 465–474.

Iasillo, C., Schmid, M., Yahia, Y., Maqbool, M.A., Descostes, N., Karadoulama, E., Bertrand, E., Andrau, J.-C., and Jensen, T.H. (2017). ARS2 is a general suppressor of pervasive transcription. Nucleic Acids Res. *45*, 10229–10241.

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.

Izaurralde, E., Lewis, J., McGuigan, C., Jankowska, M., Darzynkiewicz, E., and Mattaj, I.W. (1994). A nuclear cap binding protein complex involved in pre-mRNA splicing. Cell *78*, 657–668.

Jaeger, J. (2011). The gap gene network. Cell. Mol. Life Sci. *68*, 243–274.

James Kent, W., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

Jeck, W.R., Sorrentino, J.A., Wang, K., Slevin, M.K., Burd, C.E., Liu, J., Marzluff, W.F., and Sharpless, N.E. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA *19*, 141–157.

Jiao, X., Xiang, S., Oh, C., Martin, C.E., Tong, L., and Kiledjian, M. (2010). Identification of a quality-control mechanism for mRNA 5′-end capping. Nature *467*, 608–611.

Jimeno, S. (2002). The yeast THO complex and mRNA export factors link RNA metabolism with transcription and genome instability. EMBO J. *21*, 3526–3535.

John, H.A., Birnstiel, M.L., and Jones, K.W. (1969). RNA-DNA Hybrids at the Cytological Level. Nature *223*, 582–587.

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

Jonkers, I., and Lis, J.T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. Nat. Rev. Mol. Cell Biol. *16*, 167–177.

Jonkers, I., Kwak, H., and Lis, J.T. (2014a). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. ELife *3*.

Jonkers, I., Kwak, H., and Lis, J.T. (2014b). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. ELife *2014*, 1–25.

Joseph, B., Kondo, S., and Lai, E.C. (2018a). Short cryptic exons mediate recursive splicing in Drosophila. Nat. Struct. Mol. Biol. *25*, 365–371.

Joseph, B., Kondo, S., and Lai, E.C. (2018b). Short cryptic exons mediate recursive splicing in Drosophila. Nat. Struct. Mol. Biol. *25*, 365–371.

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.

Kamakaka, R.T., and Kadonaga, J.T. (1994). The soluble nuclear fraction, a highly efficient transcription extract from Drosophila embryos. Methods Cell Biol. *44*, 225–235.

Karim, F.D., and Thummel, C.S. (1991). Ecdysone coordinates the timing and amounts of E74A and E74B transcription in Drosophila. Genes Dev. *5*, 1067–1079.

Katahira, J., and Yoneda, Y. (2009). Roles of the TREX complex in nuclear export of mRNA. RNA Biol. *6*, 149–152.

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

Kellinger, M.W., Song, C.-X., Chong, J., Lu, X.-Y., He, C., and Wang, D. (2012). 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription. Nat. Struct. Mol. Biol. *19*, 831–833.

Kelly, S., Georgomanolis, T., Zirkel, A., Diermeier, S., O'Reilly, D., Murphy, S., Längst, G., Cook, P.R., and Papantonis, A. (2015). Splicing of many human genes involves sites embedded within introns. Nucleic Acids Res. *43*, 4721–4732.

Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S., and Karolchik, D. (2010). BigWig and BigBed: Enabling browsing of large distributed datasets. Bioinformatics *26*, 2204–2207.

Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: Diversification, exon definition and function. Nat. Rev. Genet. *11*, 345–355.

Khodor, Y.L., Rodriguez, J., Abruzzi, K.C., Tang, C.H.A., Marr, M.T., and Rosbash, M. (2011a). Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. Genes Dev. *25*, 2502–2512.

Khodor, Y.L., Rodriguez, J., Abruzzi, K.C., Tang, C.-H.A., Marr, M.T., and Rosbash, M. (2011b). Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. Genes Dev. *25*, 2502–2512.

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

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

Kimura, H., Tao, Y., Roeder, R.G., and Cook, P.R. (1999). Quantitation of RNA Polymerase II and Its Transcription Factors in an HeLa Cell: Little Soluble Holoenzyme but Significant Amounts of Polymerases Attached to the Nuclear Substructure. Mol. Cell. Biol. *19*, 5383–5392.

Koga, M., Hayashi, M., and Kaida, D. (2015). Splicing inhibition decreases phosphorylation level of Ser2 in Pol II CTD. Nucleic Acids Res. *43*, 8258–8267.

Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X.S., and Ahringer, J. (2009). Differential chromatin marking of introns and expressed exons by H3K36me3. Nat. Genet. *41*, 376–381.

Kornberg, R.D. (2005). Mediator and the mechanism of transcriptional activation. Trends Biochem. Sci. *30*, 235–239.

Kornblihtt, A.R., Schor, I.E., Allo, M., and Blencowe, B.J. (2009). When chromatin meets splicing. Nat. Struct. Mol. Biol. *16*, 902–903.

Korneta, I., and Bujnicki, J.M. (2012). Intrinsic Disorder in the Human Spliceosomal Proteome. PLoS Comput. Biol. *8*, e1002641.

Kramer, M.C., Liang, D., Tatomer, D.C., Gold, B., March, Z.M., Cherry, S., and Wilusz, J.E. (2015). Combinatorial control of *Drosophila* circular RNA expression by intronic repeats, hnRNPs, and SR proteins. Genes Dev. *29*, 2168–2182.

Kuehner, J.N., Pearson, E.L., and Moore, C. (2011). Unravelling the means to an end: RNA polymerase II transcription termination. Nat. Rev. Mol. Cell Biol. *12*, 283–294.

Kulaeva, O.I., Nizovtseva, E.V., Polikanov, Y.S., Ulianov, S.V., and Studitsky, V.M. (2012). Distant Activation of Transcription: Mechanisms of Enhancer Action. Mol. Cell. Biol. *32*, 4892–4897.

Kulish, D., and Struhl, K. (2001). TFIIS Enhances Transcriptional Elongation through an Artificial Arrest Site In Vivo. Mol. Cell. Biol. *21*, 4162–4168.

Kuraoka, I., Ito, S., Wada, T., Hayashida, M., Lee, L., Saijo, M., Nakatsu, Y., Matsumoto, M., Matsunaga, T., Handa, H., et al. (2008). Isolation of XAB2 Complex Involved in Pre-mRNA Splicing, Transcription, and Transcription-coupled Repair. J. Biol. Chem. *283*, 940–950.

Kwak, H., and Lis, J.T. (2013). Control of Transcriptional Elongation. Annu. Rev. Genet. *47*, 483–508.

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.

Kwasnieski, J.C., Orr-Weaver, T.L., and Bartel, D.P. (2019a). Early genome activation in *Drosophila* is extensive with an initial tendency for aborted transcripts and retained introns. Genome Res. *29*, 1188–1197.

Kwasnieski, J.C., Orr-Weaver, T.L., and Bartel, D.P. (2019b). Early genome activation in Drosophila is extensive with an initial tendency for aborted transcripts and retained introns. Genome Res. *29*, 1188–1197.

Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992). CpG islands as gene markers in the human genome. Genomics *13*, 1095–1107.

Larson, M.H., Mooney, R.A., Peters, J.M., Windgassen, T., Nayak, D., Gross, C.A., Block, S.M., Greenleaf, W.J., Landick, R., and Weissman, J.S. (2014). A pause sequence enriched at translation start sites drives transcription dynamics in vivo. Science *344*, 1042–1047.

Lasko, P. (2012). mRNA Localization and Translational Control in Drosophila Oogenesis. Cold Spring Harb. Perspect. Biol. *4*, a012294–a012294.

Lécuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., Hughes, T.R., Tomancak, P., and Krause, H.M. (2007). Global Analysis of mRNA Localization Reveals a Prominent Role in Organizing Cellular Architecture and Function. Cell *131*, 174–187.

Lee, T.I., and Young, R.A. (2013). Transcriptional Regulation and its Misregulation in Disease. Cell *152*, 1237–1251.

Lehninger, A.L., Nelson, D.L., and Cox, M.M. (2013). Lehninger principles of biochemistry (New York: W.H. Freeman).

Lejeune, F., and Maquat, L.E. (2005). Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. Curr. Opin. Cell Biol. *17*, 309–315.

Lesk, A.M. (1969). Why does DNA contain thymine and RNA uracil? J. Theor. Biol. *22*, 537–540.

Lewis, J.D., Izaurralde, E., Jarmolowski, A., McGuigan, C., and Mattaj, I.W. (1996). A nuclear cap-binding complex facilitates association of U1 snRNP with the cap-proximal 5' splice site. Genes Dev. *10*, 1683–1698.

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 *25*, 2078–2079.

Li, X.-Y., Harrison, M.M., Villalta, J.E., Kaplan, T., and Eisen, M.B. (2014). Establishment of regions of genomic activity during the Drosophila maternal to zygotic transition. ELife *3*.

Li, Y., Chen, Z.-Y., Wang, W., Baker, C.C., and Krug, R.M. (2001). The 3′-end-processing factor CPSF is required for the splicing of single-intron pre-mRNAs in vivo. RNA *7*, 920–931.

Liang, H.-L., Nien, C.-Y., Liu, H.-Y., Metzstein, M.M., Kirov, N., and Rushlow, C. (2008). The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. Nature *456*, 400–403.

Lim, L.P., and Burge, C.B. (2001). A computational analysis of sequence features involved in recognition of short introns. Proc. Natl. Acad. Sci. *98*, 11193–11198.

Long, M., and Deutsch, M. (1999). Intron—exon structures of eukaryotic model organisms. Nucleic Acids Res. *27*, 3219–3228.

Lott, S.E., Villalta, J.E., Schroth, G.P., Luo, S., Tonkin, L.A., and Eisen, M.B. (2011a). Noncanonical compensation of zygotic X transcription in early Drosophila melanogaster development revealed through single-embryo RNA-Seq. PLoS Biol. *9*.

Lott, S.E., Villalta, J.E., Schroth, G.P., Luo, S., Tonkin, L.A., and Eisen, M.B. (2011b). Noncanonical Compensation of Zygotic X Transcription in Early Drosophila melanogaster Development Revealed through Single-Embryo RNA-Seq. PLoS Biol. *9*, e1000590.

Lu, F., Portz, B., and Gilmour, D.S. (2019). The C-Terminal Domain of RNA Polymerase II Is a Multivalent Targeting Sequence that Supports Drosophila Development with Only Consensus Heptads. Mol. Cell *73*, 1232-1242.e4.

Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M., and Misteli, T. (2010). Regulation of Alternative Splicing by Histone Modifications. Science *327*, 996–1000.

Ludwig, M.Z., Manu, Kittler, R., White, K.P., and Kreitman, M. (2011). Consequences of Eukaryotic Enhancer Architecture for Gene Expression Dynamics, Development, and Fitness. PLoS Genet. *7*, e1002364.

Lund, E., and Dahlberg, J.E. (1992). Cyclic 2 ', 3 "-Phosphates and Nontemplated Nucleotides at the 3" End of Spliceosomal U6 Small Nuclear RNAs. Science *255*.

Madhani, H.D. (2013). The frustrated gene: origins of eukaryotic gene expression. Cell *155*, 744–749.

Makarova, O.V., Makarov, E.M., Urlaub, H., Will, C.L., Gentzel, M., Wilm, M., and Lührmann, R. (2004). A subset of human 35S U5 proteins, including Prp19, function prior to catalytic step 1 of splicing. EMBO J. *23*, 2381–2391.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal *17*, 10–12.

Martin, R.M., Rino, J., Carvalho, C., Kirchhausen, T., and Carmo-Fonseca, M. (2013). Live-Cell Visualization of Pre-mRNA Splicing with Single-Molecule Sensitivity. Cell Rep. *4*, 1144–1155.

Martinho, R.G., Guilgur, L.G., and Prudêncio, P. (2015). How gene expression in fast-proliferating cells keeps pace. BioEssays News Rev. Mol. Cell. Dev. Biol. *37*, 514–524.

Masuda, S. (2005). Recruitment of the human TREX complex to mRNA during splicing. Genes Dev. *19*, 1512–1517.

de la Mata, M., Alonso, C.R., Kadener, S., Fededa, J.P., Blaustein, M., Pelisch, F., Cramer, P., Bentley, D., and Kornblihtt, A.R. (2003). A slow RNA polymerase II affects alternative splicing in vivo. Mol. Cell *12*, 525–532.

Matlin, A.J., Clark, F., and Smith, C.W.J. (2005). Understanding alternative splicing: towards a cellular code. Nat. Rev. Mol. Cell Biol. *6*, 386–398.

Mattick, J.S., and Gagen, M.J. (2001). The Evolution of Controlled Multitasked Gene Networks: The Role of Introns and Other Noncoding RNAs in the Development of Complex Organisms. Mol. Biol. Evol. *18*, 1611–1630.

Mayer, A., Di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J.A., and Churchman, L.S. (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell *161*, 541–554.

McDaniel, S.L., Gibson, T.J., Schulz, K.N., Fernandez Garcia, M., Nevil, M., Jain, S.U., Lewis, P.W., Zaret, K.S., and Harrison, M.M. (2019). Continued Activity of the Pioneer Factor Zelda Is Required to Drive Zygotic Genome Activation. Mol. Cell *74*, 185-195.e4.

McGettigan, P.A. (2013). Transcriptomics in the RNA-seq era. Curr. Opin. Chem. Biol. *17*, 4–11.

McGinty, R.K., and Tan, S. (2015). Nucleosome Structure and Function. Chem. Rev. *115*, 2255–2273.

McGrail, J.C., Krause, A., and O'Keefe, R.T. (2009). The RNA binding protein Cwc2 interacts directly with the U6 snRNA to link the nineteen complex to the spliceosome during pre-mRNA splicing. Nucleic Acids Res. *37*, 4205–4217.

McKnight, S.L., and Miller Jr., O.L. (1976). Ultrastructural patterns of RNA synthesis during early embryogenesis of Drosophila melanogaster. Cell *8*, 305–319.

McManus, C.J., Duff, M.O., Eipper-Mains, J., and Graveley, B.R. (2010). Global analysis of trans-splicing in Drosophila. Proc. Natl. Acad. Sci. *107*, 12975–12979.

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S.D., Gregersen, L.H., Munschauer, M., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. Nature *495*, 333–338.

Miller, O.L., and Beatty, B.R. (1969). Visualization of Nucleolar Genes. Science *164*, 955–957.

Mitrea, D.M., and Kriwacki, R.W. (2016). Phase separation in biology; functional organization of a higher order. Cell Commun. Signal. *14*.

Moore, M.J., Query, C.C., and Sharp, P.A. (1993). Splicing of precursors to mRNAs by the spliceosome.

Morris, D.P., and Greenleaf, A.L. (2000). The Splicing Factor, Prp40, Binds the Phosphorylated Carboxyl-terminal Domain of RNA Polymerase II. J. Biol. Chem. *275*, 39935–39943.

Mourier, T. (2003). Eukaryotic Intron Loss. Science *300*, 1393–1393.

Müller, F., and Tora, L. (2014). Chromatin and DNA sequences in defining promoters for transcription initiation. Biochim. Biophys. Acta BBA - Gene Regul. Mech. *1839*, 118–128.

Muniz, L., Deb, M.K., Aguirrebengoa, M., Lazorthes, S., Trouche, D., and Nicolas, E. (2017). Control of Gene Expression in Senescence through Transcriptional Read-Through of Convergent Protein-Coding Genes. Cell Rep. *21*, 2433–2446.

Myer, V.E., and Young, R.A. (1998). RNA Polymerase II Holoenzymes and Subcomplexes. J. Biol. Chem. *273*, 27757–27760.

Naggan Perl, T., Schmid, B.G.M., Schwirz, J., and Chipman, A.D. (2013). The Evolution of the knirps Family of Transcription Factors in Arthropods. Mol. Biol. Evol. *30*, 1348–1357.

Nakatsu, Y., Asahina, H., Citterio, E., Rademakers, S., Vermeulen, W., Kamiuchi, S., Yeo, J.-P., Khaw, M.-C., Saijo, M., Kodo, N., et al. (2000). XAB2, a Novel Tetratricopeptide Repeat Protein Involved in Transcription-coupled DNA Repair and Transcription. J. Biol. Chem. *275*, 34931–34937.

Neves, G., Zucker, J., Daly, M., and Chess, A. (2004). Stochastic yet biased expression of multiple Dscam splice variants by individual cells. Nat. Genet. *36*, 240–246.

Ni, J.Z., Grate, L., Donohue, J.P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T.A., Blume, J.E., and Ares, M. (2007). Ultraconserved elements are associated with homeostatic

control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev. *21*, 708–718.

Nielsen, H., and Johansen, S.D. (2009). Group I introns: Moving in new directions. RNA Biol. *6*, 375–383.

Niwa, M., and Berget, S.M. (1991). Mutation of the AAUAAA polyadenylation signal depresses in vitro splicing of proximal but not distal introns. Genes Dev. *5*, 2086–2095.

Nogales, E., Louder, R.K., and He, Y. (2017). Structural Insights into the Eukaryotic Transcription Initiation Machinery. Annu. Rev. Biophys. *46*, 59–83.

Nojima, T., Gomes, T., Grosso, A.R.F., Kimura, H., Dye, M.J., Dhir, S., Carmo-Fonseca, M., and Proudfoot, N.J. (2015a). Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. Cell *161*, 526–540.

Nojima, T., Gomes, T., Grosso, A.R.F., Kimura, H., Dye, M.J., Dhir, S., Carmo-Fonseca, M., and Proudfoot, N.J. (2015b). Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. Cell *161*, 526–540.

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

Nojima, T., Rebelo, K., Gomes, T., Grosso, A.R., Proudfoot, N.J., and Carmo-Fonseca, M. (2018). RNA Polymerase II Phosphorylated on CTD Serine 5 Interacts with the Spliceosome during Co-transcriptional Splicing. Mol. Cell *72*, 369-379.e4.

Nonet, M.L., and Young, R.A. (1989). Intragenic and extragenic suppressors of mutations in the heptapeptide repeat domain of Saccharomyces cerevisiae RNA polymerase II. Genetics *123*, 715–724.

Nudler, E., Mustaev, A., Goldfarb, A., and Lukhtanov, E. (1997). The RNA–DNA Hybrid Maintains the Register of Transcription by Preventing Backtracking of RNA Polymerase. Cell *89*, 33–41.

Ohler, U., Liao, G., Niemann, H., and Rubin, G.M. (2002). [No title found]. Genome Biol. *3*, research0087.1.

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.

Pabis, M., Neufeld, N., Steiner, M.C., Bojic, T., Shav-Tal, Y., and Neugebauer, K.M. (2013). The nuclear cap-binding complex interacts with the U4/U6{middle dot}U5 tri-snRNP and promotes spliceosome assembly in mammalian cells. RNA *19*, 1054–1063.

Pai, A.A., Henriques, T., McCue, K., Burkholder, A., Adelman, K., and Burge, C.B. (2017). The kinetics of pre-mRNA splicing in the Drosophila genome: influence of gene architecture.

Pai, A.A., Paggi, J., Adelman, K., and Burge, C.B. (2018a). Numerous recursive sites contribute to accuracy of splicing of long introns in flies.

Pai, A.A., Paggi, J.M., Yan, P., Adelman, K., and Burge, C.B. (2018b). Numerous recursive sites contribute to accuracy of splicing in long introns in flies. PLoS Genet. *14*, 1–24.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40*, 1413–1415.

Pandya-Jones, A., and Black, D.L. (2009). Co-transcriptional splicing of constitutive and alternative exons. RNA *15*, 1896–1908.

Pankratz, M., Busch, M., Hoch, M., Seifert, E., and Jackle, H. (1992). Spatial control of the gap gene knirps in the Drosophila embryo by posterior morphogen system. Science *255*, 986–989.

Papasaikas, P., and Valcárcel, J. (2016). The Spliceosome: The Ultimate RNA Chaperone and Sculptor. Trends Biochem. Sci. *41*, 33–45.

Papasaikas, P., Tejedor, J.R., Vigevani, L., and Valcárcel, J. (2015). Functional Splicing Network Reveals Extensive Regulatory Potential of the Core Spliceosomal Machinery. Mol. Cell *57*, 7–22.

Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y. (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. Am. J. Hum. Genet. *102*, 11–26.

Patel, A.A., and Steitz, J.A. (2003). Splicing double: insights from the second spliceosome. Nat. Rev. Mol. Cell Biol. *4*, 960–970.

Peng, C. (1997). Mitotic and G2 Checkpoint Control: Regulation of 14-3-3 Protein Binding by Phosphorylation of Cdc25C on Serine-216. Science *277*, 1501–1505.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295.

Pilot, F. (2006). Developmental control of nuclear morphogenesis and anchoring by charleston, identified in a functional genomic screen of Drosophila cellularisation. Development *133*, 711–723.

Pimenta-Marques, A., Tostões, R., Marty, T., Barbosa, V., Lehmann, R., and Martinho, R.G. (2008). Differential requirements of a mitotic acetyltransferase in somatic and germ line cells. Dev. Biol. *323*, 197–206.

Porrua, O., Boudvillain, M., and Libri, D. (2016). Transcription Termination: Variations on Common Themes. Trends Genet. *32*, 508–522.

Prescott, E.M., and Proudfoot, N.J. (2002). Transcriptional collision between convergent genes in budding yeast. Proc. Natl. Acad. Sci. *99*, 8796–8801.

Price, D.H. (2018). Transient pausing by RNA polymerase II. Proc. Natl. Acad. Sci. *115*, 4810–4812.

Pritchard, D.K., and Schubiger, G. (1996). Activation of transcription in Drosophila embryos is a gradual process mediated by the nucleocytoplasmic ratio. Genes Dev. *10*, 1131–1142.

Proudfoot, N.J. (1976). Sequence analysis of the 3′ non-coding regions of rabbit α- and β-globin messenger RNAs. J. Mol. Biol. *107*, 491–525.

Proudfoot, N.J. (2011). Ending the message: poly(A) signals then and now. Genes Dev. *25*, 1770–1782.

Proudfoot, N.J. (2016). Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. Science *352*, aad9926–aad9926.

Prudêncio, P., Rebelo, K., Grosso, A.R., Martinho, R.G., and Carmo-Fonseca, M. (2019). Analysis of Mammalian Native Elongating Transcript sequencing (mNET-seq) high-throughput data. Methods 1–7.

Pu, M., Ni, Z., Wang, M., Wang, X., Wood, J.G., Helfand, S.L., Yu, H., and Lee, S.S. (2015). Trimethylation of Lys36 on H3 restricts gene expression change during aging and impacts life span. Genes Dev. *29*, 718–731.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Raddatz, G., Guzzardo, P.M., Olova, N., Fantappie, M.R., Rampp, M., Schaefer, M., Reik, W., Hannon, G.J., and Lyko, F. (2013). Dnmt2-dependent methylomes lack defined DNA methylation patterns. Proc. Natl. Acad. Sci. *110*, 8627–8631.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165.

Rasmussen, E.B., and Lis, J.T. (1993). In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. Proc. Natl. Acad. Sci. *90*, 7923–7927.

De Renzis, S., Elemento, O., Tavazoie, S., and Wieschaus, E.F. (2007b). Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo. PLoS Biol. *5*, 1036–1051.

Richard, P., and Manley, J.L. (2009). Transcription termination by nuclear RNA polymerases. Genes Dev. *23*, 1247–1269.

Richmond, T.J., and Davey, C.A. (2003). The structure of DNA in the nucleosome core. Nature *423*, 145–150.

Rienzo, M., and Casamassimi, A. (2016). Integrator complex and transcription regulation: Recent findings and pathophysiology. Biochim. Biophys. Acta BBA - Gene Regul. Mech. *1859*, 1269–1280.

