Biocomputational tools for transcriptome-wide analyses of RNA-binding proteins

J. Antonio Paternina Osorio

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Normale Supérieure

Biocomputational tools for transcriptome-wide analyses of RNA-binding proteins

Soutenue par

J. Antonio Paternina Osorio

Le 6 Octobre 2020

École doctorale n°

École Doctorale Complexité du vivant ED515

Spécialité

Biologie computationnelle

ENS

Composition du jury :

M. Hervé LE HIR Équipe Expression des ARN messagers eucaryotes, École normale supérieure	*Directeur de thèse*
M. Auguste GENOVESIO Équipe Bio-imagerie computationnelle et bioinformatique, École normale supérieure	*Directeur de thèse*
Mme Mihaela ZAVOLAN Key regulators of gene expression and cell identity, Biozentrum, Université de Bâle	*Rapporteur*
M. Pierre NICOLAS Unité MaIAGE, Institut national de recherche pour l'agriculture, l'alimentation et l'environnement	*Rapporteur*
M. Antonin MORILLON Équipe ARN non-codant, epigénétique et fluidité du génome, Insitut Curie	*Président*
Mme Silvia BOTTINI Medical Data Laboratory, Université Côte d'Azur	*Examinateur*
Mme Morgane THOMAS-CHOLLIER Équipe Biologie computationnelle de systèmes, École normale supérieure	*Membre invité*

Acknowledgments

Spanish poet Antonio Machado coined the verse:

"*Caminante no hay camino, se hace camino al andar* (Traveler, there is no road; you make your own path as you walk)".

Although it is commonly used as a metaphor for life-changing voyages, I consider it an appropriate description of the PhD pursuit.

Contrary to popular belief, we do not walk the PhD path on our own—and I wish I could have realized this earlier. I would like to thank my supervisors Hervé Le Hir and Auguste Genovesio, first for your vote of confidence, and foremost for your guidance, patience, and support during this project. The working dynamic where we could openly discuss and decide how to move forward was essential for the advancement of this project.

I would like to thank the members of my thesis committee Frédéric Devaux, Claude Thermes, and Morgane Thomas-Chollier for your guidance through this project, both at the scientific and the personal levels. Our fruitful conversations steered the project to this direction. Without your input, these pages would not look the same.

During this adventure, I have crossed paths with wonderful people and friends. As I had the chance to be part of two teams, my odds of meeting brilliant people were double.

To the members of team Le Hir, thank you for treating me as part of the lab despite not being a *true wet* biologist. To Quentin Alasseur, thank you not only for generating the data that allowed this work to exist, but for being there when we were both frustrated by CLIP. To Lucía Morillo, I know my scripts will be left in good hands. Special thanks to the former members that paved the way for this work: Rémi Hocq and Leïla Bastianelli. To the members of team Genovesio, thank you for creating the best everyday environment. I will cherish the time we spent together inside and outside the lab. You made this challenge easier to carry out, and for this I am grateful. To Nikita Menezes and Kasia Radomska, thank you for being the ones I could confide in and look for unconditional support.

Thank you to all the kind people in the Institute that shared part of this adventure with me. To the IT service and Mathieu Bahin that made this project

possible at the technical level. The HR service that helped me navigate the administrative labyrinth of French bureaucracy. To SPIBENS and my friends from other labs that expanded my social horizons. Thank you to all the people that helped me prepare the presentation that ultimately funded my PhD, and thank you to the École Doctorale Complexité du Vivant for supporting this project.

I will allow myself to write a few words in Spanish to thank my family and friends. A mi madre Dafnis y a Vicenç, gracias por su apoyo incondicional en los momentos más cruciales durante los últimos años. Mi vida en París sería infinitamente más complicada sin ustedes. Nadie podrá negar que han contribuido de manera esencial a este trabajo. A mis amigos en el extranjero, gracias por estar allí y brindarme un segundo hogar cada vez que nos reencontramos. A Nicolás, que estuvo presente sin estarlo durante los últimos meses, ofreciéndome su apoyo y cariño a pesar de la distancia. And to Justin, thank you for your sound advice and being a constant source of joy, even through the darkest times.