Rigo, F., and Martinson, H.G. (2008). Functional Coupling of Last-Intron Splicing and 3′-End Processing to Transcription In Vitro: the Poly(A) Signal Couples to Splicing before Committing to Cleavage. Mol. Cell. Biol. *28*, 849–862.

Robberson, B.L., Cote, G.J., and Berget, S.M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol. Cell. Biol. *10*, 84–94.

Roeder, R.G., and Rutter, W.J. (1969). Multiple Forms of DNA-dependent RNA Polymerase in Eukaryotic Organisms. Nature *224*, 234–237.

Rogozin, I.B., Carmel, L., Csuros, M., and Koonin, E.V. (2012). Origin and evolution of spliceosomal introns. Biol. Direct *7*, 11.

Rondón, A.G., Jimeno, S., García-Rubio, M., and Aguilera, A. (2003). Molecular Evidence That the Eukaryotic THO/TREX Complex Is Required for Efficient Transcription Elongation. J. Biol. Chem. *278*, 39037–39043.

Rose, A.B. (2008). Intron-Mediated Regulation of Gene Expression. In Nuclear Pre-MRNA Processing in Plants, A.S.N. Reddy, and M. Golovkin, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 277–290.

Rosonina, E. (2004). Analysis of the requirement for RNA polymerase II CTD heptapeptide repeats in pre-mRNA splicing and 3'-end cleavage. RNA *10*, 581–589.

Ross, W., Gosink, K.K., Salomon, J., Igarashi, K., Zou, C., Ishihama, A., Severinov, K., and Gourse, R.L. (1993). A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. Science *262*, 1407–1413.

Rothe, M., Pehl, M., Taubert, H., and Jäckle, H. (1992). Loss of gene function through rapid mitotic cycles in the Drosophila embryo. Nature *359*, 156–159.

Rougvie, A.E., and Lis, J.T. (1988). The RNA polymerase II molecule at the 5′ end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. Cell *54*, 795–804.

Russell, C.S., Ben-Yehuda, S., Dix, I., Kupiec, M., and Beggs, J.D. (2000). Functional analyses of interacting factors involved in both pre-mRNA splicing and cell cycle progression in Saccharomyces cerevisiae. RNA N. Y. N *6*, 1565–1572.

Rymond, B.C., and Rosbash, M. (1985). Cleavage of 5′ splice site and lariat formation are independent of 3′ splice site in yeast mRNA splicing. Nature *317*, 735–737.

Saeki, H., and Svejstrup, J.Q. (2009). Stability, Flexibility, and Dynamic Interactions of Colliding RNA Polymerase II Elongation Complexes. Mol. Cell *35*, 191–205.

Sandler, J.E., Irizarry, J., Stepanik, V., Dunipace, L., Amrhein, H., and Stathopoulos, A. (2018). A Developmental Program Truncates Long Transcripts to Temporally Regulate Cell Signaling. Dev. Cell *47*, 773-784.e6.

Sanfilippo, P., Wen, J., and Lai, E.C. (2017). Landscape and evolution of tissue-specific alternative polyadenylation across Drosophila species. Genome Biol. *18*.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. *74*, 5463–5467.

Santos-Pereira, J.M., and Aguilera, A. (2015). R loops: new modulators of genome dynamics and function. Nat. Rev. Genet. *16*, 583–597.

Saunders, A., Core, L.J., Sutcliffe, C., Lis, J.T., and Ashe, H.L. (2013). Extensive polymerase pausing during Drosophila axis patterning enables high-level and pliable transcription. Genes Dev. *27*, 1146–1158.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. *103*, 1412–1417.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science *270*, 467–470.

Schilbach, S., Hantsche, M., Tegunov, D., Dienemann, C., Wigge, C., Urlaub, H., and Cramer, P. (2017). Structures of transcription pre-initiation complex with TFIIH and Mediator. Nature *551*, 204–209.

Schlackow, M., Nojima, T., Gomes, T., Dhir, A., Carmo-Fonseca, M., and Proudfoot, N.J. (2017). Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs. Mol. Cell *65*, 25–38.

Schmucker, D., Clemens, J.C., Shu, H., Worby, C.A., Xiao, J., Muda, M., Dixon, J.E., and Zipursky, S.L. (2000). Drosophila Dscam Is an Axon Guidance Receptor Exhibiting Extraordinary Molecular Diversity. Cell *101*, 671–684.

Schneider, M., Will, C.L., Anokhina, M., Tazi, J., Urlaub, H., and Lührmann, R. (2010). Exon Definition Complexes Contain the Tri-snRNP and Can Be Directly Converted into B-like Precatalytic Splicing Complexes. Mol. Cell *38*, 223–235.

Schones, D.E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic Regulation of Nucleosome Positioning in the Human Genome. Cell *132*, 887–898.

Schwalb, B., Michel, M., Zacher, B., Frühauf, K., Demel, C., Tresch, A., Gagneur, J., and Cramer, P. (2016). TT-seq maps the human transient transcriptome. Science *352*, 1225–1228.

Schwartz, S., Meshorer, E., and Ast, G. (2009). Chromatin organization marks exon-intron structure. Nat. Struct. Mol. Biol. *16*, 990–995.

Seoighe, C., and Korir, P.K. (2011). Evidence for intron length conservation in a set of mammalian genes associated with embryonic development. BMC Bioinformatics *12*.

Shah, S., Takei, Y., Zhou, W., Lubeck, E., Yun, J., Eng, C.-H.L., Koulena, N., Cronin, C., Karp, C., Liaw, E.J., et al. (2018). Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron seqFISH. Cell *174*, 363-376.e16.

Shao, W., Kim, H.-S., Cao, Y., Xu, Y.-Z., and Query, C.C. (2012). A U1-U2 snRNP Interaction Network during Intron Definition. Mol. Cell. Biol. *32*, 470–478.

Shcherbakova, I., Hoskins, A.A., Friedman, L.J., Serebrov, V., Corrêa, I.R., Xu, M.-Q., Gelles, J., and Moore, M.J. (2013). Alternative Spliceosome Assembly Pathways Revealed by Single-Molecule Fluorescence Microscopy. Cell Rep. *5*, 151–165.

Shearwin, K., Callen, B., and Egan, J. (2005). Transcriptional interference – a crash course. Trends Genet. *21*, 339–345.

Shermoen, A.W., and O'Farrell, P.H. (1991). Progression of the cell cycle through mitosis leads to abortion of nascent transcripts. Cell *67*, 303–310.

Shi, Y., and Manley, J.L. (2015). The end of the message: multiple protein–RNA interactions define the mRNA polyadenylation site. Genes Dev. *29*, 889–897.

Shin, C., and Manley, J.L. (2002). The SR Protein SRp38 Represses Splicing in M Phase Cells. Cell *111*, 407–417.

Sibley, C.R., Emmett, W., Blazquez, L., Faro, A., Haberman, N., Briese, M., Trabzuni, D., Ryten, M., Weale, M.E., Hardy, J., et al. (2015). Recursive splicing in long vertebrate genes. Nature *521*, 371–375.

Sibon, O.C.M., Stevenson, V.A., and Theurkauf, W.E. (1997). DNA-replication checkpoint control at the Drosophila midblastula transition. Nature *388*, 93–97.

Sims, R.J., Millhouse, S., Chen, C.-F., Lewis, B.A., Erdjument-Bromage, H., Tempst, P., Manley, J.L., and Reinberg, D. (2007). Recognition of Trimethylated Histone H3 Lysine 4 Facilitates the Recruitment of Transcription Postinitiation Factors and Pre-mRNA Splicing. Mol. Cell *28*, 665–676.

Singh, B.N., and Hampsey, M. (2007). A Transcription-Independent Role for TFIIB in Gene Looping. Mol. Cell *27*, 806–816.

Skourti-Stathaki, K., Kamieniarz-Gdula, K., and Proudfoot, N.J. (2014). R-loops induce repressive chromatin marks over mammalian gene terminators. Nature *516*, 436–439.

Small, S., Kraut, R., Hoey, T., Warrior, R., and Levine, M. (1991). Transcriptional regulation of a pair-rule stripe in Drosophila. Genes Dev. *5*, 827–839.

Smathers, C.M., and Robart, A.R. (2019). The mechanism of splicing as told by group II introns: Ancestors of the spliceosome. Biochim. Biophys. Acta BBA - Gene Regul. Mech. *1862*, 194390.

Solomon, M.J., Larsen, P.L., and Varshavsky, A. (1988). Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene. Cell *53*, 937–947.

Southern, E.M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. *98*, 503–517.

Spies, N., Nielsen, C.B., Padgett, R.A., and Burge, C.B. (2009). Biased Chromatin Signatures around Polyadenylation Sites and Exons. Mol. Cell *36*, 245–254.

Stein, J.A., Broihier, H.T., Moore, L.A., and Lehmann, R. (2002). Slow as molasses is required for polarized membrane growth and germ cell migration in Drosophila. Dev. Camb. Engl. *129*, 3925–3934.

Sterner, D.A., Carlo, T., and Berget, S.M. (1996). Architectural limits on split genes. Proc. Natl. Acad. Sci. *93*, 15081–15085.

Steurer, B., Janssens, R.C., Geverts, B., Geijer, M.E., Wienholz, F., Theil, A.F., Chang, J., Dealy, S., Pothof, J., van Cappellen, W.A., et al. (2018). Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA Polymerase II. Proc. Natl. Acad. Sci. *115*, E4368–E4376.

Sutton, R.E., and Boothroyd, J.C. (1986). Evidence for Trans splicing in trypanosomes. Cell *47*, 527–535.

Swinburne, I.A., and Silver, P.A. (2008). Intron Delays and Transcriptional Timing during Development. Dev. Cell *14*, 324–330.

Sykes, P.J., Neoh, S.H., Brisco, M.J., Hughes, E., Condon, J., and Morley, A.A. (1992). Quantitation of targets for PCR by use of limiting dilution. BioTechniques *13*, 444–449.

Tadros, W., and Lipshitz, H.D. (2009). The maternal-to-zygotic transition: a play in two acts. Development *136*, 3033–3042.

Takagaki, Y., Seipelt, R.L., Peterson, M.L., and Manley, J.L. (1996). The Polyadenylation Factor CstF-64 Regulates Alternative Processing of IgM Heavy Chain Pre-mRNA during B Cell Differentiation. Cell *87*, 941–952.

Takashima, Y., Ohtsuka, T., Gonzalez, A., Miyachi, H., and Kageyama, R. (2011). Intronic delay is essential for oscillatory expression in the segmentation clock. Proc. Natl. Acad. Sci. *108*, 3300–3305.

Talbert, P.B., Meers, M.P., and Henikoff, S. (2019). Old cogs, new tricks: the evolution of gene expression in a chromatin context. Nat. Rev. Genet. *20*, 283–297.

Talerico, M., and Berget, S.M. (1994). Intron definition in splicing of small Drosophila introns. Mol. Cell. Biol. *14*, 3434–3445.

Tatomer, D.C., Elrod, N.D., Liang, D., Xiao, M.-S., Jiang, J.Z., Jonathan, M., Huang, K.-L., Wagner, E.J., Cherry, S., and Wilusz, J.E. (2019). The Integrator complex cleaves nascent mRNAs to attenuate transcription. Genes Dev. *33*, 1525–1538.

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

The modENCODE Consortium, Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., et al. (2010). Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. Science *330*, 1787–1797.

Thummel, C.S., Burtis, K.C., and Hogness, D.S. (1990). Spatial and temporal patterns of E74 transcription during Drosophila development. Cell *61*, 101–111.

Tian, B., and Manley, J.L. (2017). Alternative polyadenylation of mRNA precursors. Nat. Rev. Mol. Cell Biol. *18*, 18–30.

Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcárcel, J., and Guigó, R. (2009). Nucleosome positioning as a determinant of exon recognition. Nat. Struct. Mol. Biol. *16*, 996–1001.

Tillo, D., and Hughes, T.R. (2009). G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics *10*.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Urano, Y., Iiduka, M., Sugiyama, A., Akiyama, H., Uzawa, K., Matsumoto, G., Kawasaki, Y., and Tashiro, F. (2006). Involvement of the Mouse *Prp19* Gene in Neuronal/Astroglial Cell Fate Decisions. J. Biol. Chem. *281*, 7498–7514.

Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R., and Johnson, R. (2018). Towards a complete map of the human long non-coding RNA transcriptome. Nat. Rev. Genet. *19*, 535–548.

Valencia, P., Dias, A.P., and Reed, R. (2008). Splicing promotes rapid and efficient mRNA export in mammalian cells. Proc. Natl. Acad. Sci. *105*, 3386–3391.

Vasil, V., Clancy, M., Ferl, R.J., Vasil, I.K., and Hannah, L.C. (1989). Increased Gene Expression by the First Intron of Maize *Shrunken-1* Locus in Grass Species. Plant Physiol. *91*, 1575–1579.

Vaz-Drago, R., Pinheiro, M.T., Martins, S., Enguita, F.J., Carmo-Fonseca, M., and Custodio, N. (2015). Transcription-coupled RNA surveillance in human genetic diseases caused by splice site mutations. Hum. Mol. Genet. *24*, 2784–2795.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. (1995). Serial Analysis of Gene Expression. Science *270*, 484–487.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The Sequence of the Human Genome. Hum. GENOME *291*, 51.

Venters, C.C., Oh, J.-M., Di, C., So, B.R., and Dreyfuss, G. (2019). U1 snRNP Telescripting: Suppression of Premature Transcription Termination in Introns as a New Layer of Gene Regulation. Cold Spring Harb. Perspect. Biol. *11*, a032235.

Vermeulen, M., Mulder, K.W., Denissov, S., Pijnappel, W.W.M.P., van Schaik, F.M.A., Varier, R.A., Baltissen, M.P.A., Stunnenberg, H.G., Mann, M., and Timmers, H.Th.M. (2007). Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4. Cell *131*, 58–69.

Villa, T. (2005). The Isy1p component of the NineTeen Complex interacts with the ATPase Prp16p to regulate the fidelity of pre-mRNA splicing. Genes Dev. *19*, 1894–1904.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Wang, Y.V., Tang, H., and Gilmour, D.S. (2005). Identification In Vivo of Different Rate-Limiting Steps Associated with Transcriptional Activators in the Presence and Absence of a GAGA Element. Mol. Cell. Biol. *25*, 3543–3552.

Wery, M., Gautier, C., Descrimes, M., Yoda, M., Vennin-Rendos, H., Migeot, V., Gautheret, D., Hermand, D., and Morillon, A. (2018). Native elongating transcript sequencing reveals global anti-correlation between sense and antisense nascent transcription in fission yeast. RNA *24*, 196–208.

Wild, T., and Cramer, P. (2012). Biogenesis of multisubunit RNA polymerases. Trends Biochem. Sci. *37*, 99–105.

Will, C.L., and Luhrmann, R. (2011). Spliceosome Structure and Function. Cold Spring Harb. Perspect. Biol. *3*, a003707–a003707.

Wilusz, J.E., Sunwoo, H., and Spector, D.L. (2009). Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. *23*, 1494–1504.

Windhager, L., Bonfert, T., Burger, K., Ruzsics, Z., Krebs, S., Kaufmann, S., Malterer, G., L'Hernault, A., Schilhabel, M., Schreiber, S., et al. (2012). Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. Genome Res. *22*, 2031–2042.

Wong, J.J.-L., Ritchie, W., Ebner, O.A., Selbach, M., Wong, J.W.H., Huang, Y., Gao, D., Pinello, N., Gonzalez, M., Baidya, K., et al. (2013). Orchestrated Intron Retention Regulates Normal Granulocyte Differentiation. Cell *154*, 583–595.

Wu, C.-S., Yu, C.-Y., Chuang, C.-Y., Hsiao, M., Kao, C.-F., Kuo, H.-C., and Chuang, T.-J. (2014). Integrative transcriptome sequencing identifies trans-splicing events with important roles in human embryonic stem cell pluripotency. Genome Res. *24*, 25–36.

Yamaguchi, Y., Shibata, H., and Handa, H. (2013). Transcription elongation factors DSIF and NELF: Promoter-proximal pausing and beyond. Biochim. Biophys. Acta BBA - Gene Regul. Mech. *1829*, 98–104.

Yearim, A., Gelfman, S., Shayevitch, R., Melcer, S., Glaich, O., Mallm, J.-P., Nissim-Rafinia, M., Cohen, A.-H.S., Rippe, K., Meshorer, E., et al. (2015). HP1 Is Involved in Regulating the Global Impact of DNA Methylation on Alternative Splicing. Cell Rep. *10*, 1122–1134.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. *11*, 377–394.

Young, R.A. (1991). RNA Polymerase II. Annu. Rev. Biochem. *60*, 689–715.

Yuan, F., Hankey, W., Wagner, E.J., Li, W., and Wang, Q. (2019). Alternative polyadenylation of mRNA and its role in cancer. Genes Dis.

Yuryev, A., Patturajan, M., Litingtung, Y., Joshi, R.V., Gentile, C., Gebara, M., and Corden, J.L. (1996). The C-terminal domain of the largest subunit of RNA polymerase II interacts with a novel set of serine/arginine-rich proteins. Proc. Natl. Acad. Sci. *93*, 6975–6980.

Zallen, J.A., and Wieschaus, E. (2004). Patterned Gene Expression Directs Bipolar Planar Polarity in Drosophila. Dev. Cell *6*, 343–355.

Zeitlinger, J., Stark, A., Kellis, M., Hong, J.-W., Nechaev, S., Adelman, K., Levine, M., and Young, R.A. (2007). RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nat. Genet. *39*, 1512–1516.

Zeng, C., and Berget, S.M. (2000). Participation of the C-Terminal Domain of RNA Polymerase II in Exon Definition during Pre-mRNA Splicing. Mol. Cell. Biol. *20*, 8290–8301.

Zenklusen, D., Larson, D.R., and Singer, R.H. (2008). Single-RNA counting reveals alternative modes of gene expression in yeast. Nat. Struct. Mol. Biol. *15*, 1263–1271.

Zeytuni, N., and Zarivach, R. (2012). Structural and Functional Discussion of the Tetra-Trico-Peptide Repeat, a Protein Interaction Module. Structure *20*, 397–405.

Zhou, M., and Law, J.A. (2015). RNA Pol IV and V in gene silencing: Rebel polymerases evolving away from Pol II's rules. Curr. Opin. Plant Biol. *27*, 154–164.

Zhou, H.-L., Hinman, M.N., Barron, V.A., Geng, C., Zhou, G., Luo, G., Siegel, R.E., and Lou, H. (2011). Hu proteins regulate alternative splicing by inducing localized histone hyperacetylation in an RNA-dependent manner. Proc. Natl. Acad. Sci. *108*, E627–E635.

Zhou, Q., Li, T., and Price, D.H. (2012). RNA Polymerase II Elongation Control. Annu. Rev. Biochem. *81*, 119–143.

Zhu, J., He, F., Hu, S., and Yu, J. (2008). On the nature of human housekeeping genes. Trends Genet. *24*, 481–484.

Zhu, J., Liu, M., Liu, X., and Dong, Z. (2018). RNA polymerase II activity revealed by GRO-seq and pNET-seq in Arabidopsis. Nat. Plants *4*, 1112–1123.

Zhu, L., Zhang, Y., Zhang, W., Yang, S., Chen, J., and Tian, D. (2009). Patterns of exon-intron architecture variation of genes in eukaryotic genomes. *12*, 1–12.

# Appendix

eLIFE elifesciences.org

RESEARCH ARTICLE

# Requirement for highly efficient pre-mRNA splicing during *Drosophila* early embryonic development

**Leonardo Gastón Guilgur[1,2,3], Pedro Prudêncio[1,2,3], Daniel Sobral[1], Denisa Liszekova[1], André Rosa[1], Rui Gonçalo Martinho[1,2,3]***

[1]Instituto Gulbenkian de Ciência, Oeiras, Portugal; [2]Departamento de Ciências Biomédicas e Medicina, Universidade do Algarve, Faro, Portugal; [3]IBB-Institute for Biotechnology and Bioengineering, Centro de Biomedicina Molecular e Estrutural, Universidade do Algarve, Faro, Portugal

**Abstract** *Drosophila* syncytial nuclear divisions limit transcription unit size of early zygotic genes. As mitosis inhibits not only transcription, but also pre-mRNA splicing, we reasoned that constraints on splicing were likely to exist in the early embryo, being splicing avoidance a possible explanation why most early zygotic genes are intronless. We isolated two mutant alleles for a subunit of the NTC/Prp19 complexes, which specifically impaired pre-mRNA splicing of early zygotic but not maternally encoded transcripts. We hypothesized that the requirements for pre-mRNA splicing efficiency were likely to vary during development. Ectopic maternal expression of an early zygotic pre-mRNA was sufficient to suppress its splicing defects in the mutant background. Furthermore, a small early zygotic transcript with multiple introns was poorly spliced in wild-type embryos. Our findings demonstrate for the first time the existence of a developmental pre-requisite for highly efficient splicing during *Drosophila* early embryonic development and suggest in highly proliferative tissues a need for coordination between cell cycle and gene architecture to ensure correct gene expression and avoid abnormally processed transcripts.
DOI: 10.7554/eLife.02181.001

***For correspondence:** rmartinho@igc.gulbenkian.pt

**Competing interests:** The authors declare that no competing interests exist.

**Funding:** See page 18

**Received:** 30 December 2013
**Accepted:** 09 March 2014
**Published:** 22 April 2014

**Reviewing editor**: Elisa Izaurralde, Max Planck Institute Development Biology, Germany

Copyright Guilgur et al. This article is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited.

## Introduction

Timing and coordination of biological processes is crucial for cellular homeostasis and normal development. *Drosophila melanogaster* embryonic development starts with thirteen nuclear divisions without cytokinesis (syncytial blastoderm), these divisions being among the fastest known for any animal embryonic system (***Foe and Alberts, 1983***). *Drosophila* syncytial blastoderm formation relies on the maternally encoded gene products loaded into the egg during oogenesis (***Tadros and Lipshitz, 2009***). After fertilization, as nuclei enter interphase 14, the maternal to zygotic transition (MZT) occurs in which the soma suddenly becomes transcriptionally active and many of the maternally encoded gene products are rapidly degraded (***Anderson and Lengyel, 1979***; ***McKnight and Miller, 1976***; ***Yasuda et al., 1991***). While the major burst of zygotic transcription occurs once the nuclei arrest in interphase 14, there is an initial wave of zygotic gene expression during the syncytial nuclear divisions 8–13 (***Pritchard and Schubiger, 1996***; ***ten Bosch et al., 2006***). Due to the extreme speed of syncytial nuclear divisions, a limitation to the size of early zygotic transcriptional units has been suggested (***McKnight and Miller, 1976***; ***Rothe et al., 1992***; ***Shermoen and O'Farrell, 1991***). Consistently, approximately 70% of early zygotic genes are small in size and intronless (***De Renzis et al., 2007***). As only 20% of *Drosophila* genes are intronless, it has been proposed that small intronless genes have an important selective advantage for transcription during the syncytial blastoderm formation (***De Renzis et al., 2007***).

**eLife digest** When a fertilized egg develops into an embryo, the expression of many genes must be carefully timed and coordinated. Researchers regularly use a type of fruit fly called *Drosophila* to study development because it is small, it has a short lifespan, and its whole genome sequence is already known. The development of a *Drosophila* embryo starts with the nucleus of the fertilized egg, which contains most of the cell's genetic material, dividing 13 times in quick succession, without the cell itself splitting. These divisions are amongst the fastest known for any animal, and given the fast developmental speed, the embryo must efficiently express all genes it needs to stay alive. Because cell division is known to inhibit gene expression this raises an interesting conundrum about the way fast cell proliferation and gene expression are coordinated.