All of you can claim this work, for it would not exist without you.

Abstract

Post-transcriptional Gene Expression Regulation is a complex network that involves RNA-binding proteins and non-coding RNAs to orchestrate the complex life of mRNAs. In metazoans, the Exon Junction Complex (EJC) is a multi-protein complex deposited onto mRNAs exon junctions during splicing. The EJC interacts with numerous factors and is important for coupling pre-mRNA splicing with mRNA nuclear export, localization, translation, and decay. Despite its central role in gene expression and in organism development, the comprehensive map of EJC binding sites is lacking. Crosslinking and immunoprecipitation coupled with high-throughput sequencing (CLIP-seq) aims to identify transcriptome-wide RNA-protein interactions in vivo. Yet, current trends in CLIP-seq data analysis gravitate towards painting a global landscape rather than characterizing individual binding sites. However, we observed that current peak callers applied to EJC CLIP data yield results with limited reproducibility and sensibility.

During my PhD, we developed a dedicated strategy to detect EJC signal enrichment at the exon level. By aggregating data from several replicates, we built a list of robust genes with reproducible EJC loading rate. Within robust genes, we assigned a robustness score to each exon according to frequency of detection across replicates. We found that the exon robustness score was correlated to the thymidine (T) content of EJC binding sites. Assuming this was due to cross-linking chemistry, we corrected the score for the T content and found exons with either high or low detection rates. The last suggests that EJC loading is not homogeneous along a transcript, but rather differential. Thus, we established an unprecedented binding site map of the EJC in living cells validated by statistical tools. Crossing this map with other information showed that EJC loading is independent of transcript expression levels or known gene functional annotations. Although the scope of this work does not include possible explanations for this differential loading, it presents a first reproducible and specific data analysis pipeline to detect EJC-loaded exons.

Altogether, our contribution is twofold. First, we proposed a robust way to detect EJC signal enrichment at the exon level and demonstrated quantitatively that our approach is more reproducible and more sensitive compared to conventional tools. Second, we proved that the EJC can be present on some, and absent on other exons of the same transcript suggesting that EJC loading is a regulated process following a code that remains to be discovered.

Résumé

La régulation post-transcriptionnelle de l'expression des gènes est un réseau d'interactions impliquant de nombreuses protéines de liaison à l'ARN et des ARN non-codants afin d'orchestrer la vie complexe des ARN messagers (ARNm). Chez les métazoaires, le complexe EJC (*Exon Junction Complex*) est un complexe multiprotéique déposé sur la jonction exonique des ARNm pendant l'épissage. L'EJC interagit avec de nombreux facteurs et est important pour le couplage fonctionnel entre l'épissage et l'export du noyau, la localisation, la traduction et la dégradation des ARNm. Malgré son rôle central dans la régulation génique et le développement de l'organisme, aucune carte exhaustive des sites de liaison de l'EJC n'a encore été établie. La méthode de CLIP (Cross-Linking and Immunoprécipitation) associée au séquençage à haut-débit (CLIP-seq) permet d'identifier les sites de liaison protéine à l'ARN in vivo. Cependant, les analyses des données de CLIP-seq ont permettent aujourd'hui d'obtenir une vue globale plutôt qu'une caractérisation individuelle des sites de liaison d'une protéine. En effet, les détecteurs de pics conventionnels appliqués aux données de CLIP de l'EJC produisent des résultats dont la reproductibilité et la sensibilité sont limitées.

Durant ma thèse, nous avons développé une stratégie dédiée à la détection du signal de l'EJC au niveau exonique. En agrégeant les informations de différents réplicas, nous avons généré une liste de gènes reproductibles. Au sein de ces gènes, nous avons trouvé une forte corrélation entre la robustesse de détection des exons et le contenu en thymidine (T) au niveau des sites de liaison. Posant l'hypothèse que ceci est un effet du photopontage, nous avons corrigé le score de robustesse par le contenu en T et avons ainsi clairement montré que l'EJC est déposé sur certains exons et pas sur d'autres. Par conséquent, le complexe EJC est déposé de manière différentielle le long d'un même transcrit. Nous avons ainsi établi une carte des sites de liaisons de l'EJC sans précédent. L'intégration de données supplémentaires a montré que le dépôt de l'EJC est indépendant de l'abondance du transcrit et n'est pas expliqué par des annotations fonctionnelles connues du gène. Bien que ce travail n'a pas permis à ce stade d'identifier les raisons de ce dépôt différentiel, nous présentons une première méthode d'analyse spécifique et reproductible des exons liés à un EJC par CLIP-seq.

Les deux contributions principales de ce travail sont donc les suivantes. Premièrement, nous proposons une méthode robuste pour détecter l'enrichissement du signal de l'EJC à l'échelle de l'exon, en démontrant quantitativement que celle-ci est plus reproductible et plus sensible que les solutions offertes par les outils actuels.

Deuxièmement, nous prouvons que, au sein d'un même transcrit, l'EJC peut être présent sur des exons, et absent d'autres, suggérant que le dépôt de l'EJC est un processus régulé suivant un code qui reste à découvrir.

Summary

PART I

STATE OF THE ART

Chapter 1

Introduction

1.1 Context

Post-transcriptional gene regulation (PTGR) controls gene expression at the RNA level: processing (splicing, base chemical modifications, and cleavage), nuclear export, sub-cellular localization, translation, and decay. The cell orchestrates a high complexity network of over 1500 RNA-binding proteins (RBPs), which are proteins that bind RNA directly or indirectly. These interactions are fundamental for cell homeostasis, differentiation, and organism development. A failure in the PTGR network results in pathological phenotypes. To this day, however, the cellular functions of many RBPs have not been characterized. Moreover, a precise map of most RBP binding sites and the dynamics of their interaction with RNA remain to be elucidated.

The Exon Junction Complex (EJC) is a multi-RBP complex and a central actor in the PTGR network. Assembled onto messenger RNA (mRNA) by the splicesome, it escorts mature transcripts to the cytoplasm and exerts key roles in localization, translation and degradation. The spliceosome-dependent assembly positions the EJC near the exon junction, estimated to be around the 24th nucleotide upstream of the exon junction. Its structural features allow the complex to stably bind RNA in a sequence independent manner. It thus serves as an anchor for interactions with other RBPs, which grants it the versatility to intervene in all stages of mRNA life. Mutations of EJC components result in morphological and neurological disorders, highlighting its importance during embryonic development and neurogenesis. Despite its central role in gene expression regulation, many questions about the functional impact of EJC binding remain open.

1.2 Problematic

The versatility of the EJC hinders the study of its multiple functions. The current knowledge about its assembly and roles have been elucidated with low throughput molecular biology techniques. This limits the study of the impact of EJC binding on the regulation of specific genes. These approaches ignore its potential role in different

stages of the life of different transcripts. For instance, it is not known whether the presence of the EJC on all the exons of a transcript is necessary for proper gene expression. Conversely, in case the EJC assembly is targeted towards specific exons, within specific genes, the factors driving its deposition remain unknown. An exhaustive transcriptome-wide map of EJC binding sites is extremely valuable to dissect the rules that dictate its deposition, its functional impact, and, potentially, its regulation in different cellular contexts.

crosslinking and immunoprecipitation (CLIP) coupled with high-throughput sequencing aims at discovering the binding site landscape of an RBP at the transcriptome scale. This method consists in inducing protein-RNA covalent bonds in vivo with UV light, which allows to perform immunoprecipitation (IP) under stringent conditions. During IP, antibodies against the RBP of interest purify the protein along with the bound RNA fragments. Sequencing and mapping these fragments to the transcriptome result in local signal enrichment, which corresponds to the binding sites of the RBP. Yet, crosslinking and IP efficiency majorly limit obtaining RNA fragment libraries that represent all specific RBP binding sites. Moreover, the analysis of CLIP-seq data is a challenging task in terms of sensitivity, specificity, and reproducibility.