The first step of gene expression involves a length of DNA being transcribed to produce an intermediate molecule called a messenger RNA (mRNA), which is then translated to produce a protein. However, some mRNA molecules contain regions called 'introns' that are not translated and must instead be removed via a time-consuming process called 'splicing' before the protein is produced.

At first a *Drosophila* embryo uses mRNA molecules that were spliced and packaged inside the egg by the mother, but later it starts to make its own mRNA molecules. The very first mRNA molecules made by the early embryo tend to be short and lack introns. The shortness of these molecules is thought to reflect the fact there is not enough time to produce longer mRNA molecules. Is the same 'need for speed' also responsible for the lack of introns in these molecules?

Now, Guilgur et al. have tested this hypothesis by manipulating a gene named *fandango*, which codes for part of the cellular machinery that removes introns from mRNA molecules, in fruit flies. These mutant fruit flies had less of the Fandango protein than wild-type flies and while they passed through the early stages of development normally, they later developed defects—such as abnormally shaped cells. Guilgur et al. revealed that *fandango* mutants fail to splice out the introns in the mRNA molecules that are made in the early embryo, whereas similar mRNA molecules from the mother were spliced as normal. Further experiments suggested that wild-type embryos struggled to correctly splice an untypical early gene that had multiple introns.

Together the findings of Guilgur et al. suggest that when nuclei (or cells) are dividing rapidly, there is a strong selective pressure to splice mRNA molecules quickly in the short time between the divisions. Furthermore, this pressure appears to have shaped the architecture of the earliest genes expressed in the *Drosophila* embryo, which is why the first mRNA molecules produced by the embryo itself tend not to contain introns.

DOI: 10.7554/eLife.02181.002

In yeast, *Drosophila*, and human cells, pre-mRNA splicing is mostly co-transcriptional (***Ameur et al., 2011***; ***Khodor et al., 2011***), with in vivo splicing rates being in the order of 30 s to approximately 3 min once the intron is transcribed (***Alexander et al., 2010***; ***Huranova et al., 2010***; ***Schmidt et al., 2011***). As most early zygotic transcripts are intronless (***De Renzis et al., 2007***), syncytial blastoderm interphases can be as short as 4 to 5 min, and given the fact that mitosis inhibits splicing (***Shin and Manley, 2002***), we hypothesized that further to the selective pressure for small transcriptional units, there is also a pressure to avoid pre-mRNA splicing during early zygotic expression. We isolated two mutant alleles for a subunit of the NTC/Prp19 complexes, known to be important for efficient spliceosome activation, which specifically impaired pre-mRNA splicing of early zygotic but not maternally encoded transcripts. We showed that the differential splicing defects were not related to any particular structure/sequence of the early zygotic transcripts or differential association of spliceosomal components to the NTC/Prp19 complexes. Ectopic maternal expression of an early zygotic transcript in a mutant background was sufficient to suppress its splicing defects, suggesting that they were dependent on the developmental context of gene expression. We reasoned that constraints on pre-mRNA splicing are present during *Drosophila* early embryonic development. Consistently, a small early zygotic transcript with four introns was poorly spliced in wild-type embryos. Such constraints on pre-mRNA splicing are a likely explanation why most early zygotic genes are intronless and suggest that highly proliferative tissues need coordination between cell cycle and gene architecture for correct gene expression and avoidance of abnormally processed transcripts. Our results strongly argue in

favor of a developmental pre-requisite for highly efficient splicing during fast development, therefore we propose that the requirement for overall splicing efficiency is likely to vary during development.

## Results and discussion

### *Drosophila* Fandango/Xab2 is required for blastoderm cellularization

Previously we isolated a collection of maternal mutants defective in blastoderm cellularization and/or germ-band extension (***Pimenta-Marques et al., 2008***). Complementation group 7 contained two different mutant alleles with similar defects in blastoderm cellularization. Through deficiency mapping and a candidate gene approach we concluded that both were allelic to the uncharacterized coding gene CG6197 (Flybase). To confirm the mutants' identity, we rescued their zygotic lethality, female sterility (germ-line clones), and blastoderm cellularization defects (maternal mutant embryos) using a genomic fragment construct that contained a wild-type copy of CG6197 (***Figure 1—figure supplement 1A***, data not shown). Both isolated alleles of CG6197 showed identical phenotypes: maternal mutant embryos (hereafter referred to as mutant embryos) showed normal syncytial nuclear divisions (***Figure 1A,B***) but subsequently failed to elongate the cortical nuclei, which became mislocalized during blastoderm cellularization (***Figure 1C–F***, quantification in ***Figure 1G***). The blastoderm cellularization phenotype was remarkably similar to that described for *kugelkern*/*charleston* mutant embryos (***Brandt et al., 2006***; ***Pilot et al., 2006***). Based on the observed phenotypes, we named the corresponding gene *fandango*, after the Iberian folk dance.

*fandango* encodes the *Drosophila* ortholog of yeast SYF1 (synthetic lethal with cdc41) (***Russell et al., 2000***) and human XAB2 (XPA binding protein 2) (***Nakatsu et al., 2000***; ***Kuraoka et al., 2008***). These proteins were described as subunits of the NTC/Prp19 complexes, which are important for spliceosome stabilization and activation (***Chan et al., 2003***; ***Chang et al., 2009***; ***Hogg et al., 2010***). Fandango protein has multiple tetratricopeptide repeat (TPR) motifs, which is a protein–protein interaction module (***Zeytuni and Zarivach, 2012***). Sequencing both alleles of *fandango* (*fand*$^{1}$ and *fand*$^{2}$) revealed distinct mutations within the *fandango* open reading frame (ORF). *fand*$^{1}$ contained a missense point mutation in a highly conserved residue within TPR domain VII (from an alanine to a valine; A401V), whereas *fand*$^{2}$ contained a microdeletion of 18 nucleotides within TPR domain VI, which deleted six conserved amino acids from position 355 to 360 (***Figure 1—figure supplement 1B***). In total protein extracts, both *fand*$^{1}$ and *fand*$^{2}$ mutant embryos showed a significant reduction in Fandango protein levels compared to control (***Figure 1I***). fandango mRNA levels, analyzed by real-time qPCR, were similar between control and *fand*$^{1}$ mutant embryos (***Figure 1J***), suggesting that the mutation did not alter the stability of the encoding pre-mRNA.

### *Drosophila* Fandango/Xab2 is differentially required for splicing of maternal and early zygotic pre-mRNAs

As noted above *fandango* maternal mutant embryos and *kugelkern* (*kuk*) mutant embryos showed remarkably similar blastoderm cellularization defects (***Brandt et al., 2006***; ***Pilot et al., 2006***). Since *fandango* encodes a protein whose yeast and human orthologs are required for efficient spliceosome activity, we hypothesized that Fandango was required for splicing of *kuk* transcripts. *kuk* encodes two different transcripts, which vary in intron size (***Figure 2A***). Both transcripts are predicted to encode the same protein. Analysis of publicly available modENCODE transcriptome datasets (***Graveley et al., 2011***) suggested that the large *kuk* transcript was maternally expressed, whereas the small *kuk* transcript was only expressed zygotically. Through RT-PCR analysis we confirmed that similarly to control maternal genes (*nanos* and *oskar*) the large *kuk* transcript was maternally expressed (being present in unfertilized eggs), whereas the small *kuk* transcript was exclusively zygotically expressed (being present only in fertilized eggs) as the case of well-known early zygotic genes (*even-skipped* and *krüppel*) (***Figure 2—figure supplement 1A***).

To investigate by RT-PCR whether Fandango was required for splicing of *kuk* pre-mRNAs, specific sets of primers (exon–exon, e–e; intron–exon, i–e) were designed for each *kuk* transcript, taking advantage of a longer 3′UTR in the small *kuk* transcript (***Figure 2A***). Surprisingly, whereas *fandango* embryos showed significant splicing defects of the small zygotic *kuk* transcript, the large maternal *kuk* transcript was correctly spliced (***Figure 2B***; ***Figure 2—figure supplement 1B***). Splicing defects were fully rescued by a genomic fragment construct that contained a wild type copy of *fandango* (***Figure 2—figure supplement 1C***). The differential requirement of Fandango for splicing of *kuk* transcripts prompted us

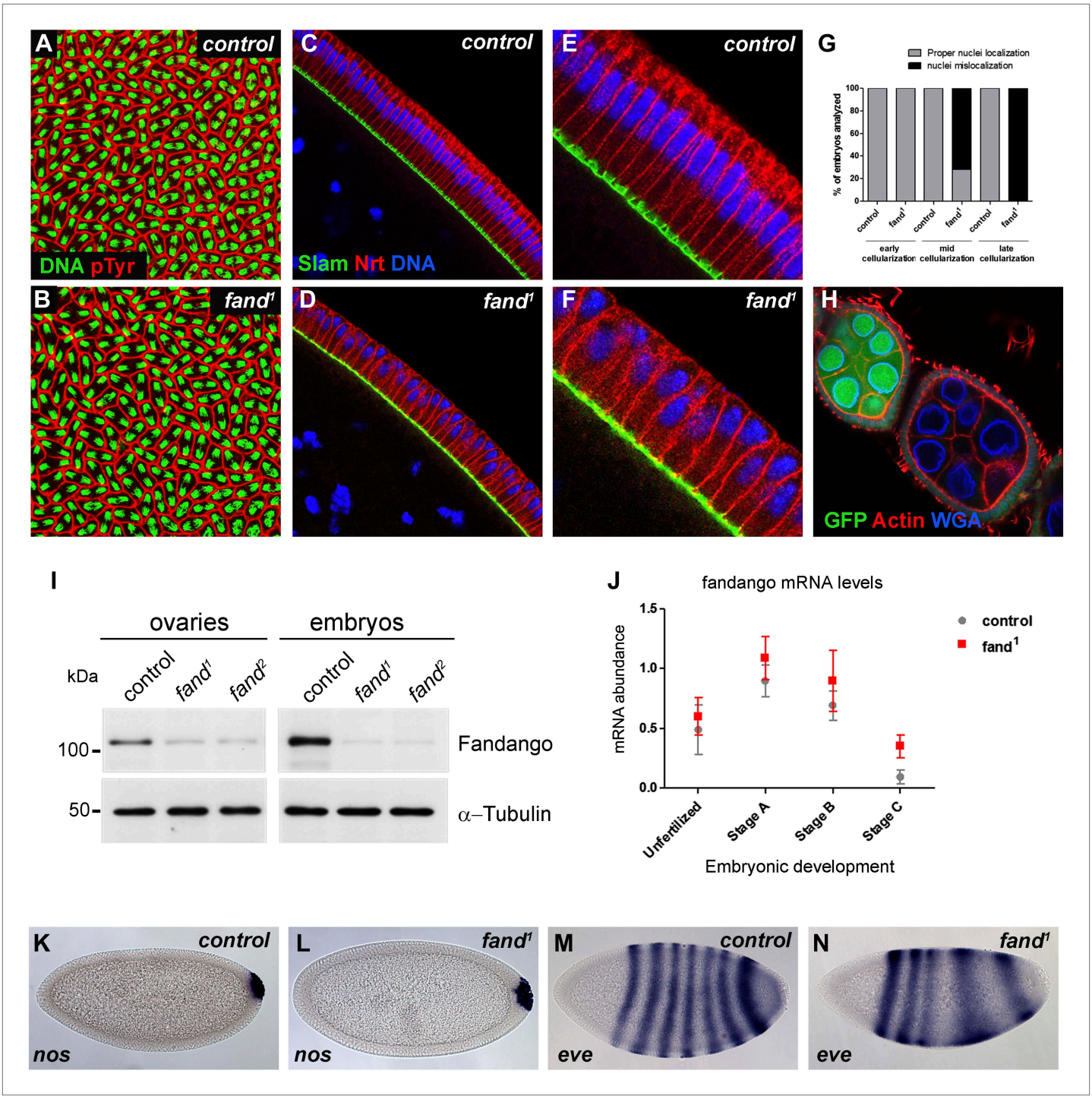



**Figure 1**. Drosophila Fandango/Xab2 is required for blastoderm cellularization. (**A** and **B**) Panels show embryos with normal syncytial blastoderm nuclear divisions in control embryos (*hs*-FLP; FRT42B) (**A**) and *fand*$^{1}$ germ-line clone embryos (*hs*-FLP; FRT42B *fand*$^{1}$, maternal mutant) (**B**). Embryos were stained for DNA (green) and p-Tyrosine (red). (**C**–**F**) Panels show blastoderm cellularized embryos. Control embryos showed normal epithelial architecture with elongated nuclei and columnar cell shape (**C**). *fand*$^{1}$ germ-line clone mutant embryos showed abnormal epithelial architecture, the cortical nuclei failed to elongate and became mislocalized (**D**). (**E** and **F**) Magnification of **C** and **D,** respectively. Embryos were stained for Slam (green), Neurotactin (red), and DNA (blue). (**G**) Quantification of *fandango* maternal mutant embryo phenotype during blastoderm cellularization. Early cellularization: control: 100% normal (n = 44), *fand*$^{1}$: 100% normal (n = 49); mid cellularization: control: 100% normal (n = 25), *fand*$^{1}$: 28% normal (n = 21); late cellularization: control: 100% normal (n = 42), *fand*$^{1}$: 0% normal (n = 38). (**H**) Maternally controlled oogenesis was normal in *fandango* mutant clones. Absence of endogenous nGFP (green) indicated that the cells were homozygous for *fand*$^{1}$ mutation. Ovaries were stained for F-actin (red) and WGA (blue). (**I**) Western blot of

*Figure 1. Continued on next page*

*Figure 1. Continued*

whole protein extracts from embryos and ovaries mutant for *fand*$^{1}$ and *fand*$^{2}$ alleles (germ-line clones) showed a clear reduction in Fandango protein levels compared to control tissues. It should be noticed that due to experimental constraints the total protein extracts from mutant ovaries included not only signal from mutant germ-line cells (homozygous for *fand*$^{1}$), but also the tightly associated heterozygote somatic follicle cells. α-Tubulin was used as a loading control. (**J**) Real-time qPCR analysis showed no significant differences in fandango mRNA levels between control and *fand*$^{1}$ embryos during development (Two-way ANOVA p>0.05 ns.). fandango mRNA levels were normalized with β-actin mRNA levels. (**K**–**N**) in situ hybridization for *nanos* RNA (maternal) and *even-skipped* RNA (early zygotic) in blastoderm cellularized embryos. Both control (**K**) and *fand*$^{1}$ mutant (**L**) embryos showed normal *nos* localization pattern in the pole cells. *fand*$^{1}$ embryos (**N**) showed A–P patterning defects of *eve* compared to control embryos (**M**).

DOI: 10.7554/eLife.02181.003

The following figure supplements are available for figure 1:

**Figure supplement 1**. *fandango* mutant alleles contain changes in highly conserved amino acids.

DOI: 10.7554/eLife.02181.004

to investigate more than 20 other maternal and early zygotic genes. RT-PCR analysis of *fandango* embryos invariably showed splicing defects of early zygotic but not maternally encoded transcripts (***Figure 2C***, data not shown). High-throughput transcriptome sequencing (RNAseq) confirmed that splicing of early zygotic but not maternally encoded gene products was affected in *fandango* embryos (***Figure 2D***, ***Figure 2—figure supplement 2A***). Maternal transcripts, whose intron size was equivalent to those observed in early zygotic transcripts, were unaffected (***Figure 2—figure supplement 2B***), which showed that Fandango was not specifically rate limiting for splicing of small introns. Comparison analysis of 5′ and 3′ splice site consensus sequences between maternal and zygotic pre-mRNA transcripts showed no significant differences (***Figure 2—figure supplement 2C***) and the two populations of transcripts displayed a similarly heterogeneous exon–intron structure (***Figure 2—figure supplement 2D***). RT-PCR analysis of maternally encoded transcripts from wild-type and *fandango* mutant ovaries (germ-line clones) also failed to detect splicing defects (***Figure 2—figure supplement 1D***). This suggested that the absence of splicing defects of maternally encoded transcripts in *fandango* embryos was not due to specific degradation of unspliced transcripts during oogenesis.

The differential requirement of Fandango for splicing of early zygotic encoded transcripts is fully consistent with the observation that maternally controlled oogenesis, primordial germ-cell formation, and syncytial nuclear divisions were normal in *fandango* mutants (***Figure 1A,B,H,K,L***), whereas the first detectable phenotype only occurred during zygotically controlled blastoderm cellularization (***Figure 1C–F***). Despite the fact that clonal analysis of the female germ line for both alleles of *fandango* showed normal oogenesis and egg laying (***Figure 1H***) (data not shown), Fandango protein levels were significantly reduced in the mutant ovaries (germ-line clones) (***Figure 1I***). *fandango* embryos also failed to initiate germ-band extension after blastoderm cellularization (data not shown). It was previously shown that anterior–posterior (A–P) patterning is required for germ-band extension (***Zallen and Wieschaus, 2004***). Consistently, *fandango* embryos showed A–P patterning defects in the early zygotic pair-rule gene *even-skipped* (***Figure 1M,N***).

## Fandango is similarly associated with the NTC/Prp19 complexes during oogenesis and early embryonic development

The highly conserved NTC/Prp19 and NTC/Prp19-related complexes are essential for pre-mRNA splicing as they facilitate the formation and progression between distinct spliceosome conformations during the splicing reaction (***Chan et al., 2003***; ***Hogg et al., 2010***).

Endogenous Fandango and Prp19 physically interacted in the early embryo (***Figure 3A***). Moreover, both endogenous Fandango and Prp19 physically interacted with endogenous ISY1 and CDC5L (***Figure 3A***), confirming that Fandango is a *bona fide* subunit of *Drosophila* NTC/Prp19 complexes. Immunoprecipitation of Myc-tagged Fandango and Myc-tagged Prp19 from embryonic protein extracts also identified an identical group of interacting proteins (***Table 1***; ***Supplementary file 1***). Whereas Myc-Fandango mostly interacted with the NTC/Prp19-related complex subunits, Myc-Prp19 interacted principally with the NTC/Prp19 complex subunits. This illustrated that, as in humans, distinct but interacting NTC/Prp19 complexes exist in *Drosophila*, in agreement with the recent suggestion that a remarkable degree of conservation of distinct splicing complexes exists among metazoans (***Herold et al., 2009***).

The differential requirements of Fandango for pre-mRNA splicing of maternal and early zygotic transcripts potentially suggest distinct interactions between Fandango and other splicing proteins

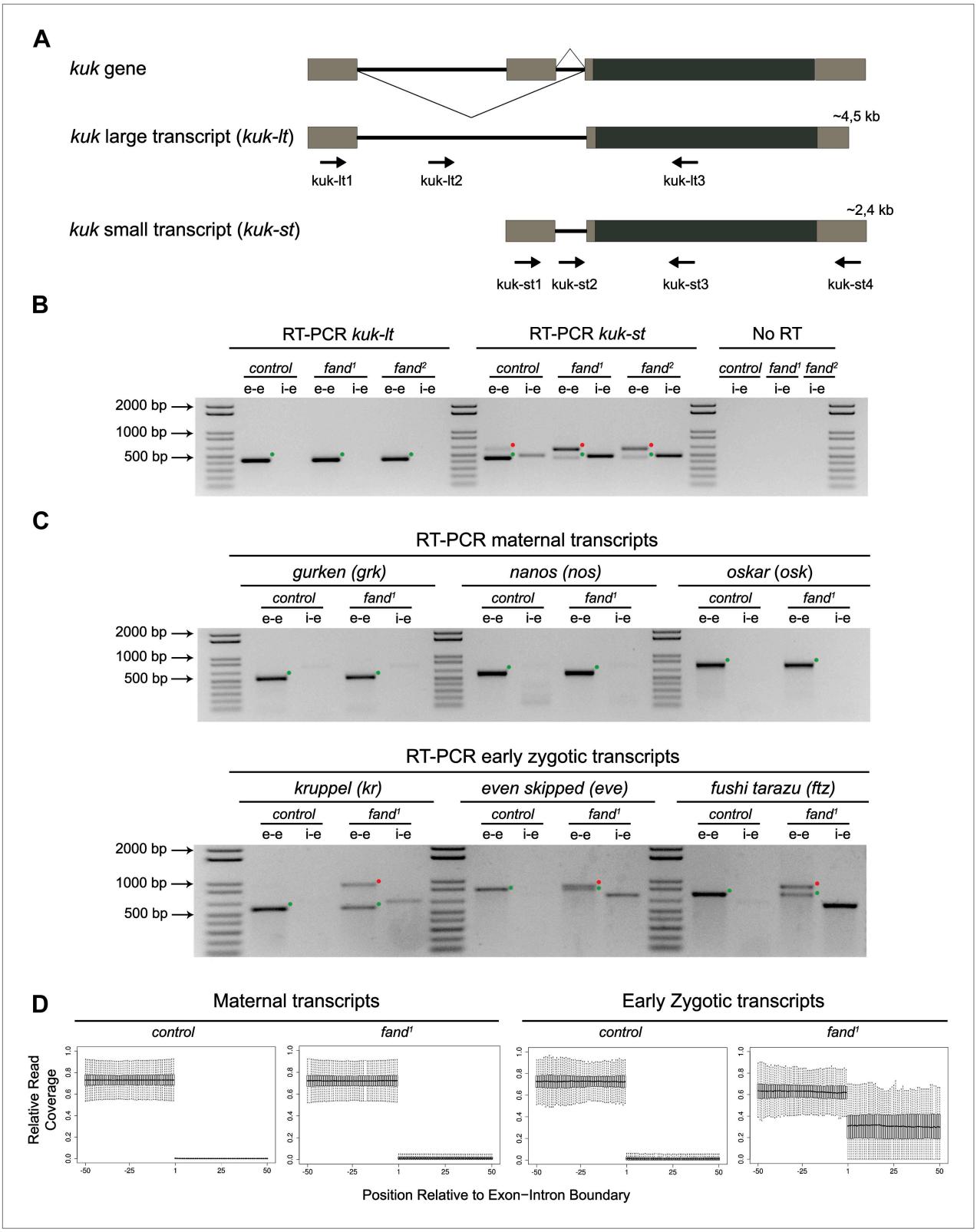



**Figure 2**. Splicing of early zygotic but not maternally encoded pre-mRNAs is affected in *fandango* mutants. (**A**) The *kugelkern* (*kuk*) locus encodes two transcripts of different size, *kuk-lt* containing a large intron and *kuk-st* with a short intron. Orientation and position of primers used for splicing analysis is indicated (arrows). (**B**) RT-PCR analysis of *kuk* transcripts. Control embryos yielded PCR products in the size predicted for the properly spliced forms of both *kuk* transcripts using exon–exon (e–e) primers (green dots, *kuk-lt*: 431 bp and *kuk-st*: 437 bp). *fandango* maternal mutant embryos

*Figure 2. Continued on next page*

*Figure 2. Continued*

(*fand*$^{1}$ and *fand*$^{2}$ alleles) showed splicing defects only in the *kuk-st* transcript; PCR products were detected by e–e primers in the size expected for intron retention (red dots, *kuk-st*: 596 bp) and by intron–exon (i–e) primers (*kuk-st*: 474 bp). Splicing of the *kuk-lt* was not affected in *fandango* mutant background; PCR products were only detected with e–e primers in the predicted size for the correctly spliced pre-mRNA (green dots, *kuk-lt*: 431 bp). 'No RT' controls (only total RNA as template) yielded no amplification, meaning there was no contamination with genomic DNA in the samples tested. (**C**) RT-PCR analysis of maternal and early zygotic genes. Maternal transcripts were properly spliced, in both, control and *fand*$^{1}$ mutant embryos; PCR products were only detected using e–e primers (green dots, *grk*: 527, *nos*: 581, *osk*: 762 bp). In contrast, early zygotic transcripts were correctly spliced only in control embryos (green dots, *kr*: 559, *eve*: 828, *ftz*: 753 bp). *fand*$^{1}$ mutant embryos yielded PCR products in the size predicted for intron retention with e–e primers (red dots, *kr*: 932, *eve*: 899, *ftz*: 900 bp) and with i–e primers (*kr*: 629, *eve*: 720, *ftz*: 595 bp). All PCR bands showed in the panels were cloned and sequenced to confirm their identity. Green dots indicate correctly spliced transcripts, red dots indicate unspliced transcripts (intron retention). (**D**) RNA-Seq data confirmed that zygotic but not maternally encoded transcripts displayed a large fraction of splicing defects (intron retention) in *fand*$^{1}$ mutant embryos. The panel shows box plot of the distribution of numbers of reads per bp relative to the total number of reads falling inside a 100 bp window centered around the 5'splice sites of zygotic (n = 408 splice sites from 270 genes) or maternal genes (n = 5876 splice sites from 2048 genes).