Over the years, several attempts to establish the EJC binding landscape have been performed. The results suggest that the EJC is not loaded homogeneously on the exons of a transcript, but rather on specific locations. However, these studies have generally bypassed the reproducibility of binding site detection. As CLIP protocols increasingly gained resolution, the reproducibility of individual binding sites became harder to assess. Thus, despite the experimental advances, a reproducible, high-resolution, transcriptome-wide map of the EJC binding sites has not yet been established.

1.3 Plan

In this manuscript, we will dedicate Chapter 2 — The Exon Junction Complex: a cornerstone in gene regulation — to introduce the importance of RBPs and the state of the art on the current knowledge about the EJC.

In Chapter 3 — Mapping RNA-binding proteins with CLIP —, we will summarize the principle of the CLIP protocol, as well the current data analysis strategies. We will then overview the insights gained with EJC CLIP data and their limitations. Finally, we will state the main objectives of this work.

In Chapter 4 — Acquiring single-nucleotide EJC CLIP data —, we will present a high-resolution CLIP protocol that allows to assign EJC binding sites with higher precision. Next, we will present a data pre-processing strategy that accelerated the production of EJC CLIP libraries.

In Chapter 5 — Peak calling with high resolution: a tale of reproducibility —, we will present the results of currently available binding site detection strategies on

our data. We will highlight its limitations in reproducibility, which are a general phenomenon for CLIP data.

In Chapter 6 — Reproducibility first: introducing an EJC-tailored pipeline —, we will present the strategy developed to overcome these limitations. Next, the insights we gained thanks to this approach are presented in Chapter 7 — Studying the behavior of the EJC with CLIP data.

We dedicate Chapter 8 — Discussion — to discuss these results and present some perspective work.

Chapter 9 — Annexes — contains the article corresponding to the CLIP protocol presented in Chapter 4, as well as an additional contribution to a separate project.

Chapter 2

The Exon Junction Complex: a cornerstone in gene regulation

2.1 Beyond DNA: regulating gene expression at the RNA level

2.1.1 The complexity of the transcriptome

The central dogma of biology relies on three molecular pillars: DNA, as genetic information storage; RNA, as the information transmitter, and protein as the machinery executing most of the functions for the life of the cell. Although the process seems straightforward, each step of gene expression involves hundreds of different factors that orchestrate and fine-tune the whole process. Rather than a 3-step linear recipe, gene expression is the result of an intricate network of strongly regulated molecular interactions (M. J. Moore and Proudfoot 2009).

In eukaryotes, protein coding genes are divided into distinct alternating segments: exons and introns. As they are transcribed into a premature messenger RNA (pre-mRNA) in the nucleus, the nascent molecule undergoes splicing, which consists in excising the intron segments and linking the exons together. Additionally, a methylated guanosine triphosphate is added to the 5'-end to form the 5'-cap, and a poly-adenosine (poly-A) tail added to the 3'-end. Together, these modifications constitute the mature form of the mRNA. Mature transcripts are then exported to the cytoplasm, where they are translated into proteins by the ribosome, and ultimately degraded.

The number of human protein coding genes is estimated to be around 20000 by GENCODE annotation standards (Frankish et al. 2019). A typical human gene is made of 8 exons, but the number of exons per gene ranges from 1 to 149. Thus, most pre-mRNAs experience multiple splicing events during maturation. Moreover, alternative splicing events include or exclude particular exons, which generates several transcript *isoforms* for a single gene. This ultimately increases protein diversity and function without the need to increase the number of genes (Keren, Lev-Maor, and Ast 2010). Transcriptome-wide data has revealed that around 95% of human multi-exon genes undergo alternative splicing (Pan et al. 2008), and has shown that isoform abundance is cell-type dependent. Alternative splicing is only one example that illustrates the complexity of the transcriptome. Contrary to DNA, several RNA species exist in different degrees of abundance. Only 2% of the total RNA in a cell correspond to mRNA, which in turn presents various levels of abundance across different genes and the isoforms of the same gene. The processes that shape and determine these complex transcriptome dynamics are collectively known as Post-transcriptional Gene Expression Regulation (PTGR). The main effectors of PTGR are RNA-binding proteins (RBPs), which interact with transcripts and with each other to ensure proper gene expression.