DOI: 10.7554/eLife.02181.005

The following figure supplements are available for figure 2:

**Figure supplement 1**. Splicing of early zygotic but not maternally encoded pre-mRNAs is affected in *fandango* mutants.

DOI: 10.7554/eLife.02181.006

**Figure supplement 2**. Early zygotic but not maternally encoded pre-mRNAs shows significant intron retention in *fandango* mutants.

DOI: 10.7554/eLife.02181.007

during oogenesis and early embryonic development. Nevertheless, immunoprecipitation of Myc-Fandango specifically expressed in the female germ line during oogenesis and in the early embryo identified a virtually identical group of interacting proteins: mostly subunits of the NTC/Prp19-related complex, and to a lesser extent, subunits of the NTC/Prp19 complex (***Table 1***; ***Supplementary file 1***). These results showed that Fandango physically interacts with a similar group of splicing proteins during oogenesis and in the early embryo.

To better understand the splicing defects observed in *fandango* embryos, we investigated if the integrity of NTC/Prp19 complexes was affected in this mutant. Size-exclusion chromatography showed detectable changes in the integrity of NTC/Prp19 complexes in *fandango* embryos (***Figure 3B***), with a significant reduction in the levels of ISY1 protein (***Figure 3C***). ISY1 is a NTC/Prp19-related complex subunit (***Figure 3A***). The loss of integrity of the ISY1-positive ~600–800 kDa NTC/Prp19 complex (***Figure 3B***) and concomitant reduction in the stability of some of their subunits, most likely impaired efficient activation of the spliceosome (***Villa and Guthrie, 2005***) and were likely explanations for the splicing defects observed in *fandango* embryos. In agreement with the suboptimal spliceosome activation hypothesis, intron retention was the main splicing defect of early zygotic transcripts in *fandango* embryos (***Figure 2B,C***, ***Figure 2—figure supplement 2B***; data not shown).

Levels of ISY1 were similarly affected in *fandango* mutants during oogenesis and in the early embryo (***Figure 3C***), suggesting this decrease did not explain the differential requirements of Fandango for splicing of early zygotic and maternally encoded transcripts. Mutant clonal analysis of a stronger allele of *fandango* (nonsense mutation), showed a complete loss of the female germ line in adult ovaries (data not shown). This demonstrated that the two isolated alleles of *fandango* are hypomorphic and suggested that Fandango was required, albeit at lower levels, for splicing of maternal transcripts. We concluded it is unlikely that a differential expression and/or association of core components of the spliceosome could potentially explain the differential requirements for Fandango between oogenesis and the early embryo. The most likely explanation is that Fandango is quantitatively (but not qualitatively) differentially required during early embryonic development.

## Reduction in Fandango levels affects mainly its splicing function

Transcriptional elongation can affect co-transcriptional splicing (***de la Mata et al., 2003***; ***Ip et al., 2011***). It was recently shown that Syf1, the yeast ortholog of Fandango, is also important for RNApol II transcriptional activity (***Chanarat et al., 2011***; ***David et al., 2011***), therefore we decided to investigate transcription in *fandango* embryos. Three intronless early zygotic genes (*nullo*, *snail*, and *scute*) and two early zygotic genes with introns (*even-skipped* and *tailless*) were selected for further analysis by real-time qPCR. During mid/late-syncytial blastoderm (stage B) (***Figure 4A***, 'Materials and methods'), no significant differences in transcript abundance were observed between control and

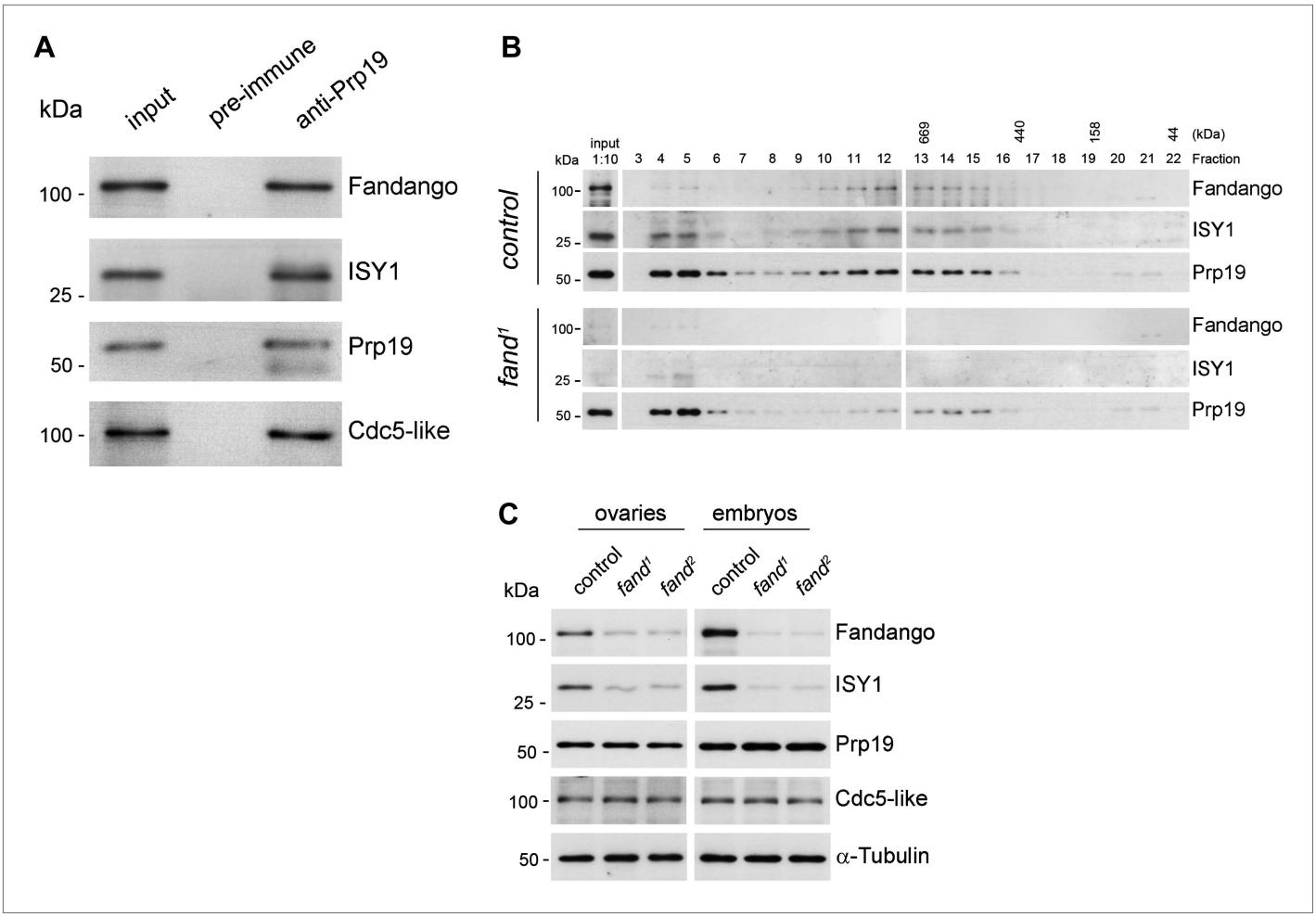



**Figure 3**. Fandango physically interacts with a similar group of splicing proteins during oogenesis and embryogenesis. (**A**) Pull down assay from nuclear-enriched protein extracts using a polyclonal antibody of Prp19. Endogenous Prp19 interacts physically with Fandango and other subunits of the NTC/Prp19 complexes (ISY1 and CDC5L). Pre-immune serum was used in the control. (**B**) Size-exclusion chromatography of control and *fand*$^{1}$ mutant protein extracts from 0–3 hr embryo collections using a Superose 6 10/300 column. After separation, each fraction was analyzed by Western blot. NTC/Prp19 complexes subunits (Prp19, Fandango, and ISY1) were part of a ~600–800 kDa complex and also co-purified in a significantly larger complex (fraction 4 and 5). *fand*$^{1}$ mutant protein extracts showed a significant reduction in levels of Fandango and ISY1 subunits and a size reduction of the Prp19-positive ~600–800 kDa complex. (**C**) Western-blot analysis of total protein extracts from ovaries (left) and 0–3 hr embryos (right) from control and both *fandango* alleles, showed a reduction of Fandango and ISY1 protein levels in both tissues. Protein levels of Prp19 and CDC5L were not affected. α-Tubulin was used as loading control. Fandango Western blot is the same as shown in ***Figure 1I***.

DOI: 10.7554/eLife.02181.008

*fandango* (***Figure 4—figure supplement 1A***), whereas embryos mutant for *grapes* showed the expected reduction of transcript levels (***Figure 4—figure supplement 1A***; ***Sibon et al., 1997***).

During transcriptional elongation, RNApol II is specifically phosphorylated on the Ser2 residue of its carboxy-terminal domain (CTD) (***Hsin and Manley, 2012***). In agreement with the onset of early zygotic transcription, we observed a significant increase in RNApol II CTD Ser2 phosphorylation as the embryo developed from early/mid-syncytial blastoderm (stage A), into mid/late-syncytial blastoderm (stage B), and blastoderm cellularization (interphase 14) (stage C) (***Figure 4A,B***). Both control and *fandango* embryos showed a similar increase in global levels of RNApol II CTD Ser2 phosphorylation (***Figure 4B,C***). As transcriptional regulation during interphase 14 (stage C) relies on correct expression of early zygotic genes and degradation of many maternal RNAs (MZT) (***Tadros and Lipshitz, 2009***), we concluded that transcriptional changes at this stage (***Figure 4—figure supplement 1A***) were most likely a consequence of the widespread defects occurring during mid/late-syncytial blastoderm. Altogether, we concluded that the observed reduction in Fandango levels affects mainly its splicing function.

**Table 1.** LC-MS analysis of co-immunoprecipitation assays from ovaries and embryos

| Drosophila | | | Fandango-myc | | | | Prp19-myc | |
|---|---|---|---|---|---|---|---|---|
| CG | gene | | ovaries | | embryos | | embryos | |
| | | *Human/yeast* | rep1 | rep2 | rep1 | rep2 | rep1 | rep2 |
| prp19 complex | | | | | | | | |
| CG5519 | prp19 | PRP19/Prp19 | + | + | + | + | +++ | ++ |
| CG6905 | cdc5-like | CDC5L/Cef1 | + | + | + | + | +++ | ++ |
| CG1796 | Tango4 | PLRG1/Prp46 | + | + | + | + | + | + |
| CG4980 | - | BCAS2/Snt309 | – | + | - | - | + | + |
| CG12135 | c12.1 | CWC15/cwc15 | + | – | – | - | – | – |
| Prp19 related | | | | | | | | |
| CG6197 | Fandango | Xab2/Syf1 | +++ | +++ | +++ | +++ | + | + |
| CG31368 | – | AQR/– | +++ | +++ | +++ | +++ | + | + |
| CG4886 | cyp33 | PPIE/– | ++ | ++ | + | ++ | + | + |
| CG9667 | – | ISY1/ISY1 | + | + | + | + | + | + |
| CG8264 | Bx42 | SNW1/Prp45 | + | + | + | + | + | + |
| CG14641 | – | RBM22/Cwc2 | – | + | + | + | + | – |
| CG3193 | Crn | CRNKL1/Clf1 | – | + | – | – | + | – |
| CG13892 | cypl | PPIL1/- | – | – | – | + | – | – |
| CG1639 | l(1)10Bb | BUD31/Bud31 | – | – | – | – | + | – |

Co-immunoprecipitations were performed using total protein extracts from the different tissues expressing Myc-tagged Fandango or Myc-tagged-Prp19. Human and yeast homologues and the different sub-complexes are shown as described in (***Herold et al., 2009***). (–), (+), (++), (+++) correspond to 0, 1–9, 10–19, and >20 non-repeated peptides respectively. None of the proteins shown were detected in the negative controls (for detailed LC-MS analysis see ***Supplementary file 1***).

DOI: 10.7554/eLife.02181.009

## Ectopic maternal expression of an early zygotic transcript in the mutant background was sufficient to suppress its splicing defects

To investigate if the differential requirement of Fandango for splicing of early zygotic and maternally encoded transcripts potentially resulted from distinct transcript sequences, we generated an early zygotic *kuk* transcript (*kuk-LacZ*) under the control of an UAS/Gal4 inducible promoter, where the open reading frame (ORF) was replaced by LacZ (***Figure 5A***, see 'Materials and methods' for more details). As expected, when this construct was expressed zygotically, it was correctly spliced in control but not in *fandango* embryos (***Figure 5B***). In contrast, splicing of the *kuk-LacZ* construct occurred normally in both control and *fandango* mutants when it was expressed maternally (***Figure 5B***). Since maternal expression of an early zygotic transcript, in a *fandango* mutant background, was enough to suppress its splicing defects, we concluded that the differential requirement of Fandango for splicing of early zygotic transcripts was most likely due to the developmental context of gene expression and not a particularity in the early zygotic pre-mRNA sequences. Consistently, we failed to detect differences related to intron size, splice sites consensus, and exon–intron structure between maternal and zygotic transcripts (***Figure 2—figure supplement 2B–D***).

## A small early zygotic transcript with multiple introns was poorly spliced in wild-type embryos

*fandango* mutants showed a significant reduction in Fandango and ISY1 protein levels (***Figure 3C***), which most likely impaired efficient activation of the spliceosome (***Villa and Guthrie, 2005***). Since mitosis inhibits splicing (***Shin and Manley, 2002***), pre-mRNA splicing of early zygotic transcripts needs to be highly efficient for these genes to be correctly expressed. This suggests the existence of a developmental pre-requisite for highly efficient splicing, so that a suboptimal activation of the spliceosome would specifically impair pre-mRNA splicing of early zygotic but not maternal transcripts. Wild-type embryos already showed a detectable amount of intron retention in early zygotic transcripts

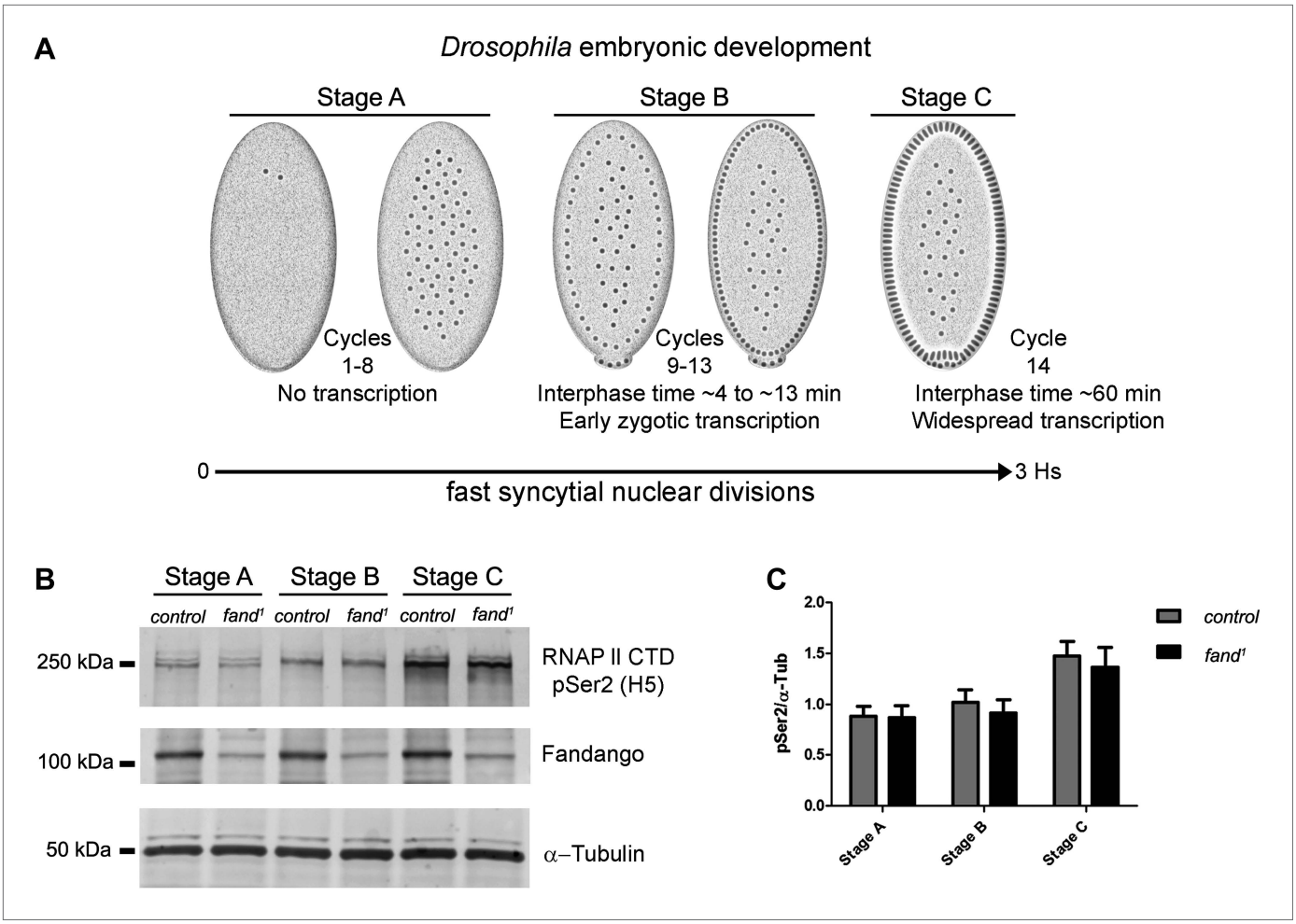



**Figure 4**. Early zygotic transcription is not affected during mid/late-syncytial blastoderm in *fandango* mutants. (**A**) Embryos were divided into three different groups according to developmental stage ('Materials and methods'), stage A: early/mid-syncytial blastoderm embryos, stage B: mid/late-syncytial blastoderm embryos, and stage C: blastoderm cellularization embryos. (**B**) Western blot for RNApol II CTD Ser2 phosphorylation levels. Control and *fand*$^{1}$ embryos showed a similar increase in the global levels of RNApol II CTD Ser2 phosphorylation over the course of early embryonic development. α-Tubulin was used as a loading control. (**C**) Quantification of the CTD Ser2 phosphorylation from five independent western blot assays showed no significant difference at any of the embryonic developmental stages analyzed (Two-way ANOVA p>0.05 ns.).

DOI: 10.7554/eLife.02181.010

The following figure supplements are available for figure 4:

**Figure supplement 1**. Early zygotic transcription is not affected during mid/late-syncytial blastoderm in *fandango* mutants.

DOI: 10.7554/eLife.02181.011

(***Figure 2B,D***, ***Figure 2—figure supplement 2B***), which was dramatically exacerbated in *fandango* embryos (***Figure 2B,D***, ***Figure 2—figure supplement 2B***).

We hypothesized that regardless of transcript size, there was also a constraint on pre-mRNA splicing of early zygotic transcripts in wild-type embryos. We generated a gene where the 5′UTR sequence including the intron of the small zygotic *kuk* transcript was quadruplicate to test this hypothesis (***Figure 6A,D*** see 'Materials and methods' for more details). Quadruplicate introns were linked by in-frame LacZ coding sequences, and the entire construct (*4x intron kuk-LacZ*) was under the control either of an endogenous early zygotic minimal promoter (*nullo-4x intron kuk-LacZ*, ~2.5 Kb) (***Figure 6A***) or an inducible UAS/Gal4 promoter (*UAS-4x intron kuk-LacZ*, ~2.5 Kb) (***Figure 6D***). The total size of the encoded pre-mRNAs was comparable to many other endogenous early zygotic genes (e.g., *kugelkern*, *runt*, *krupple*). As a control, we introduced point mutations in the splice sites of these constructs to generate comparable intronless transcripts (*no intron kuk-LacZ*) (***Figure 6A,D***).

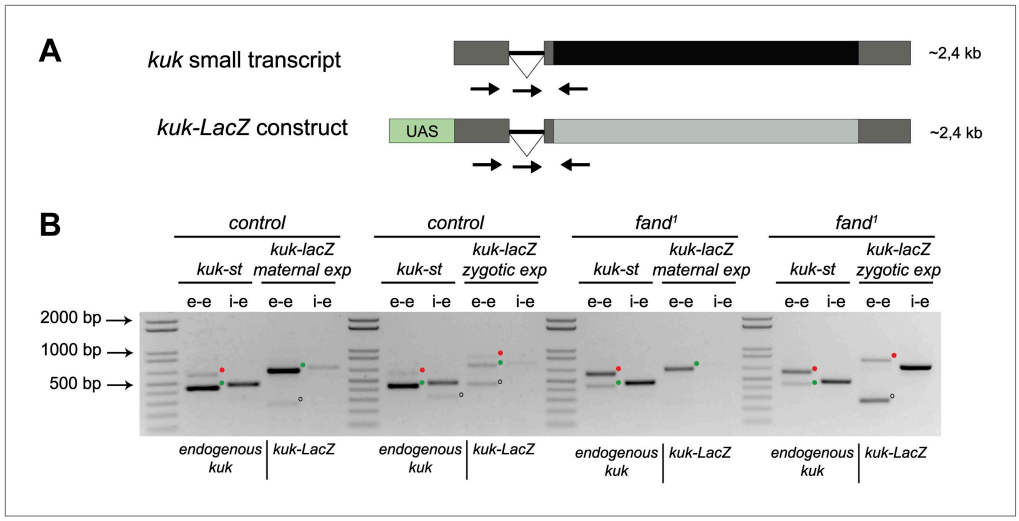



**Figure 5**. Ectopic maternal expression of an early zygotic transcript in the mutant background is sufficient to suppress its splicing defects. (**A**) The *kuk-LacZ* construct was built using the 5′UTR, the intron and the 3′UTR of the *kuk* small transcript (dark gray), and replacing the *kuk* ORF (black) by the LacZ coding sequence (light gray). To induce the expression of this construct it was put under the control of the UAS promoter (green) to drive the tissue specific expression with GAL4 drivers. Orientation and position of primers used for splicing analysis is indicated (arrows). (**B**) RT-PCR analysis of the *kuk-LacZ* construct. When it was zygotically expressed, it was correctly spliced in control but not in *fand*[1] embryos (similarly to the endogenous small *kuk* transcript). Intron retention with e–e primers (red dots, *kuk-st*: 596 bp and *kuk-LacZ*: 869 bp) and a PCR product with i–e primers (751 bp) were observed in the mutant. When it was maternally expressed, *kuk-LacZ* construct was correctly spliced both in control and *fand*[1] embryos, being detected just the spliced form of the construct (green dots, *kuk-st*: 437 bp and *kuk-LacZ*: 713 bp). In contrast, the endogenous zygotically expressed small *kuk* transcript (*kuk-st*) is still poorly spliced in *fand*[1] embryos carrying the *kuk-LacZ* construct. Open circles indicate unspecific PCR products (confirmed by sequencing). Green dots indicate correctly spliced transcripts, whereas red dots indicate unspliced transcripts (intron retention).
DOI: 10.7554/eLife.02181.012

Only the first intron (int1) of the *4x intron kuk-LacZ* construct was correctly spliced when it was zygotically expressed in wild-type embryos under the control of an endogenous early zygotic minimal promoter (***Figure 6B***). Likewise, when the *4x intron kuk-LacZ* construct was early zygotically expressed under the control of the inducible promoter UAS/Gal4 there were similar splicing defects (intron retention) (***Figure 6E***).

Measurement of in vivo kinetics of mRNA splicing showed that half-lives for splicing reactions are <1 min for the first intron, but 2–8 min for both second and third introns (***Audibert et al., 2002***). Hence, splicing of two or more introns requires more time than transcription and becomes rate limiting. Consistent with the hypothesis of a temporal constraint on pre-mRNA splicing, when the *4x intron kuk-LacZ* construct was zygotically expressed, the splicing defects of the firstly transcribed 5′-localized introns (Int1 and Int2) were significantly weaker than those observed in the later transcribed 3′-localized introns (Int3 and Int4) (***Figure 6E***). Importantly, maternal expression of this construct was sufficient to significantly suppress its splicing defects (***Figure 6E***). Real-time qPCR analysis showed that these constructs were equivalently zygotically and maternally expressed (***Figure 6C,F***). This suggested that splicing did not quantitatively impair early zygotic transcription, which was consistent with the observation that the rates of transcriptional elongation proceed independently of splicing (***Brody et al., 2011***).

We showed that in wild-type embryos pre-mRNA splicing imposed significant constraints on early zygotic expression, which is a likely explanation why most early zygotic genes are intronless (***De Renzis et al., 2007***). Although a moderate decrease in the length of syncytial blastoderm interphases (seen in *grapes* mutant embryos [***Sibon et al., 1997***]) was not sufficient to induce splicing defects in otherwise wild-type embryos (data not shown), we hypothesize that avoidance of pre-mRNA splicing during early zygotic expression is a consequence of the extremely short interphases and frequent mitotic cycles. Similarly to *Drosophila*, mosquito *Aedes aegypti* and the zebrafish *Danio rerio* early zygotic transcripts

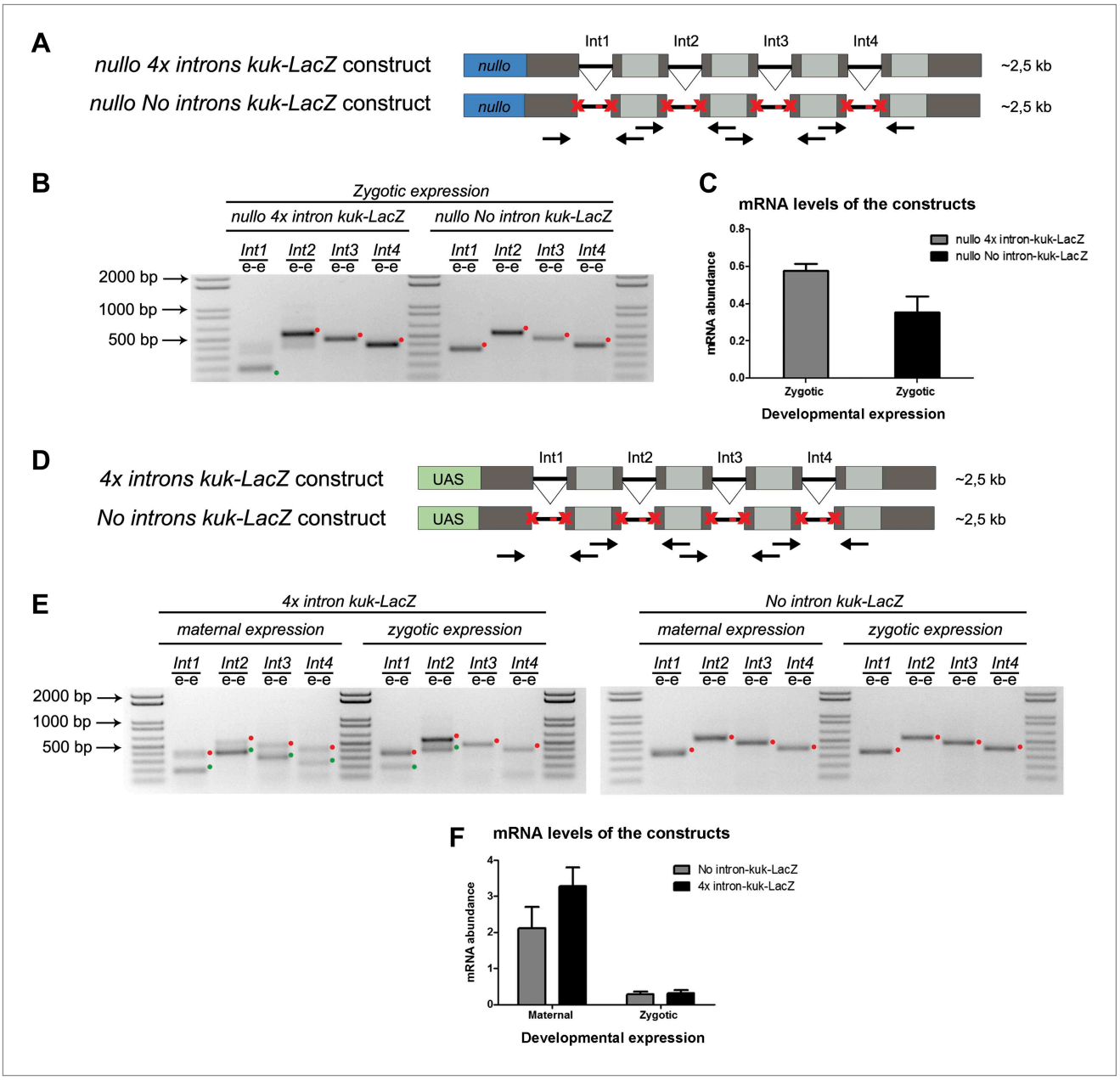



**Figure 6**. A small early zygotic transcript containing four introns is poorly spliced in wild-type embryos. (**A** and **D**) The *4x intron kuk-LacZ* construct was a variant of the *kuk-LacZ* that contains four copies of *kuk* small transcript intron (dark gray). Each intron is separated by 201 nucleotides of an in frame Lac-Z sequence (light gray). The *no intron kuk-LacZ* construct has all splice sites present in the *4x intron kuk-LacZ* construct mutated to thymidines. The constructs were fused to a *nullo* minimal promoter (blue) (**A**), or fused to an inducible UAS promoter (green) (**D**). Orientation and position of primers used for splicing analysis is indicated (arrows). (**B**) RT-PCR analysis showed significant splicing defects (intron retention) of the *4x intron kuk-LacZ* construct when expressed under the control of an endogenous early zygotic promoter (*nullo* promoter). The first intron was correctly spliced, being detected mainly the PCR product corresponding to the spliced form (green dot). The remaining introns (second, third, and fourth) were completely unspliced (red dots, intron retention). In the intronless (*no intron kuk-LacZ*) construct, under the control of the same *nullo* promoter were only observed PCR bands whose sizes correspond to unspliced forms (red dots, intron retention). (**C**) Real-time qPCR analysis showed that the *4x intron kuk-LacZ* and *no intron kuk-LacZ* constructs were expressed to the same extent when under the control of the *nullo* minimal promoter (*t* test p>0.05 ns.). (**E**) RT-PCR analysis of the *4x intron kuk-LacZ* construct showed significant splicing defects (intron retention) when zygotically expressed in wild-type embryos under the control of an inducible UAS promoter. Although the most 5′-localized introns (first and second) were still partially spliced, being observed two PCR bands corresponding to the spliced (green dots, int1: 191 and int2: 385 bp), and unspliced forms (red dots, int1: 347 and int2: 541 bp). The furthest 3′-localized introns (third and fourth) were completely unspliced, being only observed one PCR band with the size corresponding to intron retention (red dots, int3: 463 and int4: 385 bp). Maternal expression of the *4x intron kuk-LacZ* construct was sufficient to significantly suppress splicing defects in

*Figure 6. Continued on next page*

*Figure 6. Continued*

the four introns analyzed (green dots, spliced forms: int1: 191, int2: 385, int3: 307, int4: 229 bp; red dots, unspliced forms: int1: 347, int2: 541, int3: 463, int4: 385 bp). Zygotic and maternal expression of the *no intron kuk-LacZ* construct only showed PCR bands with sizes corresponding to unspliced forms (red dots, intron retention). (**F**) Real-time qPCR analysis showed that the *4x intron kuk-LacZ* and *no intron kuk-LacZ* constructs were expressed to the same extent both maternally (Two-way ANOVA p>0.05 ns.) and zygotically (p>0.05 ns.) in wild-type embryos. All PCR bands shown in these panels were cloned and sequenced to confirm their identity. Green dots indicate correctly spliced transcripts, red dots indicate unspliced transcripts (intron retention).

DOI: 10.7554/eLife.02181.013

are frequently intronless when compared with the rest of the transcriptome (***Biedler et al., 2012***; ***Heyn et al., 2014***). This suggests that highly proliferative tissues need coordination between cell cycle and gene architecture for correct gene expression and avoidance of abnormally processed transcripts.

Our results highlight cell cycle constraints during early embryonic development as a force capable of driving changes in gene architecture of multicellular organisms. In unicellular organisms intron paucity correlates with a bias toward the 5′ ends, whereas introns from multicellular genomes are evenly distributed throughout the genes (***Mourier and Jeffares, 2003***). This suggests that similar constraints on gene architecture are also likely to exist in yeast and other fast-dividing single cell eukaryotes.

The way splicing efficiency might be regulated through changes in constitutive spliceosome factors and how it might influence differential gene expression is a new area of interest. In this study, we present experimental evidence supporting the hypothesis of a requirement for highly efficient pre-mRNA splicing during early embryonic development. Since the NTC/Prp19 complexes are known to be important for efficient spliceosome activation, and our mutant alleles specifically impaired pre-mRNA splicing of early zygotic but not maternally encoded transcripts, we propose that overall requirements for splicing efficiency are likely to vary during development, being the NTC/Prp19 complexes a key modulator of spliceosome activation rates. In agreement with this hypothesis, Prp19 expression varies during neuronal differentiation (***Urano et al., 2006***).

In plants it was recently shown that removal of retained introns regulates translation in rapidly developing gametophytes (***Boothby et al., 2013***). In *Drosophila*, a sub-population of early zygotic transcripts with introns similarly showed some degree of intron retention in wild-type embryos (***Figure 2B,D***, ***Figure 2—figure supplement 2B***). Our results also suggest that the pre-requisite for highly efficient splicing during early embryonic development is paradoxically also likely to play an important regulatory role in the expression of a subset of early zygotic transcripts, which further supports the possibility that modulation of spliceosome activation per se is important for differential regulation of gene expression during development.

# Materials and methods

## Fly work and genetics

Flies were raised using standard techniques. The *fandango* alleles were isolated in a previously reported maternal screen (***Pimenta-Marques et al., 2008***). Maternal mutant embryos and germ-line mutant clones were generated using the FLP/FRT $ovo^D$ system (***Chou and Perrimon, 1992***). Germ-line clones of $fand^1$ and $fand^2$ were established by crossing FRT42B $fand^1$/CyO or FRT42B $fand^2$/CyO virgins to *hs*-FLP; FRT42B $ovo^D$/CyO males and the progeny was heat shocked once at 37°C for 1 hr during second and third larval instar stages. As control we generated germ-line clones with FRT42B by crossing FRT42B/CyO virgins to *hs*-FLP; FRT42B $ovo^D$/CyO males and followed by the heat shock procedure described before.

To generate homozygous mutant clones in ovaries for $fand^1$ (negative for nuclear GFP label, nGFPminus) we used FLP/FRT to induce mitotic recombination. Females *y, w, hs*-FLP; FRT42B nGFP/CyO *hs-hid* flies were crossed with *w*; FRT42B, $fand^1$/CyO *hs-hid* males. Recombination was induced by 1-hr heat shock at 37°C at second and third instar larval stage. Adult ovaries were harvested from 4–5-day-old females and subsequently processed for immunofluorescence.

Viability and phenotypes of *fandango* alleles were complemented by crossing a transgenic fly carrying a genomic fragment construct that contained a wild-type copy of CG6197 (*wt-fandango*). w; FRT42B, $fand^1$/CyO virgins were crossed to *wt-fandango*; FRT42B, $fand^2$/CyO males; reciprocal crosses were also performed. Offspring were counted to determine viability. Rescue of maternal phenotypes

(cellularized blastoderm defects and splicing defects in early zygotic transcripts) was also analyzed in embryos laid by F1 *wt-fandango*/ +; FRT42B, *fand*$^{1}$/FRT42B *fand*$^{2}$ females. Germ-line clones of *fand*$^{1}$ and *fand*$^{2}$ were also rescued (cellularized blastoderm defects and splicing defects in early zygotic transcripts) by a copy of *wt-fandango* in the third chromosome. FRT42B *fand*$^{1}$/CyO; *wt-fandango* or FRT42B *fand*$^{2}$/CyO; *wt-fandango* virgins were crossed with *hs*-FLP; FRT42B *ovo*$^{D}$/CyO males and heat shock performed as described above.

To induce maternal and zygotic expression of the *UAS-kuk-LacZ* construct in control and *fandango* maternal mutant embryos, we performed the following crosses:

Maternal expression in control genetic background: virgin females +/+; *Nanos-Gal4, UAS-kuk-LacZ/TM6B* crossed with wild-type males.

Zygotic expression in control genetic background: virgin females +/+; *actin-Gal4/TM6B* crossed with +/+; *UAS-kuk-LacZ* males.

Maternal expression in *fandango* maternal mutant genetic background: firstly, virgin females FRT42B *fand*$^{1}$/CyO; *Nanos-Gal4, UAS-kuk-LacZ/TM6B* crossed with *hs*-FLP; FRT42B *ovo*$^{D}$/CyO males, and heat shocked as described above. After eclosion, Cy$^{+}$ and Tb$^{+}$ females were selected from the progeny and crossed to wild-type males.

Zygotic expression in *fandango* maternal mutant genetic background: firstly, virgin females FRT42B *fand*$^{1}$/CyO; *actin-Gal4/TM6B* were crossed with *hs*-FLP; FRT42B *ovo*$^{D}$/CyO males, and heat shocked as described. After eclosion, Cy$^{+}$ and Tb$^{+}$ virgin females were selected from the progeny and crossed to +/+; *UAS-kuk-LacZ* males.

To induce maternal and zygotic expression of the *4x intron kuk-LacZ* and *no intron kuk-LacZ* constructs we performed following crosses:

Maternal expression: firstly, virgin females +/+; *actin-Gal4/TM6B* crossed with +/+; *UAS-4xintron-kuk-LacZ/TM6B* or *UAS-nointron-kuk-LacZ/TM6B* males. After eclosion, females Tb$^{+}$ (+/+; *actin-Gal4/UAS-4xintron-kuk-LacZ* or +/+; *actin-Gal4/UAS-nointron-kuk-LacZ*) were selected and crossed with wild-type males.

Zygotic expression: virgin females +/+; *actin-Gal4/TM6B* were crossed with +/+; *UAS-4xintron-kuk-LacZ/TM6B* or *UAS-nointron-kuk-LacZ/TM6B* males.

To analyze zygotic expression of the *4x intron kuk-LacZ* and *no intron kuk-LacZ* constructs under the control of the minimal promoter of the gene *nullo*, females carrying the corresponding construct were selected and crossed with wild-type males.

To drive embryonic and ovarian expression of Myc-tagged Fandango and Myc-tagged Prp19 proteins, *Nanos-Gal4* homozygous virgins were crossed with UAS-*Fandango-6xMyc* males or *UAS-Prp19-6xMyc/TM6B* males, respectively. After eclosion females (in case of Myc-Fandango) or Tb$^{+}$ females (in case of Myc-Prp19) were selected, dissected ovaries from 4–5-day-old females, or laid embryos after a cross with wild-type males were used for protein extraction.

## Cloning of *fandango* alleles

To identify the gene responsible for lethality in *fandango* alleles, we performed a complementation analysis using the Bloomington 2R Deficiency kit.

Deficiency Df(2R)CX1 (covering an interval from cytological band 49C1 to 50D2, Bloomington stock number 442) failed to complement zygotic viability of both *fandango* alleles (complementation group 7). All additional 22 overlapping deficiencies complemented both *fandango* alleles. The cytological interval between bands 50B4-B6 (comprising 6 genes) was not covered by the 22 deficiencies. We cloned and sequenced these 6 genes from genomic DNA of both control and *fandango* alleles and identified mutations in gene CG6197 in both *fandango* alleles.

To confirm the identity of our mutants, we digested DNA from genomic clone (BACR14P04, Flybase) with restriction enzymes *XbaI* and *EcoRI* to generate a genomic fragment comprising the wild-type gene sequence of CG6197 (*wt-fandango*). Then we cloned the fragment into pCasper vector and used it to generate transgenic stocks (Bestgene, Chino Hills, CA, USA). A genomic wild-type copy of CG6197 under the control of its endogenous promoter fully complemented all known phenotypes in both *fandango* alleles.

## Immunohistochemistry

0–3 hr (after egg laying) embryos, both maternally mutant for *fandango* and control, were fixed and stained using standard procedures (***Pimenta-Marques et al., 2008***). For Neurotactin and Slam

immunostaining, the fixation procedure was modified: embryos were added to boiling heat fix solution (68 mM NaCl +0.1% Triton) and stirred for 1 min, then cooled by adding an equal volume of cooled fix solution. Immunostaining for oogenesis phenotypic analysis was performed as described in ***Guilgur et al. (2012)***. Following primary antibodies used were: mouse anti-Neurotactin clone BP106 at 1:133 (DSHB, Iowa City, Iowa, USA); mouse anti-pTyr at 1:1000 (9411; Cell Signaling, Danvers, MA, USA), and rabbit anti-slam at 1:1000 (Ruth Lehman Lab). For F-actin staining, a 5-min incubation with phalloidin-Rhodamine at 1:200 dilution (Sigma, St Louis, MO, USA; stock concentration 1 mg/ml) was employed at room temperature. For DNA staining, we used SYTOX Green (Invitrogen, Grand Island, NY, USA) at 1:5000 dilution with 5 mg/ml RNase A in PBT (PBS+0.1% Tween-20) for 30 min at room temperature. Cy3- or Cy5-conjugated secondary antibodies were used at 1:1000 dilution (Jackson ImmunoResearch, West Grove, PA, USA) and anti-rabbit Alexa Fluor 488 at 1:1000 dilution (Molecular Probes, Grand Island, NY, USA).

## Generation of constructs and cloning

The *kuk-LacZ* construct was synthesized using the 5′UTR and intron of the kuk small transcript (kuk-RB, Flybase). The kuk ORF was replaced by the LacZ coding sequence and was followed by the 3′ UTR of the original transcript (***Figure 5A***). The *kuk-LacZ* construct was fused to a UASg promoter (Gateway system, Invitrogen, Grand Island, NY, USA).

The *4x intron kuk-LacZ* construct was synthesized using 4 repeats of the fragment of 5′UTR and intron of the kuk small transcript, separated by 201 nucleotides of in-frame LacZ sequence. The stop codon is followed by the 3′UTR kuk small transcript sequence and 300 bp of the 3′-located genomic region to promote transcriptional termination (***Figure 6A,D***). In the case of the *no intron kuk-LacZ*, all splice sites (meaning 5′ splice site, branch point, and 3′ splice site) were mutated to thymidines (***Figure 6A,D***). To induce expression of these constructs, they were fused to UAS promoter or *nullo* minimal promoter. The *4x intron kuk-LacZ* and *no intron kuk-LacZ* constructs were cloned into pWALIUM22.

Fandango open reading frame, *kuk-LacZ*, *4x intron kuk-LacZ,* and *no intron kuk-LacZ* constructs were synthetized (GenScript, Piscataway, NJ, USA). Prp19 open reading frame was cloned into pDONR221 from DGC gold BDGP clone LD09231.

Prp19 and Fandango ORFs were subcloned into a vector containing the UASp promoter and 6x C-terminal Myc-tag (Gateway, Invitrogen, Grand Island, NY, USA). All constructs were then used to generate transgenic flies stocks (BestGene, Chino Hills, CA, USA).

## RT-PCR

Total RNA was extracted from 0–3 hr (after egg laying) embryos, unfertilized embryos, and 4-day-old female ovaries mutant for *fandango* and control (FRT42B) following standard procedures (PureLink RNA Mini Kit, Ambion, Grand Island, NY, USA). 1 µg of RNA was then used for reverse transcription with Oligo(dT)12–18 and/or random hexamers primers following the manufacturer's protocol (Transcriptor First Strand cDNA Synthesis Kit, ROCHE, Germany). Primer combinations used were designed with PrimerSelect (Lasergene, Madison, WI, USA) and PCR was performed using GoTaq DNA polymerase (Promega, Fitchburg, WI, USA). Sequences of all primers used are listed in ***Supplementary file 2***.

## Real-time qPCR

To measure transcription levels, embryos were staged in three different groups based on the embryonic morphology: stage A (embryos from cycle 1 to 8, no pole cells, and no cortical nuclei are observed); stage B (embryos from cycle 8/9 to 13, pole cells present, and cortical nuclei are observed); and stage C (embryos at interphase 14, blastoderm cellularized). Three independent replicas for each stage, containing each 10 manually selected embryos were generated. Three different genetic backgrounds were analyzed (control (FRT42B), FRT42B *fandango*, and *grapes* as positive control). To measure fandango mRNA levels, unfertilized eggs were analyzed (three replicas). To measure transcription level of the *4x intron kuk-LacZ* and *no intron kuk-LacZ* constructs, 0–3 hr (after egg laying) embryo collections were used to analyze both maternal and zygotic induced expression.

Total RNA was extracted from samples and then used for reverse transcription with Oligo(dT)12–18 as described above. Real-time mRNA quantification was performed following the manufacturer's protocol (QuantiFast SYBR Green RT-PCR Kit, Qiagen, Germany). For analysis of transcription levels of early zygotic genes (*nullo*, *snail*, *scute*, *even-skipped,* and *tailless*) the *Drosophila* QuantiTect Primer Assay (Qiagen, Germany) was used. For mRNA level measurements of fandango, *4x intron kuk-LacZ* and *no intron kuk-LacZ* constructs primers were designed with Primer3 (***Supplementary file 2***).

## Antibodies generated

Anti-Fandango and anti-Prp19 rabbit polyclonal antibodies were raised against recombinant proteins corresponding to amino acids 551–750 of Fandango/CG6197-PA, and to amino acids 20–219 of Prp19-PA, respectively (Metabion international AG, Germany). In both cases it was used His-tagged recombinant proteins as antigen and the antibodies were affinity purified.