2.1.2 The world of RNA-binding proteins

mRNA molecules do not exist as naked chains of ribonucleotides in the cell. The molecular interaction between RBPs and transcripts form messenger ribonucleoparticles (mRNPs). The RBP composition determines the life of a transcript from

transcription, to translation, until degradation, through the interaction of multiple factors along the way (Glisovic et al. 2008). Remarkably, it is estimated that over 1500 human genes encode RBPs, which corresponds to approximately 8% of protein coding genes (Gerstberger, Hafner, and Tuschl 2014). This underpins, on one side, the importance of PTGR for the cell, and on the other, its degree of complexity.

The complexity of RBPs does not rely solely on their number, but on their diversity. A census in human protein-coding genes revealed around 600 structurally different RNA-binding domains (RBDs). Grouping RBP genes by RBD families shows that most classes are represented by one or two members (Gerstberger, Hafner, and Tuschl 2014). The most represented RBDs are found among messenger RBPs (mRBPs), of which 60% contain either an RNA recognition motif (RRM), a K homology (KH) domain, a DEAD motif, a double-stranded RNA-binding motif (DSRM) or a zinc-finger domain. Moreover, several mRBPs have multiple occurrences of the same or different RBDs, RBDs combined with RNA-unrelated domains, or a single occurrence of an RBD. This grants mRBPs a *modular* design, which is thought to contribute to new RNA target adaptation in an evolutionary context.

The structural diversity of mRBPs is echoed by several types of binding modalities. RRM and KH domains recognize RNA motifs with varying degrees of specificity, according to the structural context of the protein (Lunde, C. Moore, and Varani 2007). For instance, splicing factor PTB recognizes CU motifs of the polypyrimidine track by directly interacting with at least two specific nucleotides, while the antagonist factor U2AF65 has higher affinity for U-rich regions (Cléry, Blatter, and Allain 2008). On the other hand, some proteins require low specificity to exert their functions (Hentze et al. 2018), such as mRNA translation or degradation. This is the case of the eukaryotic translation initiation factor 4F complex (eIFE4F), composed of mRBPs that bind mRNAs independently of sequence, ensuring the translation of virtually all transcripts (Prévôt, Darlix, and Ohlmann 2003). Intrinsically disordered regions (IDR), play an important role in RNA binding as well. Their specificity ranges from co-folding with specific RNA sequences to a broad-spectrum of targets. For instance, SR proteins contain arginine and serine repeat regions with no globular structure that can interact with mRNA, but their recognition of degenerate RNA motifs suggests low sequence specificity (Järvelin et al. 2016). Additionally, novel mRNA-binding modalities have been attributed to proteins lacking conventional RBDs, such as metabolic enzymes that regulate gene expression according to the metabolic state of the cell (Hentze et al. 2018). Thus, the diversity in binding modalities, with varying degrees of specificity, goes hand in hand with the multiple functions of RBPs.