## Biochemistry

Protein extracts were obtained from 0–3 hr (after egg laying) embryos or 4-day-old female ovaries. Embryos were dechorionated with 50% commercial bleach solution and ovaries dissected in PBS, samples then homogenized in NB buffer (150 mM NaCl, 50 mM Tris–HCl pH 7,5, 2 mM EDTA, 0,1% NP-40, 1 mM DTT, 10 mM NaF, and EDTA-free protease inhibitor cocktail, Roche, Germany), and centrifuge at 20000×*g* for 3 min. Supernatant was recovered and centrifuged twice.

To analyze NTC/Prp19 complex composition (***Table 1***), co-immunoprecipitation was done using protein extracts from embryo or ovary tissues expressing Myc-tagged Fandango or Prp19. Briefly, 1 mg of protein was diluted in 1 ml NB buffer and incubated with 1 µg/ml of mouse c-Myc antiboby (9E10) (Santa Cruz Biotechnology, Dallas, Texas, USA) for 1 hr at 4°C. Subsequently, 0.9 mg of Dynabeads Protein G (Invitrogen, Grand Island, NY, USA) were added to the immune complex and incubated 1 hr at 4°C. Beads were washed three times with NB buffer and protein elution performed with 50 µl of 100 mM Glycine pH 2.5 during 2 min at RT and stopped with 5 µl of 1M Tris–HCl pH 10.85. Eluted proteins were then precipitated in five times the volume of acetone at −20°C and samples analyzed by liquid chromatography coupled to tandem mass spectrometry (Mass Spectrometry Laboratory, Institute of Biochemistry and Biophysics, Poland).

To analyze NTC/Prp19 complex composition (showed in ***Figure 3A***), protein co-immunoprecipitation was performed using nuclear protein extracts (adapted from ***Kamakaka and Kadonaga, 1994***) from a collection of 0–3 hr (after egg laying) wild-type embryos (Oregon-R). 1 mg of protein extract was incubated with rabbit anti-Prp19 (1:1000 dilution) or the pre-immune (1:10,000 dilution) as control, in HNEB2 buffer (100 mM NaCl, 2,5 mM $MgCl_2$, 10 mM Tris–HCl pH 7,5, 0,5% Triton X-100, and EDTA-free protease inhibitor cocktail, Roche, Germany) during 1 hr at 4°C. The procedure was carried out as described in previous co-immunoprecipitation and eluted complexes were boiled in Laemmli sample buffer and analyzed by Western Blot.

Size-exclusion chromatography was performed in protein extracts of 0–3 hr (after egg laying) embryo collections from FRT42B *fand*$^{1}$ mutants or control (FRT42B). Extracts were prepared as described before in NB2 buffer (150 mM NaCl, 50 mM Tris–HCl pH 7,5, 2 mM EDTA, 0,01% NP-40, 1 mM DTT, and EDTA-free protease inhibitor cocktail, Roche, Germany). Subsequently, 2 mg of protein extract were fractionated using Superose 6 10/300 GL column (GE Healthcare, United Kingdom) in NB2 buffer and fractions collected and analyzed by Western blot.

To analyze protein amount in ovaries and embryos (showed in ***Figures 1I and 3C***), embryos were dechorionated and ovaries dissected as described above. Samples were homogenized in PBS supplemented with EDTA-free protease inhibitor cocktail (Roche, Germany) and centrifuged at 20000×*g* for 3 min at 4°C. Supernatant was collected and protein concentration determined using the Bradford method (BioRad, Hercules, CA, USA). Samples were immediately boiled in Laemmli sample buffer and 10 µg of protein was run in SDS-PAGE gel and analyzed by immunoblot.

Levels of RNApol II CTD Ser2 phosphorylation were analyzed in embryos dechorionated and manually selected at specific developmental stages based on the embryonic morphology (as described above). 15 embryos were selected for each stage and protein sample was obtained by lysing the embryos with a needle in Laemmli sample buffer and heating for 5 min at 100°C. Protein amounts corresponding to ~7 embryos were running in SDS-PAGE and analyzed by immunoblot. Five independent replicas were analyzed.

Antibodies used were: polyclonal rabbit anti-Prp19 at 1:8000 dilution; polyclonal rabbit anti-Fandango at 1:1000 dilution; mouse anti-alpha-Tubulin Dm1A at 1:50,000 dilution (Sigma, St Louis, MO, USA); mouse anti-RNA Polymerase II H5 at 1:500 dilution (MMS-129R, Covance, Princeton, NJ, USA); rabbit anti-ISY1 at 1:500 dilution (ab121250; Abcam, United Kingdom); and mouse anti-CDC5L [2136C1a] at 1:200 dilution (ab51320; Abcam, United Kingdom).

## Mass spectrometry

Peptides mixtures were analyzed by LC-MS-MS/MS (liquid chromatography coupled to tandem mass spectrometry) using Nano-Acquity (Waters, Milford, MA, USA) LC system and Orbitrap Velos mass

spectrometer (Thermo Electron Corp., San Jose, CA, USA). Prior to analysis, proteins were subjected to standard 'in-solution digestion' procedure, during which proteins were reduced with 100 mM DTT (for 30 min at 56°C), alkylated with 0,5 M iodoacetamide (45 min in darkroom at room temperature), and digested overnight with trypsin (sequencing Grade Modified Trypsin—Promega V5111). The peptide mixture was applied to an RP-18 precolumn (nanoACQUITY Symmetry C18—Waters 186003514) using water containing 0,1% TFA as mobile phase, then transferred to nano-HPLC RP-18 column (nanoACQUITY BEH C18–Waters 186003545) using an acetonitrile gradient (0%–35% AcN in 180 min) in the presence of 0.05% formic acid with a flow rate of 250 nl/min. The column outlet was directly coupled to the ion source of the spectrometer, operating in the regime of data dependent MS to MS/MS switch. A blank run ensuring no cross contamination from previous samples preceded each analysis.

Raw data were processed by Mascot Distiller followed by Mascot Search (Matrix Science, London, UK, on-site license) against Flybase database. Search parameters for precursor and product ions mass tolerance were 100 ppm and 0.6 Da, respectively, enzyme specificity: trypsin, missed cleavage sites allowed: 0, fixed modification of cysteine by carbamidomethylation, and variable modification of methionine oxidation. Peptides with Mascot Score exceeding the threshold value corresponding to <5% False Positive Rate, calculated by Mascot procedure, and with the Mascot score above 30 were considered to be positively identified.

Human orthologs were determined using DSRC Integrative Ortholog Prediction Tool (DIOPT) (http://www.flyrnai.org/cgi-bin/DRSC_orthologs.pl). Only scores above two were considered such as the best matches when there was more than one match per input.

## High-throughput transcriptome sequencing (RNAseq)

Total RNA was isolated from 0–3 hr collections of *fandango* maternal mutant and control (FRT42B) embryos using TRIzol Reagent (Invitrogen, Grand Island, NY, USA), following standard protocol. DNAse I (Promega, Fitchburg, WI, USA) treatment was performed during 30 min at 37°C. DNAse was extracted by Phenol-Chloroform extraction; the RNA was precipitated with ethanol, and dissolved it in 25 µl of DEPC water. Bioanalyzer testing was used to analyze quality and concentration of the samples and made up to the volume to 100 µl with water, 50 µl of 7.5 M NH4OAc added, 0.5 µl of glycogen, and 250 µl of absolute ethanol. cDNA library was generated applying the standard Illumina protocol for RNA-Seq (polyA RNAs) and sequenced with an Illumina HiSeq (Oklahoma Medical Research Foundation, Oklahoma City, OK, USA). These generated RNA-Seq data for two biological replicates each of wild-type and *fandango* mutant (*fand*$^{1}$), consisting of about 150 million illumina paired-end 100 bp reads per sample.

Paired-end reads were mapped with tophat (***Trapnell et al., 2009***) version 2.0.3 against the *Drosophila melanogaster* BDGP5 reference genome, using Flybase gene annotations downloaded from Ensembl e66 as guide.

To analyze splicing defects, we first extracted exon–intron boundaries from gene annotations. To avoid potential confounding effects, we removed all boundaries that had overlapping exon sequence (from different genes or transcripts). Subsequent analysis used this set of 'safe' exon–intron boundaries.

For coverage plots in ***Figure 2D***, we also excluded boundaries where the intron or the exon were less than 50 b long. At each base within 50 bp either side of a splice site we count the number of reads that overlap that base, then divide by the total number of reads within the 100 bp centered around the splice site. To minimize noise, we require that at least 50 reads fall within the −50:50 window around the exon–intron boundary (reads that only partially overlap the window are also counted).

To determine the frequency of splicing defects for each boundary, we extracted all reads that overlap the 5′ splice site by at least 10 bp to either side of the boundary. We classified each read as correctly spliced (if the read is split from the 5′ to the 3′ splice site), unspliced (if the read is not split) or mis-spliced (if the read is split but not matching the expected 5′ or/and 3′ sites). To reduce noise, we only include an exon–intron boundary if at least 10 fragments overlap that boundary.

To determine the exon–intron gene structure (***Figure 2—figure supplement 2D***), each gene was divided in 1000 equal segments. For each segment of each gene, we checked for the presence (or absence) of an exon in that segment. For each segment, we then plotted the frequency of exon presence in all genes. If an exon randomly appears in a given segment, it appears in ~50% of genes. For example, a set of intronless genes would produce a plot that would be always at 100%.

To determine the splice site motif (***Figure 2—figure supplement 2C***), sequences around exon–intron boundaries were extracted and motifs drawn using Weblogo.

Early zygotic and maternal genes were defined using RNA-Seq developmental gene expression data from Flybase (***Graveley et al., 2011***). A gene was defined as an early zygotic gene when its expression at 2–4 hr is at least moderate (more than 10 expression units, as defined in the Flybase dataset) and at least 5× greater than its expression at 0–2 hr (irrespective of the 0–2 hr value). Maternal genes are those genes that are not early zygotic and have high expression (more than 50 expression units) at 0–2 hr. To avoid potential artifacts, genes that have an extremely high expression (more than 1000 expression units) were not considered. Applying this definition we obtained 270 early zygotic genes (including 43 genes from ***De Renzis et al., 2007***) and 2048 maternal genes.

All scripts used for this analysis are available upon request. RNA-Seq data are available in the ArrayExpress database (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-2321.

### In situ hybridization

The procedure has been described in ***Stein et al. (2002)***. Antisense digoxigenin-labeled RNA probes were synthesized using the DIG RNA labeling Kit (Roche, Germany). *eve* and *nos* probes were made from pBluescript plasmids containing the respective cDNAs.

### Sequence alignment

Sequences were aligned using ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/) and BoxShade 3.21 (http://www.ch.embnet.org/software/BOX_form.html) for printing and shading of multiple alignment file.

### Statistical analysis

Unpaired *t* test and two-way ANOVA were performed using Prism 5.00 for Windows (GraphPad Software, San Diego, CA, USA).

## Acknowledgements

Both *fandango* mutant alleles were isolated in the laboratory of Ruth Lehmann. We thank Kohtaro Tanaka for help with the *in situs*; our colleagues, Moises Mallo, Miguel Ferreira for discussion, and suggestions that greatly improved the manuscript; Jessica Thompson and Richard Hampson for manuscript editing. Proteins were identified at the Mass Spectrometry Laboratory Institute of Biochemistry and Biophysics Polish Academy of Science.

## Additional information

### Funding


| Funder | Grant reference number | Author |
|---|---|---|
| FCT-Fundacao para a Ciencia e Tecnologia (Portugal) | PTDC/SAU-BID/111796/2009 | Rui Gonçalo Martinho |
| FCT-Fundacao para a Ciencia e Tecnologia (Portugal) | PTDC/BIA-BCM/111822/ 2009 | Rui Gonçalo Martinho |
| FCT-Fundacao para a Ciencia e Tecnologia (Portugal) | PTDC/BBB-BQB/0712/2012 | Rui Gonçalo Martinho |
| FCT-Fundacao para a Ciencia e Tecnologia (Portugal) | PEst-OE/EQB/LA0023/2013 | Rui Gonçalo Martinho |
| FCT-Fundacao para a Ciencia e Tecnologia (Portugal) | SFRH/BPD/47957/2008 | Leonardo Gastón Guilgur |
| FCT-Fundacao para a Ciencia e Tecnologia (Portugal) | SFRH/BPD/63869/2009 | Denisa Liszekova |

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

### Author contributions

LGG, Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; PP, Conception and design, Acquisition of data, Analysis and interpretation of data; DS, DL, AR, Acquisition of data, Analysis and interpretation of data; RGM, Conception and design, Analysis and interpretation of data, Drafting or revising the article

### Author ORCIDs

Rui Gonçalo Martinho, http://orcid.org/0000-0002-1641-3403

## Additional files

### Supplementary files

• Supplementary file 1. Complete list of proteins specifically co-immunoprecipitating with Myc-tagged Fandango and Myc-tagged Prp19. List of proteins co-immunoprecipitating with Myc-tagged Fandango in ovaries (germ-line) and embryos, and with Myc-tagged Prp19 in embryos. Protein extracts from embryos and ovaries not expressing the Myc-tagged proteins were used as negative controls (−). Two replica experiments were performed for each condition. Proteins were identified by LC-MS-MS and data were blasted against the Flybase database using Mascot Search (see 'Materials and methods' for detail). The presence of a protein was measured from the total number of peptides (Matches), the number of non-repeated sequence peptides (Sequences) and the respective Score according to Mascot Search. Predictions of Human protein orthologous from *Drosophila* proteins were made using DIOPT (see 'Materials and methods' for detail). Proteins were grouped in NTC/Prp19-related complex (light gray), NTC/Prp19 complex (dark gray), general spliceosomal related proteins, miscellaneous, and ribosomal proteins. Non-reproducible (only 1 peptide in one replica) and non-specific proteins (more than 1 peptide in negative controls [−]) are listed separately. (*) Spliceosomal proteins in non-reproducible table, (**) NTC/Prp19 complex or NTC/Prp19-related complex subunits present in non-reproducible and non-specific tables.

DOI: 10.7554/eLife.02181.014

• Supplementary file 2. Complete list of primers. Sequences of all the primers used in the RT-PCR and real-time qPCR assays presented in this manuscript ('Materials and methods').

DOI: 10.7554/eLife.02181.015

### Major dataset

The following dataset was generated:

| Author(s) | Year | Dataset title | Dataset ID and/or URL | Database, license, and accessibility information |
|---|---|---|---|---|
| Guilgur Leonardo, Prudêncio Pedro, Sobral Daniel, Liszekova Denisa, Rosa André, Martinho Rui | 2014 | RNA-seq of coding RNA of fandango, a Drosophila melanogaster mutant affecting splicing | http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2321/ | Publicly available at EMBL-EBI. |

## References

**Alexander RD**, Barrass JD, Dichtl B, Kos M, Obtulowicz T, Robert MC, Koper M, Karkusiewicz I, Mariconti L, Tollervey D, Kufel J, Bertrand E, Beggs JD. 2010. RiboSys, a high-resolution, quantitative approach to measure the in vivo kinetics of pre-mRNA splicing and 3′-end processing in Saccharomyces cerevisiae. *RNA* **16**:2570–2580. doi: 10.1261/rna.2162610.

**Ameur A**, Zaghlool A, Halvardson J, Wetterbom A, Gyllensten U, Cavelier L, Feuk L. 2011. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nature Structural & Molecular Biology* **18**:1435–1440. doi: 10.1038/nsmb.2143.

**Anderson KV**, Lengyel JA. 1979. Rates of synthesis of major classes of RNA in Drosophila embryos. *Developmental Biology* **70**:217–231. doi: 10.1016/0012-1606(79)90018-6.

**Audibert A**, Weil D, Dautry F. 2002. In vivo kinetics of mRNA splicing and transport in mammalian cells. *Molecular and Cellular Biology* **22**:6706–6718. doi: 10.1128/MCB.22.19.6706-6718.2002.

**Biedler JK**, Hu W, Tae H, Tu Z. 2012. Identification of early zygotic genes in the yellow fever mosquito *Aedes aegypti* and discovery of a motif involved in early zygotic genome activation. *PLOS ONE* **7**:e33933. doi: 10.1371/journal.pone.0033933.

**Boothby TC**, Zipper RS, van der Weele CM, Wolniak SM. 2013. Removal of retained introns regulates translation in the rapidly developing gametophyte of Marsilea vestita. *Developmental Cell* **24**:517–529. doi: 10.1016/j.devcel.2013.01.015.

**Brandt A**, Papagiannouli F, Wagner N, Wilsch-Bräuninger M, Braun M, Furlong EE, Loserth S, Wenzl C, Pilot F, Vogt N, Lecuit T, Krohne G, Grosshans J. 2006. Developmental control of nuclear size and shape by Kugelkern and Kurzkern. *Current Biology* **16**:543–552. doi: 10.1016/j.cub.2006.01.051.

**Brody Y**, Neufeld N, Bieberstein N, Causse SZ, Bohnlein EM, Neugebauer KM, Darzacq X, Shav-Tal Y. 2011. The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing. *PLOS Biology* **9**:e1000573. doi: 10.1371/journal.pbio.1000573.

**Chan SP**, Kao DI, Tsai WY, Cheng SC. 2003. The Prp19p-associated complex in spliceosome activation. *Science* **302**:279–282. doi: 10.1126/science.1086602.

**Chanarat S**, Seizl M, Strasser K. 2011. The Prp19 complex is a novel transcription elongation factor required for TREX occupancy at transcribed genes. *Genes & Development* **25**:1147–1158. doi: 10.1101/gad.623411.

**Chang KJ**, Chen HC, Cheng SC. 2009. Ntc90 is required for recruiting first step factor Yju2 but not for spliceosome activation. *RNA* **15**:1729–1739. doi: 10.1261/rna.1625309.

**Chou TB**, Perrimon N. 1992. Use of a yeast site-specific recombinase to produce female germline chimeras in Drosophila. *Genetics* **131**:643–653.

**David CJ**, Boyne AR, Millhouse SR, Manley JL. 2011. The RNA polymerase II C-terminal domain promotes splicing activation through recruitment of a U2AF65-Prp19 complex. *Genes & Development* **25**:972–983. doi: 10.1101/gad.2038011.

**de la Mata M**, Alonso CR, Kadener S, Fededa JP, Blaustein M, Pelisch F, Cramer P, Bentley D, Kornblihtt AR. 2003. A slow RNA polymerase II affects alternative splicing in vivo. *Molecular Cell* **12**:525–532.

**De Renzis S**, Elemento O, Tavazoie S, Wieschaus EF. 2007. Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo. *PLOS Biology* **5**:e117. doi: 10.1371/journal.pbio.0050117.

**Foe VE**, Alberts BM. 1983. Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles that precede gastrulation in Drosophila embryogenesis. *Journal of Cell Science* **61**:31–70.

**Graveley BR**, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, Brown JB, Cherbas L, Davis CA, Dobin A, Li R, Lin W, Malone JH, Mattiuzzo NR, Miller D, Sturgill D, Tuch BB, Zaleski C, Zhang D, Blanchette M, Dudoit S, Eads B, Green RE, Hammonds A, Jiang L, Kapranov P, Langton L, Perrimon N, Sandler JE, Wan KH, Willingham A, Zhang Y, Zou Y, Andrews J, Bickel PJ, Brenner SE, Brent MR, Cherbas P, Gingeras TR, Hoskins RA, Kaufman TC, Oliver B, Celniker SE. 2011. The developmental transcriptome of *Drosophila melanogaster*. *Nature* **471**:473–479. doi: 10.1038/nature09715.

**Guilgur LG**, Prudencio P, Ferreira T, Pimenta-Marques AR, Martinho RG. 2012. Drosophila aPKC is required for mitotic spindle orientation during symmetric division of epithelial cells. *Development* **139**:503–513. doi: 10.1242/dev.071027.

**Herold N**, Will CL, Wolf E, Kastner B, Urlaub H, Luhrmann R. 2009. Conservation of the protein composition and electron microscopy structure of *Drosophila melanogaster* and human spliceosomal complexes. *Molecular and Cellular Biology* **29**:281–301. doi: 10.1128/MCB.01415-08.

**Heyn P**, Kircher M, Dahl A, Kelso J, Tomancak P, Kalinka AT, Neugebauer KM. 2014. The earliest transcribed zygotic genes are short, newly evolved, and different across species. *Cell Reports* **6**:285–292. doi: 10.1016/j.celrep.2013.12.030.

**Hogg R**, McGrail JC, O'Keefe RT. 2010. The function of the NineTeen Complex (NTC) in regulating spliceosome conformations and fidelity during pre-mRNA splicing. *Biochemical Society Transactions* **38**:1110–1115. doi: 10.1042/BST0381110.

**Hsin JP**, Manley JL. 2012. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes & Development* **26**:2119–2137. doi: 10.1101/gad.200303.112.

**Huranova M**, Ivani I, Benda A, Poser I, Brody Y, Hof M, Shav-Tal Y, Neugebauer KM, Stanek D. 2010. The differential interaction of snRNPs with pre-mRNA reveals splicing kinetics in living cells. *The Journal of Cell Biology* **191**:75–86. doi: 10.1083/jcb.201004030.

**Ip JY**, Schmidt D, Pan Q, Ramani AK, Fraser AG, Odom DT, Blencowe BJ. 2011. Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. *Genome Research* **21**:390–401. doi: 10.1101/gr.111070.110.

**Kamakaka RT**, Kadonaga JT. 1994. The soluble nuclear fraction, a highly efficient transcription extract from Drosophila embryos. *Methods in Cell Biology* **44**:225–235.

**Khodor YL**, Rodriguez J, Abruzzi KC, Tang CH, Marr MT II, Rosbash M. 2011. Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes & Development* **25**:2502–2512. doi: 10.1101/gad.178962.111.

**Kuraoka I**, Ito S, Wada T, Hayashida M, Lee L, Saijo M, Nakatsu Y, Matsumoto M, Matsunaga T, Handa H, Qin J, Nakatani Y, Tanaka K. 2008. Isolation of XAB2 complex involved in pre-mRNA splicing, transcription, and transcription-coupled repair. *The Journal of Biological Chemistry* **283**:940–950. doi: 10.1074/jbc.M706647200.

**McKnight SL**, Miller OL Jnr. 1976. Ultrastructural patterns of RNA synthesis during early embryogenesis of *Drosophila melanogaster*. *Cell* **8**:305–319. doi: 10.1016/0092-8674(76)90014-3.

**Mourier T**, Jeffares DC. 2003. Eukaryotic intron loss. *Science* **300**:1393. doi: 10.1126/science.1080559.

**Nakatsu Y**, Asahina H, Citterio E, Rademakers S, Vermeulen W, Kamiuchi S, Yeo JP, Khaw MC, Saijo M, Kodo N, Matsuda T, Hoeijmakers JH, Tanaka K. 2000. XAB2, a novel tetratricopeptide repeat protein involved in transcription-coupled DNA repair and transcription. *The Journal of Biological Chemistry* **275**:34931–34937. doi: 10.1074/jbc.M004936200.

**Pilot F**, Philippe JM, Lemmers C, Chauvin JP, Lecuit T. 2006. Developmental control of nuclear morphogenesis and anchoring by charleston, identified in a functional genomic screen of Drosophila cellularisation. *Development* **133**:711–723. doi: 10.1242/dev.02251.
**Pimenta-Marques A**, Tostoes R, Marty T, Barbosa V, Lehmann R, Martinho RG. 2008. Differential requirements of a mitotic acetyltransferase in somatic and germ line cells. *Developmental Biology* **323**:197–206. doi: 10.1016/j.ydbio.2008.08.021.
**Pritchard DK**, Schubiger G. 1996. Activation of transcription in Drosophila embryos is a gradual process mediated by the nucleocytoplasmic ratio. *Genes & Development* **10**:1131–1142. doi: 10.1101/gad.10.9.1131.
**Rothe M**, Pehl M, Taubert H, Jackle H. 1992. Loss of gene function through rapid mitotic cycles in the Drosophila embryo. *Nature* **359**:156–159. doi: 10.1038/359156a0.
**Russell CS**, Ben-Yehuda S, Dix I, Kupiec M, Beggs JD. 2000. Functional analyses of interacting factors involved in both pre-mRNA splicing and cell cycle progression in *Saccharomyces cerevisiae*. *RNA* **6**:1565–1572.
**Schmidt U**, Basyuk E, Robert MC, Yoshida M, Villemin JP, Auboeuf D, Aitken S, Bertrand E. 2011. Real-time imaging of cotranscriptional splicing reveals a kinetic model that reduces noise: implications for alternative splicing regulation. *Journal of Cell Biology* **193**:819–829. doi: 10.1083/jcb.201009012.
**Shermoen AW**, O'Farrell PH. 1991. Progression of the cell cycle through mitosis leads to abortion of nascent transcripts. *Cell* **67**:303–310. doi: 10.1016/0092-8674(91)90182-X.
**Shin C**, Manley JL. 2002. The SR protein SRp38 represses splicing in M phase cells. *Cell* **111**:407–417. doi: 10.1016/S0092-8674(02)01038-3.
**Sibon OC**, Stevenson VA, Theurkauf WE. 1997. DNA-replication checkpoint control at the Drosophila midblastula transition. *Nature* **388**:93–97. doi: 10.1038/40439.
**Stein JA**, Broihier HT, Moore LA, Lehmann R. 2002. Slow as molasses is required for polarized membrane growth and germ cell migration in Drosophila. *Development* **129**:3925–3934.
**Tadros W**, Lipshitz HD. 2009. The maternal-to-zygotic transition: a play in two acts. *Development* **136**:3033–3042. doi: 10.1242/dev.033183.
**ten Bosch JR**, Benavides JA, Cline TW. 2006. The TAGteam DNA motif controls the timing of Drosophila pre-blastoderm transcription. *Development* **133**:1967–1977. doi: 10.1242/dev.02373.
**Trapnell C**, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**:1105–1111. doi: 10.1093/bioinformatics/btp120.
**Urano Y**, Iiduka M, Sugiyama A, Akiyama H, Uzawa K, Matsumoto G, Kawasaki Y, Tashiro F. 2006. Involvement of the mouse Prp19 gene in neuronal/astroglial cell fate decisions. *The Journal of Biological Chemistry* **281**:7498–7514. doi: 10.1074/jbc.M510881200.
**Villa T**, Guthrie C. 2005. The Isy1p component of the NineTeen complex interacts with the ATPase Prp16p to regulate the fidelity of pre-mRNA splicing. *Genes & Development* **19**:1894–1904. doi: 10.1101/gad.1336305.
**Yasuda GK**, Baker J, Schubiger G. 1991. Temporal regulation of gene expression in the blastoderm Drosophila embryo. *Genes & Development* **5**:1800–1812. doi: 10.1101/gad.5.10.1800.
**Zallen JA**, Wieschaus E. 2004. Patterned gene expression directs bipolar planar polarity in Drosophila. *Developmental Cell* **6**:343–355. doi: 10.1016/S1534-5807(04)00060-7.
**Zeytuni N**, Zarivach R. 2012. Structural and functional discussion of the tetra-trico-peptide repeat, a protein interaction module. *Structure* **20**:397–405. doi: 10.1016/j.str.2012.01.006.

Methods xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

# Analysis of Mammalian Native Elongating Transcript sequencing (mNET-seq) high-throughput data

Pedro Prudêncio[a,b,*,1], Kenny Rebelo[a,1], Ana Rita Grosso[a,d], Rui Gonçalo Martinho[a,b,c], Maria Carmo-Fonseca[a,*]

[a] Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal
[b] Center for Biomedical Research, Universidade do Algarve, Faro, Portugal
[c] iBiMED, Departamento de Ciências Médicas, Universidade de Aveiro, Aveiro, Portugal
[d] UCIBIO, Departamento de Ciências da Vida, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal

ARTICLE INFO

*Keywords:*
mNET-seq
RNA polymerase II
Nascent RNA
Co-transcriptional splicing

ABSTRACT

Mammalian Native Elongating Transcript sequencing (mNET-seq) is a recently developed technique that generates genome-wide profiles of nascent transcripts associated with RNA polymerase II (Pol II) elongation complexes. The ternary transcription complexes formed by Pol II, DNA template and nascent RNA are first isolated, without crosslinking, by immunoprecipitation with antibodies that specifically recognize the different phosphorylation states of the polymerase large subunit C-terminal domain (CTD). The coordinate of the 3′ end of the RNA in the complexes is then identified by high-throughput sequencing. The main advantage of mNET-seq is that it provides global, bidirectional maps of Pol II CTD phosphorylation-specific nascent transcripts and coupled RNA processing at single nucleotide resolution. Here we describe the general pipeline to prepare and analyse high-throughput data from mNET-seq experiments.

## 1. Introduction

The advent of high-throughput sequencing combined with innovative and diversified techniques to capture RNA molecules has enabled a new generation of genome-wide studies of transcription and co-transcriptional RNA processing. Native Elongating Transcript sequencing (NET-seq) was originally established in yeast to visualize the genomic position of the active site of RNA polymerase II (Pol II) by identifying the 3′ ends of the nascent RNA [1]. Because only the coordinates of the 3′ end nucleotides are recorded, single nucleotide resolution is achieved. NET-seq relies on the intrinsic stability of ternary transcription complexes (formed by Pol II, DNA template and nascent RNA) to isolate Pol II elongation complexes by immunoprecipitation without crosslinking. In the original study, a 3xFLAG epitope tag was added to the C-terminus of the third Pol II subunit (Rpb3), yeast cells were lysed and a crude whole-cell extract was used for immunoprecipitation using anti-FLAG antibodies [1,2]. Application of NET-seq to yeast revealed pervasive polymerase pausing and backtracking throughout gene transcription [1] and advanced our understanding of promoter directionality [1,3]. The NET-seq strategy was also used in bacteria to map the density of RNA polymerase, leading to the identification of novel pause sites across the genome [4,5].

In contrast to yeast and bacteria, solubilisation of Pol II complexes under native conditions is typically incomplete in metazoan cells [6]. A practical solution to this problem was found by Nojima and Proudfoot [7,8], who solubilized isolated native chromatin by extensive micrococcal nuclease (MNase) digestion and then immunoprecipitated elongation complexes using antibodies that specifically recognize the different phosphorylation states of the polymerase large subunit C-terminal domain (CTD). Although initially applied to mammalian cells, and thus termed mNET-seq, the procedure can be adapted to any organism, as recently illustrated in plants [9].

In the mNET-seq method, RNA is purified from the immunoprecipitated Pol II complexes and used to prepare a cDNA library for high-throughput Illumina sequencing (Fig. 1A). To enable directional sequencing, the 5′ hydroxyl (OH) generated by MNase digestion of RNA is first converted to a 5′ phosphate by T4 polynucleotide kinase. RNAs isolated from Pol II complexes are then size-selected on denaturing polyacrylamide gels before subsequent adapter ligation for PCR-based preparation of the cDNA library. Specific adapters are then ligated to the 5′ P and 3′ OH ends of each RNA fragment (Fig. 1A). The adapters consist of sequences used to amplify the library by PCR using generic forward

* Corresponding authors.
*E-mail address:* pprudencio@medicina.ulisboa.pt (P. Prudêncio).
[1] These authors have contributed equally to the work.

https://doi.org/10.1016/j.ymeth.2019.09.003
Received 3 May 2019; Received in revised form 17 July 2019; Accepted 1 September 2019
1046-2023/ © 2019 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

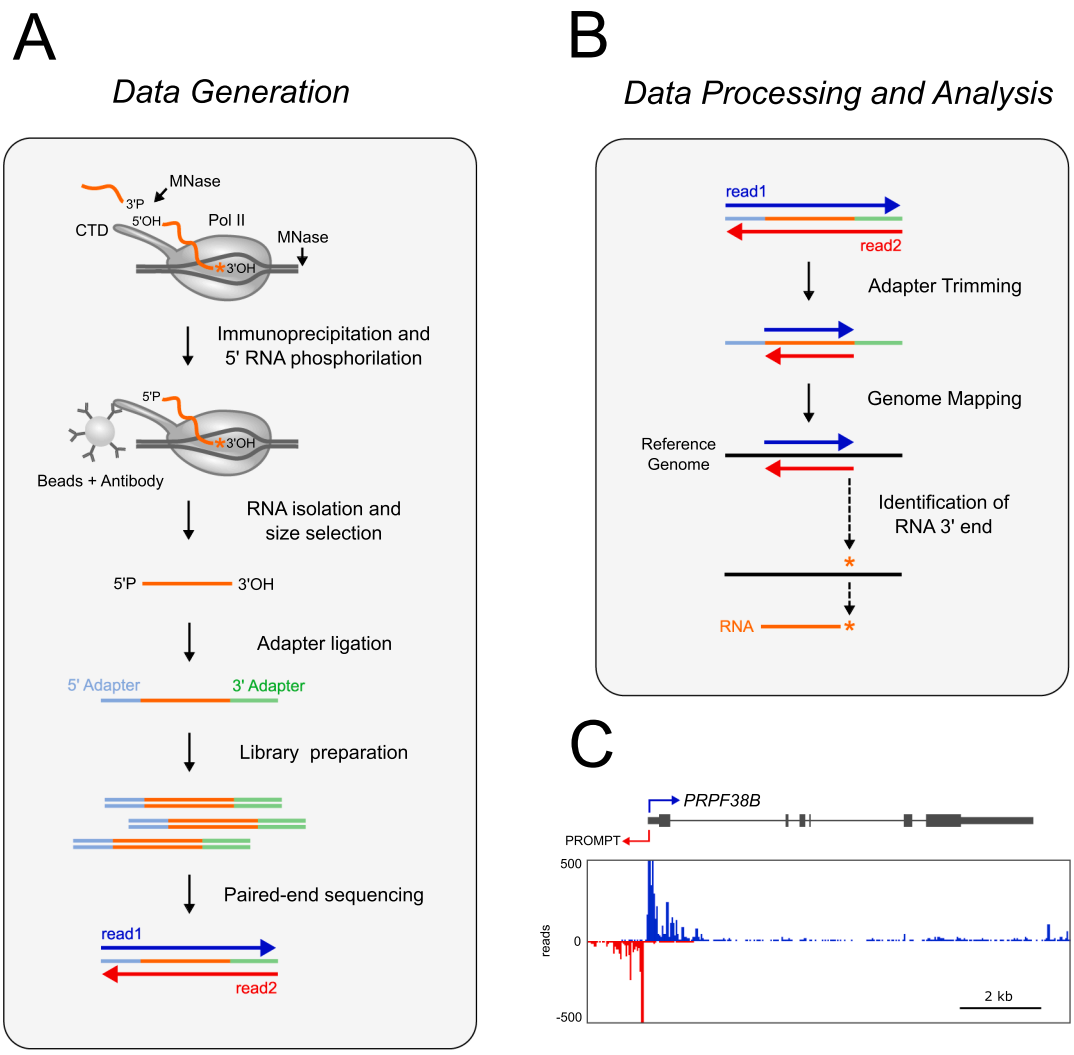



**Fig. 1.** Overview of mNET-seq. (A) Isolation of Pol II elongation complexes and library preparation. (B) Data processing. The orange asterisk denotes the nucleotide at the RNA 3′ end. (C) Visualization of mNET-seq profile along the *PRPF38B* gene. Data from HeLa cells 8WG16 mNET-seq replica1 [7] was aligned to the hg38 genome reference (GENCODEv28). Data visualized with UCSC genome browser. Blue and red arrows denote promoter bi-directionality. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and reverse primers, as well as sequences needed to associate the target nucleic acids with the sequencing instrument (e.g. the flowcell in Illumina sequencers) and, optionally, barcode sequences. After high-throughput sequencing, data must be prepared for analysis. This includes trimming of adapter sequences, mapping high quality reads to the reference genome, identification of the 3′ end nucleotide in each RNA fragment, and selection of genomic regions to be analyzed (Fig. 1B, C). In the following sections, we describe and discuss the primary analysis pipeline that we apply to data from mNET-seq experiments.

## 2. Methods

### 2.1. Quality control and adapter trimming

We use FastQC [10] for quality analysis of mNET-seq raw reads. As an example, we use the following FastQC (version 0.11.7) command to analyze data deposited in GSE106881 (GSM2856674 and GSM2856677):

```
fastqc mNET_Long_S5P_rep1_1.fastq
    fastqc mNET_Long_S5P_rep1_2.fastq
```

We further use FastQC to assess GC content, over-abundance of adapters and over-represented sequence, from which an indication of PCR duplication rate may be inferred [11]. Removal of adapter sequences and low quality reads is performed with Cutadapt [12]. We use Cutadapt (version 1.18) with an error rate of 0.05, and allow it to match 'N's in the reads to the adapter sequence; reads that are shorter than 10 bases are discarded, and the adapter is removed only once from each read. Example of Cutadapt command:

```
cutadapt -a TGGAATTCTCGGGTGCCAAGG -A GATCGTCGGACT
GTAGAACTCTGAAC –m
    10 -e 0.05 –match-read-wildcards -n 1 -o mNET_Long_
S5P_rep1_1_tr.fastq.gz –p
    mNET_Long_S5P_rep1_2_tr.fastq.gz    mNET_Long_S5P_rep1_1.
fastq
    mNET_Long_S5P_rep1_2.fastq
```

### 2.2. Mapping of reads to the reference genome

We initially used TopHat2 [13] for aligning mNET-seq reads to the reference human genome [7]. However, the currently most popular

mappers for RNA-seq data are STAR [14] and HISAT2 [15]. For STAR index generation, we set STAR (version 2.6.0b) to detect chimeric alignments with the minimum mapped length of at least 20nt on each end.

STAR index generation:

```
STAR –runMode genomeGenerate –genomeDir ./starIndex/
–genomeFastaFiles
    /genomes/human/hg38/GRCh38.primary.genome.fa
–-sjdbGTFfile
    /genomes/human/hg38/gencode.v28.annotation.gtf
```

Aligning paired-end reads:

```
STAR –runMode alignReads –genomeDir /genomes/human/
hg38/star/ –readFilesIn
    ./mNET_Long_S5P_rep1_1_tr.fastq.gz ./mNET_Long_S5P_rep1_
2_tr.fastq.gz–chimSegmentMin
    20 –outSAMtype BAM Unsorted –readFilesCommand gunzip -
c –outFileNamePrefix
    /alignments/mNET_Long_S5P_rep1_
```

The following command is used to obtain uniquely mapped reads with SAMtools (version 1.7):

```
samtools view -H mNET_Long_S5P_rep1_Aligned.out.bam
 > mNET_Long_S5P_rep1_header.sam
    samtools view -q 255 mNET_Long_S5P_rep1_Aligned.out.
bam >
    mNET_Long_S5P_rep1_unique.sam
    cat mNET_Long_S5P_rep1_header.sam mNET_Long_S5P_rep1_
unique.sam >
    mNET_Long_S5P_rep1_unique_H.sam
    samtools view -Sb -h mNET_Long_S5P_rep1_unique_H.sam >
    mNET_Long_S5P_rep1_unique.bam
    rm -f mNET_Long_S5P_rep1_unique_H.sam
    rm -f mNET_Long_S5P_rep1_unique.sam
```

An important mapping quality parameter is the percentage of mapped reads, which should always be higher than 70% [16]. We further use the RSeQC tool for quality control after mapping [17]. Only uniquely mapped reads are considered for further analysis.

### 2.3. Identification of RNA 3′ ends

NET-seq achieves single nucleotide resolution by mapping exclusively the nucleotide at the 3′ end of each immunoprecipitated RNA fragment. The full-length read sequences are discarded and only the coordinates of the 3′ end nucleotides are recorded as 1 M CIGAR strings. The RNA 3′ end corresponds to the 5′ nucleotide of read 2 in each sequencing pair, with the directionality indicated by read 1 (Fig. 1B). We do not use read 1 information because in sequencing-by-synthesis techniques accuracy decreases towards the 3′ end [18].

We developed a Python script for this purpose that is available in (https://github.com/kennyrebelo/mNET_snr). Briefly, the input alignment file (.bam) provided together with the -f argument are converted to .sam to ease subsequent parsing. For each pair of reads in the .sam file only read 2 is considered; any read that contains deletions, insertions or soft clipping information in the CIGAR string is disregarded. After obtaining a SAM flag that identifies the nucleotide at the 3′ end position, the CIGAR string is turned into “1M”. Finally, the single nucleotide resolution .sam file is converted into a .bam file. 3′ end nucleotides are obtained with Python (version 2.7.12) using the command:

```
python get_SNR_bam_ignoreSoftClip.py -f mNET_Long_S5P_rep1_
unique.bam –s
    mNET_Long_S5P_rep1 -d ./
```

### 2.4. Identification and removal of PCR internal priming and duplication events

Occasionally, the primer for reverse transcription (RT) anneals to the RNA fragment rather than to the adapter (Fig. 2A). When such internal priming occurs, the genomic sequence adjacent to the aligned read is complementary to the primer (NTGG in Fig. 2A, right panel). In order to remove reads that result from internal priming events, we developed a script that identifies the presence of the 3′ OH adapter sequence (for example, TGGAATTCTCGGGTGCCAAGG) downstream of the aligned reads (Fig. 2B). The script, which was developed and tested with SAMtools v1.7 and bedtoolsv2.27.1-1-gb87c465, is available in (https://github.com/kennyrebelo/Filtering_InternalPriming). For single-end reads, the input .bam alignment file is converted into a .bed file. /; for paired-end reads, the second read from each pair is extracted and added to a new .bam file that will then be converted to a .bed file. Iterations through the .bed file reveal the genome coordinates downstream of the 3′ OH position (i.e., the last nucleotide of single reads or the first nucleotide of the read 2 in paired reads). The number of nucleotides analysed downstream of the 3′ OH position corresponds to the adapter length provided together with the –a parameter. The next step is extracting nucleotide sequence for each .bed entry (converting the .bed file into a .fasta file). All entries matching the adapter sequence are discarded. The remaining read IDs are saved into a .txt file and are used to extract internal priming-free reads from the original alignment file. Example run for paired reads:

```
python Filter_InternalPriming.py -f /alignments/mNET_Long_
S5P_rep1_unique_sorted.bam –s
    paired -a TGG.. -g /genomes/human/hg38/GRCh38.primary.
genome.fa
```

We empirically determined that restricting the filter to the first three nucleotides of the adapter sequence (TGG in Fig. 2) detected the highest number of internal priming events. This is probably because base pairing of only a few nucleotides is sufficient for priming.

Using a pool of adapters having randomized sequences (barcodes) helps reducing PCR overamplification bias, as reads that align to the same genomic position and contain identical barcodes are likely the result of PCR duplication events. Duplicate reads can be removed using BBmap/clumpify.sh [19].

### 2.5. Distinguishing Pol II density profiles from co-transcriptional splicing

Because the final (3′ OH end) nucleotide of a nascent RNA lies at the active site of the polymerase, NET-seq provides nucleotide resolution profiles of RNA Pol II along the genome (Fig. 3A). However, the mNET-seq technique additionally detects the 3′ OH end of RNAs that are not located at the polymerase active site but associate with the Pol II elongation complex and are therefore co-immunoprecipitated [7,20]. This includes the 3′ OH ends generated by co-transcriptional cleavage of splice sites (Fig. 3B) and the free 3′ OH ends of spliceosome snRNAs (Fig. 3C). NET-seq reads mapping precisely to the last nucleotide of spliced exons correspond to splicing intermediates that are formed by cleavage at the 5′ splice site after the first splicing reaction, and reads mapping to the last nucleotide of introns correspond to released intron

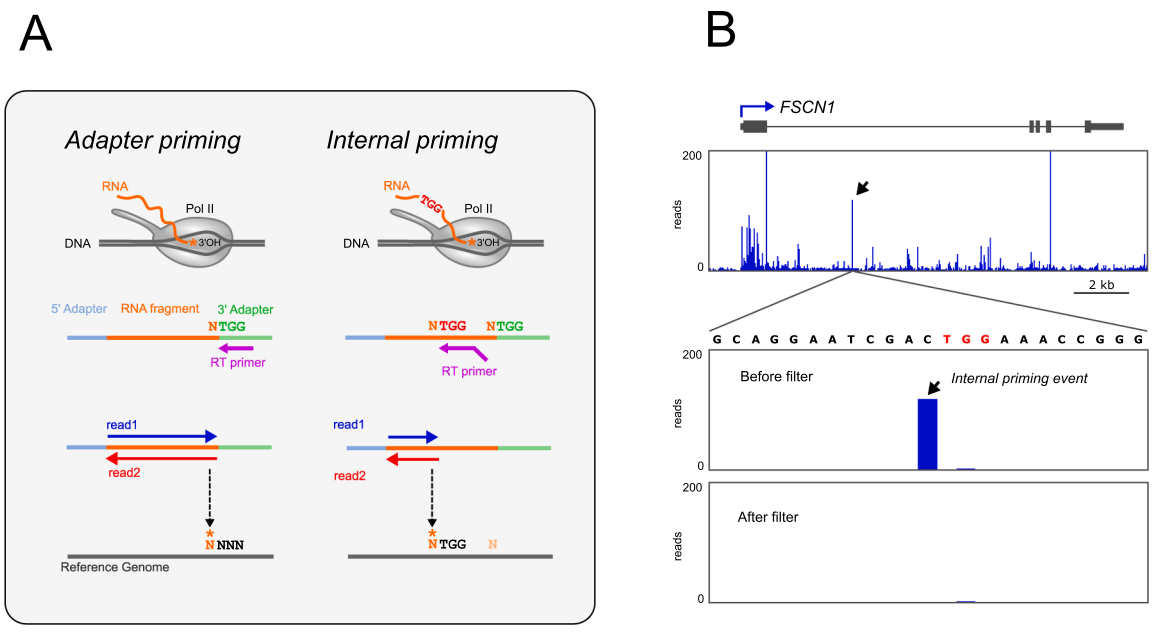



**Fig. 2.** Identification and removal of internal priming events. (A) Diagram depicts the expected base-pair complementarity between the RT primer and the adapter (left panel). Internal priming occurs when the RT primer hybridizes to the RNA sequence (right panel). (B) Visualization of mNET-seq profile along the *FSCN1* gene. The arrow denotes a spike resulting from internal priming. Data from HeLa cells long reads S5P mNET-seq replica1 [20].

lariats after completion of the splicing reaction (Fig. 3B). NET-seq reads mapping to the end of snRNA genes correspond to mature snRNAs engaged in co-transcriptional spliceosome assembly (Fig. 3C).

In order to determine RNA Pol II density profiles, only 3′ OH ends corresponding to the nucleotide at the active site of the polymerase are considered. To exclude signal from co-transcriptional splicing, reads that map to the very last nucleotide of introns and exons are discarded and the corresponding genomic positions are not considered. These reads can be removed using bedtools (version v2.27.1-1-gb87c465) together with SAMtools (version 1.7):

```
intersectBed -a mNET_Long_S5P_rep1_SNR.bam -b exons_lastNT.
bed -wa -v | samtools view - > mNET_Long_S5P_rep1_SNR_
noLastNT_temp.sam
    cat mNET_Long_S5P_rep1_header.sam mNET_Long_S5P_rep1_
SNR_noLastNT_temp.sam > mNET_Long_S5P_rep1_SNR_
noLastNT.sam
    samtools view -bS mNET_Long_S5P_rep1_SNR_noLastNT.
sam > mNET_Long_S5P_rep1_SNR_noLastNT.bam
    rm -f mNET_Long_S5P_rep1_SNR_noLastNT_temp.sam
    rm -f mNET_Long_S5P_rep1_SNR_noLastNT.sam
```

### 2.6. Selection of transcriptionally active genes

One approach to identify which genes are being transcribed in a particular cell type is to use RNA-seq data of polyadenylated RNA. Alternatively, mNET-seq read density over genes can be used to measure transcriptional activity. However, because many inactive genes maintain high levels of Pol II paused near the promoter, thus generating promoter-proximal reads, quantification of mNET-seq signal should be restricted to the gene body region. In order to identify transcriptionally active genes based on mNET-seq signal, we use a strategy adapted from GRO-seq analysis [21] that relies on read density in gene desert regions as background reference for absence of transcription. Very large intergenic regions (gene deserts) are divided into 500 kb windows, and read densities are calculated by dividing read counts in each window by the window length (in bp). Read counts per window are obtained with bedtools (version v2.27.1-1-gb87c465) using the command:

```
coverageBed -a intergenic_regions_500kb_Windows.bed –b
    Filter_IP/mNET_Long_S5P_rep1_noInternalPriming.bam
-counts >
    intergenic_regions_500kb_Windows_cov.bed
```

A density threshold is arbitrarily defined as the 90th percentile of the total read density (Fig. 4A). This threshold is then used to identify which annotated genes are transcriptionally active, based on mNET-seq signal (in RPKM) over the gene body (Fig. 4B).

### 2.7. Data visualization

Several visualization tools are available to depict mNET-seq profiles in specific genomic regions and with strand directionality. These include VING [22], IGV [23] and UCSC genome browser [24].

### 2.8. Metagene analysis

Metagene plots provide visual representations of the average mNET-seq signal at specific genomic regions. For example, to visualize Pol II density at the intron-exon boundary, we use deepTools [25] to integrate mNET-seq signal over the last 200 nucleotides of introns and the first 200 nucleotides of adjacent exons (Fig. 5A). By defining a window of 200 bp upstream and downstream of the intron-exon junction (i.e., the 3′ splice site), the analysis must be restricted to introns and exons longer than 200 nucleotides. Further biological constraints can be imposed on the genomic regions selected for metagene analysis, such as only intron-exon boundaries of transcriptionally active genes or only intron-exon boundaries of constitutively spliced exons (identified in RNA-seq poly(A) data). For comparisons between regions (for example, intron-exon boundaries of spliced versus non-spliced exons) or experimental samples (for example, intron-exon boundaries in control cells versus cells treated with a drug that inhibits splicing), normalizations should be implemented. For normalization, we divide the number of reads at each nucleotide by the total number of reads in the entire genomic region under analysis. These values are then used to calculate

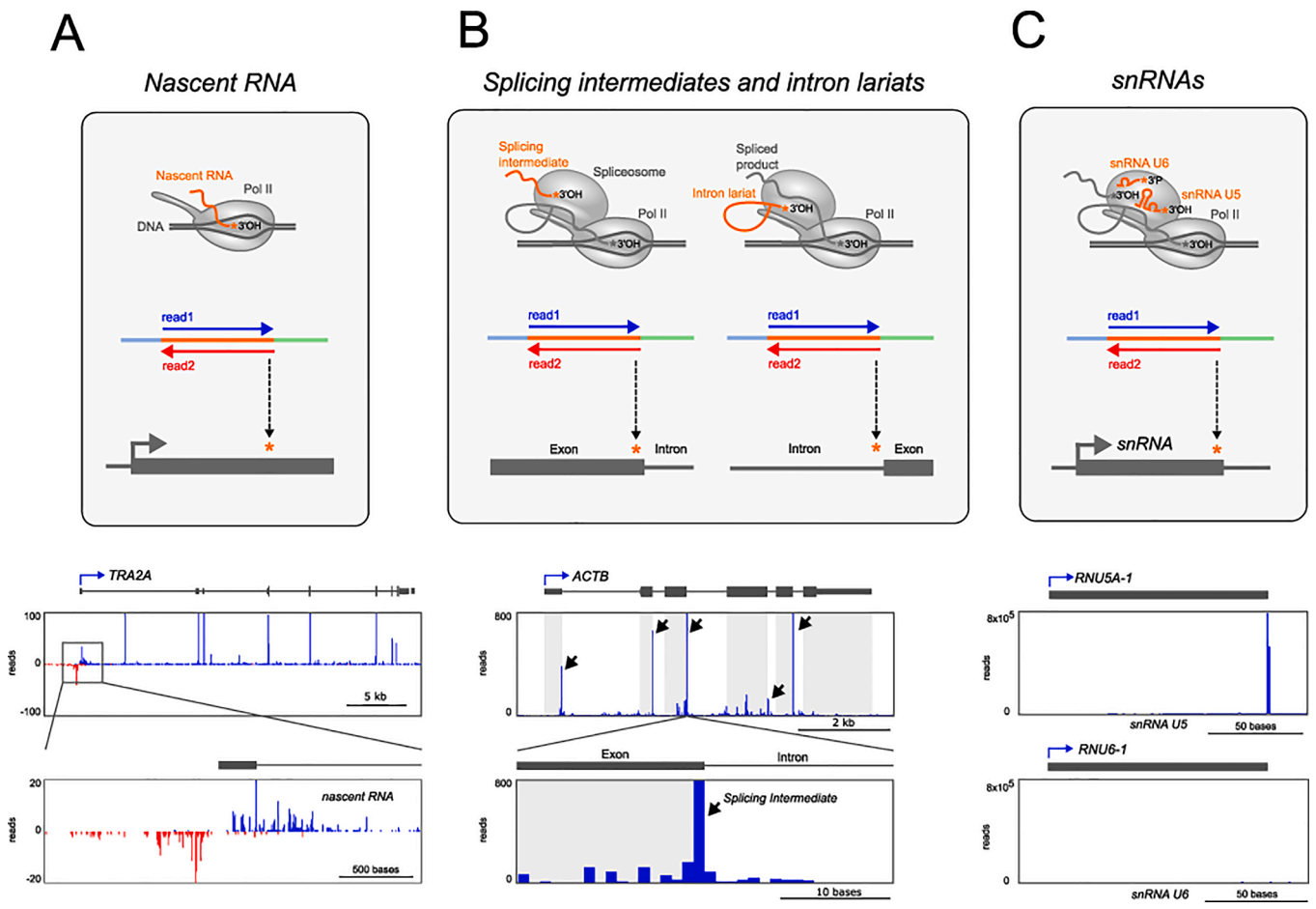



**Fig. 3.** Distinguishing mNET-seq signal from nascent RNA and co-transcriptional splicing. (A) A large fraction of mNET-seq signal corresponds to 3′ OH ends of nascent RNAs (orange asterisk, top panel). The bottom panel depicts nascent transcript profile along the *TRA2A* gene. Data from HeLa cells short reads S5P mNET-seq replica1 [20]. (B) A fraction of mNET-seq signal corresponds to 3′ OH ends of either splicing intermediates or excised intron lariats (orange asterisk, top panel). The bottom panel depicts nascent transcript profile along the *ACTB* gene using data from HeLa cells long reads S5P mNET-seq replica1 [20]. (C) An additional fraction of mNET-seq signal corresponds to 3′ OH ends of spliceosome snRNAs (orange asterisk, top panel). The bottom panel depicts nascent transcript profile along the *snRNA U5* gene. Note that in contrast to the *snRNA U5* profile, no accumulation of mNET-seq signal is detected at the end of *snRNA U6* gene, as expected since U6 snRNA contains a phosphate terminal group at the 3′ end. Data from HeLa cells long reads S5P mNET-seq replica1 [20]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

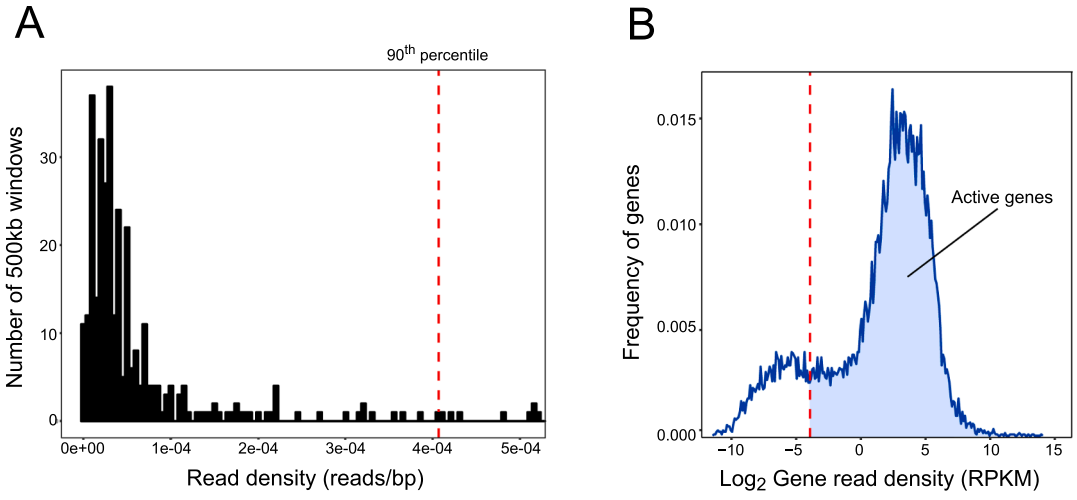



**Fig. 4.** Identification of transcriptionally active genes based on mNET-seq signal along the gene body. (A) Frequency distribution of read density in intergenic regions (gene deserts). A total of 724 windows (each 500 Kb in length) were defined in the RefSeq NCBI hg38 downloaded from UCSC table browser (accessed on the 3rd March 2019). The red dashed line represents the 90th percentile of read density in all regions analysed. (B) Frequency distribution of gene read density (RPKM) represented in Log2 scale. The 90th percentile of read density over gene deserts is set as threshold (red dashed line). A total of 13,426 genes are classified as transcriptionally active (blue area). Dataset from HeLa cells long reads S5P mNET-seq replica1 [20] (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

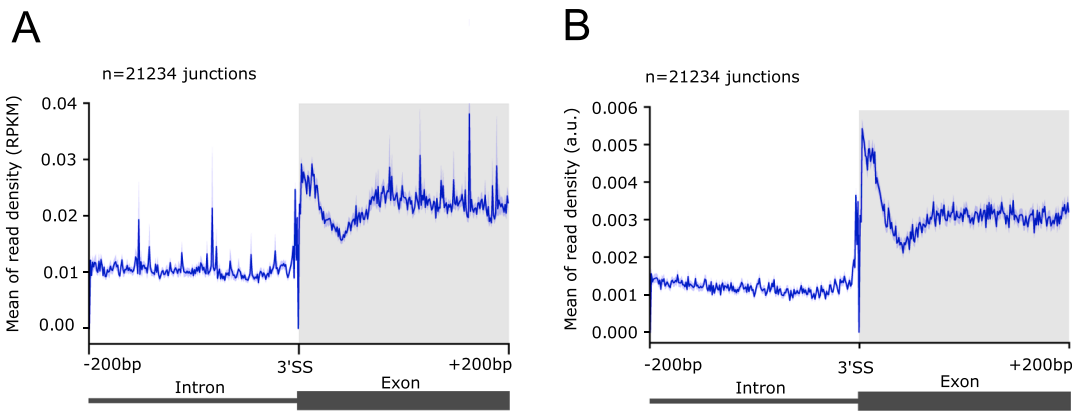



**Fig. 5.** Metagene analysis. (A) Metagene analysis for 21,234 intron-exon boundaries centred at the 3′ splice site (3′ ss). The mean of read density in RPKM was calculated for each nucleotide. (B) Read density was normalized and expressed in arbitrary units (a.u.). Dataset from HeLa cells long reads S5P mNET-seq replica1 [20]

the mean for each nucleotide, and the results are plotted in an arbitrary unit ranging from 0 to 1 (Fig. 5B).

### 2.9. Peak calling

Spikes in the density of 3′ ends at the active site of the polymerase are indicative of Pol II pausing. To identify Pol II pause positions along any given gene, Churchman and Weissman developed an algorithm that finds nucleotides where the NET-seq read density is at least three standard deviations above the mean in a local region [1]. As we systematically found significant accumulation of mNET-seq signal at the last nucleotide of spliced exons (Fig. 3B, bottom panel), we adapted this peak identifier strategy to quantify the prevalence of splicing intermediates and excised introns at genome-wide level [20]. The algorithm that we designed compares the number of reads mapping to the last nucleotide of exons and introns to the mean read density across the corresponding exon or intron. Only positions with coverage of at least 4 reads are considered. The script that is available in (https://github.com/kennyrebelo/NET_snrPeakFinder), was used with the following command:

```
python NET_snrPeakFinder.py -i mNET_Long_S5P_rep1_SNR.bam
-g gene_list.bed -s paired
```

Using this algorithm on mNET-seq datasets from HeLa cells, we found that approximately 90% of efficiently spliced exons had a 5′ splicing intermediate peak, whereas 3′ splice site peaks were rare [20]. Accumulation of mNET-seq signal corresponding to 5′ splicing intermediates and excised introns can be viewed as a proxy for co-transcriptional splicing kinetics, as the intensity of each peak depends on the lifetime of that particular RNA product. According to this view, mNET-seq reveals a significant time lag between the first and second splicing steps, whereas excised introns are rapidly degraded or dissociated from Pol II after completion of splicing.

## 3. Conclusions and perspectives

To date, the mNET-seq technique has been used in human [7,20,26], mouse [20] and plant cells [9]. In all cases, antibodies specific for Pol II with the CTD phosphorylated on serine 5 residues (S5P) immunoprecipitated abundant RNA fragments mapping precisely to the last nucleotide of exons, as expected for intermediates formed after the first splicing reaction. This suggests that the catalytically active spliceosome forms a tight complex with S5P Pol II and highlights the discovery potential of mNET-seq compared to previous techniques such as ChIP. Indeed, based largely on ChIP data, the established view was that the CTD was phosphorylated on serine 5 near the transcription start site, but shortly after transcription initiation and RNA capping these phosphates were removed [27]. According to mNET-seq data, serine 5 phosphorylation is not restricted to transcription initiation but is also present during elongation and co-transcriptional splicing. Noteworthy, a splicing-related accumulation of S5P Pol II along gene bodies was observed by ChIP in HeLa cells [28] and in yeast [29,30].

In addition to Pol II, other antibodies could in principle be used for mNET-seq. For example, mNET-seq analysis of complexes immunoprecipitated with antibodies to RNA processing factors could be useful for further studies on transcription-coupled pre-mRNA processing. Antibodies to Pol I and Pol III could also be used to determine the genomic distribution and nascent transcript profiles of these polymerases.

As an antibody-based technique, mNET-seq relies on the availability of antibodies that are specific and capable of precipitating the protein of interest. Problems related to antibody binding to off-target epitopes and differential accessibility of epitopes should always be considered, namely when investigating nascent transcripts associated with different phospho-isoforms of RNA Pol II. Another limitation of mNET-seq is treatment with MNase, which in our hands digests nascent RNA into 30–60 nucleotide-long fragments [20]. Developing new approaches to preserve longer stretches of nascent RNA associated with specific Pol II complexes is crucial for understanding how splicing kinetics is coordinated with transcription. Another challenge will be to adapt nascent RNA analysis to third generation sequencing technologies, which allow sequencing long molecules and avoid biases introduced by PCR amplification.

## Acknowledgements/Funding

We are grateful to Nick Proudfoot and Taka Nojima (University of Oxford) for sharing mNET-seq datasets and for insightful discussion.

M.C.-F. acknowledges funding from Fundação para a Ciência e Tecnologia (FCT)/ Ministério da Ciência, Tecnologia e Ensino Superior (MCTES) through Fundos do Orçamento de Estado (UID/BIM/50005/2019), and FCT/FEDER/POR Lisboa 2020, Programa Operacional Regional de Lisboa, PORTUGAL 2020 (LISBOA-01-0145-FEDER-016394).. R.G.M. is supported by FCT grants PTDC/BEX-BID/0395/2014, PTDC/BIA-BID/28441/2017, and UID/BIM/04773/2013 CBMR 1334.

## References

[1] L.S. Churchman, J.S. Weissman, Nascent transcript sequencing visualizes transcription at nucleotide resolution, Nature 469 (2011) 368–373, https://doi.org/10.

ARTICLE IN PRESS



1038/nature09652.
[2] L.S. Churchman, J.S. Weissman, Native Elongating Transcript Sequencing (NET-seq), Curr. Protoc. Mol. Biol. 98 (2012) 14.4.1–14.4.17, https://doi.org/10.1002/0471142727.mb0414s98.
[3] Y. Jin, U. Eser, K. Struhl, L.S. Churchman, The ground state and evolution of promoter region directionality, Cell 170 (2017) 889–898.e10, https://doi.org/10.1016/j.cell.2017.07.006.
[4] M.H. Larson, R.A. Mooney, J.M. Peters, T. Windgassen, D. Nayak, C.A. Gross, S.M. Block, W.J. Greenleaf, R. Landick, J.S. Weissman, A pause sequence enriched at translation start sites drives transcription dynamics in vivo, Science 344 (2014) 1042–1047, https://doi.org/10.1126/science.1251871.
[5] I.O. Vvedenskaya, H. Vahedian-Movahed, J.G. Bird, J.G. Knoblauch, S.R. Goldman, Y. Zhang, R.H. Ebright, B.E. Nickels, Interactions between RNA polymerase and the "core recognition element" counteract pausing, Science 344 (2014) 1285–1289, https://doi.org/10.1126/science.1253458.
[6] H. Kimura, Y. Tao, R.G. Roeder, P.R. Cook, Quantitation of RNA polymerase II and its transcription factors in an HeLa cell: little soluble holoenzyme but significant amounts of polymerases attached to the nuclear substructure, Mol. Cell. Biol. 19 (1999) 5383–5392, https://doi.org/10.1128/MCB.19.8.5383.
[7] T. Nojima, T. Gomes, A.R.F. Grosso, H. Kimura, M.J. Dye, S. Dhir, M. Carmo-Fonseca, N.J. Proudfoot, Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing, Cell. 161 (2015) 526–540, https://doi.org/10.1016/j.cell.2015.03.027.
[8] T. Nojima, T. Gomes, M. Carmo-Fonseca, N.J. Proudfoot, Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide, Nat. Protoc. 11 (2016) 413–428, https://doi.org/10.1038/nprot.2016.012.
[9] J. Zhu, M. Liu, X. Liu, Z. Dong, RNA polymerase II activity revealed by GRO-seq and pNET-seq in arabidopsis, Nat. Plants 4 (2018) 1112–1123, https://doi.org/10.1038/s41477-018-0280-0.
[10] S. Andrews, FASTQC. A quality control tool for high throughput sequence data. n.d. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (accessed April 15, 2019).
[11] R.M. Leggett, R.H. Ramirez-Gonzalez, B.J. Clavijo, D. Waite, R.P. Davey, Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics, Front. Genet. 4 (2013), https://doi.org/10.3389/fgene.2013.00288.
[12] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet.J. 17 (2011) 10–12, https://doi.org/10.14806/ej.17.1.200.
[13] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S.L. Salzberg, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions, Genome Biol. 14 (2013) R36, https://doi.org/10.1186/gb-2013-14-4-r36.
[14] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T.R. Gingeras, STAR: ultrafast universal RNA-seq aligner, Bioinformatics 29 (2013) 15–21, https://doi.org/10.1093/bioinformatics/bts635.
[15] D. Kim, B. Langmead, S.L. Salzberg, HISAT: a fast spliced aligner with low memory requirements, Nat. Methods 12 (2015) 357–360, https://doi.org/10.1038/nmeth.3317.
[16] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M.W. Szcześniak, D.J. Gaffney, L.L. Elo, X. Zhang, A. Mortazavi, A survey of best practices for RNA-seq data analysis, Genome Biol. 17 (2016) 13, https://doi.org/10.1186/s13059-016-0881-8.
[17] L. Wang, S. Wang, W. Li, RSeQC: quality control of RNA-seq experiments, Bioinformatics 28 (2012) 2184–2185, https://doi.org/10.1093/bioinformatics/bts356.
[18] C.W. Fuller, L.R. Middendorf, S.A. Benner, G.M. Church, T. Harris, X. Huang, S.B. Jovanovich, J.R. Nelson, J.A. Schloss, D.C. Schwartz, D.V. Vezenov, The challenges of sequencing by synthesis, Nat. Biotechnol. 27 (2009) 1013–1023, https://doi.org/10.1038/nbt.1585.
[19] Bushnell, Brian, BBMap: A Fast, Accurate, Splice-Aware Aligner, 2014. https://www.osti.gov/servlets/purl/1241166.
[20] T. Nojima, K. Rebelo, T. Gomes, A.R. Grosso, N.J. Proudfoot, M. Carmo-Fonseca, RNA polymerase II phosphorylated on CTD serine 5 interacts with the spliceosome during co-transcriptional splicing, Mol. Cell 72 (2018) 369–379.e4, https://doi.org/10.1016/j.molcel.2018.09.004.
[21] L.J. Core, J.J. Waterfall, J.T. Lis, Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters, Science 322 (2008) 1845–1848, https://doi.org/10.1126/science.1162228.
[22] M. Descrimes, Y.B. Zouari, M. Wery, R. Legendre, D. Gautheret, A. Morillon, VING: a software for visualization of deep sequencing signals, BMC Res. Notes 8 (2015) 419, https://doi.org/10.1186/s13104-015-1404-5.
[23] J.T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E.S. Lander, G. Getz, J.P. Mesirov, Integrative genomics viewer, Nat. Biotechnol. 29 (2011) 24–26, https://doi.org/10.1038/nbt.1754.
[24] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler, The human genome browser at UCSC, Genome Res. 12 (2002) 996–1006, https://doi.org/10.1101/gr.229102.
[25] F. Ramírez, D.P. Ryan, B. Grüning, V. Bhardwaj, F. Kilpert, A.S. Richter, S. Heyne, F. Dündar, T. Manke, deepTools2: a next generation web server for deep-sequencing data analysis, Nucleic Acids Res. 44 (2016) W160–W165, https://doi.org/10.1093/nar/gkw257.
[26] M. Schlackow, T. Nojima, T. Gomes, A. Dhir, M. Carmo-Fonseca, N.J. Proudfoot, Distinctive patterns of transcription and RNA processing for human lincRNAs, Mol. Cell 65 (2017) 25–38, https://doi.org/10.1016/j.molcel.2016.11.029.
[27] D. Eick, M. Geyer, The RNA polymerase II carboxy-terminal domain (CTD) code, Chem. Rev. 113 (2013) 8456–8490, https://doi.org/10.1021/cr400071f.
[28] E. Batsché, M. Yaniv, C. Muchardt, The human SWI/SNF subunit Brm is a regulator of alternative splicing, Nat. Struct. Mol. Biol. 13 (2006) 22–29, https://doi.org/10.1038/nsmb1030.
[29] R.D. Alexander, S.A. Innocente, J.D. Barrass, J.D. Beggs, Splicing-dependent RNA polymerase pausing in yeast, Mol. Cell 40 (2010) 582–593, https://doi.org/10.1016/j.molcel.2010.11.005.
[30] K.T. Chathoth, J.D. Barrass, S. Webb, J.D. Beggs, A splicing-dependent transcriptional checkpoint associated with prespliceosome formation, Mol. Cell 53 (2014) 779–790, https://doi.org/10.1016/j.molcel.2014.01.017.