2.1.3 A vast orchestra of RBPs plays the PTGR symphony

These functions are crucial for cell homeostasis. The importance of RBPs was pinpointed by a large-scale census that revealed the high abundance of RBP-coding transcripts in various tissues (Gerstberger, Hafner, and Tuschl 2014). RBP genes make up to 20% of the expressed transcriptome, while transcription factors add up to 3%, despite comprising a similar number of genes. Although a high proportion of transcribed RBPs correspond to ribosomal proteins (12-13%), a significant fraction of

expressed genes corresponds to mRBPS (4-5%). Moreover, only 6% of all RBPs show tissue specificity. This proportion changes dramatically when considering separately the different types of RBP. While most ribosomal proteins and core components of the splicing machinery, and 68% of mRBPs are ubiquitous, several mRBPs portray some tissue specificity. These tissue-specific RBPs are restricted to a handful of tissues: brain, muscle, bone marrow, liver and adult testis. This highlights the essential cellular role of RBPs, where they constitute the backbone of every step of gene expression regulation.

Splicing and alternative splicing

The splicing pf pre-mRNA is catalyzed by the multi-RNP machinery of the spliceosome. First, the bond at the 5'-splice site (5'-SS) is cleaved through a nucleophilic attack of the first intron nucleotide by a distal adenosine within the branch point sequence (BPS) in the intron, producing a free exon in 5' and an intron lariat. Next, the first nucleotide of the 3' exon at the 3'-SS is attacked by the free 5' exon, resulting in the concatenation of both exons and the excision of the intron lariat (Wahl, Will, and Lührmann 2009, see Fig. 2.1a).

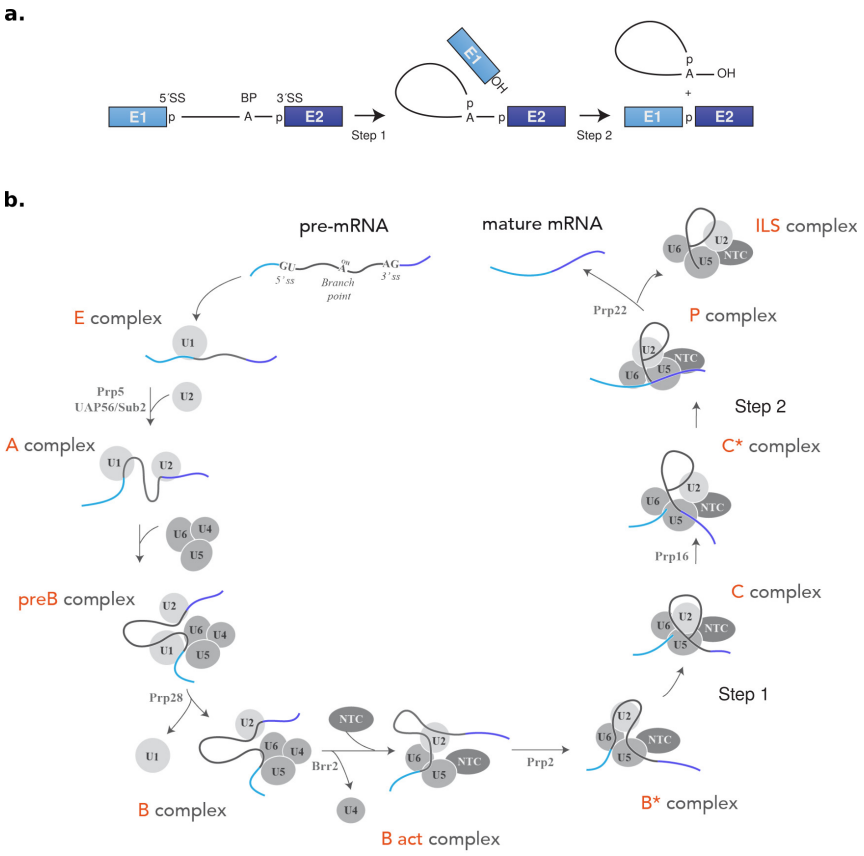

Figure 2.1: **a.** Representation of the two main splicing reactions: cleavage of the 5'-SS by the BPS adenosine, and cleavage of the 3'-SS by the 5'-end of exon 1; E1: 5'-exon (exon 1); E2: 3'-exon (exon2). **b.** A more detailed representation of the splicing reaction with the corresponding spliceosome conformation. Adapted from: Will and Lührmann 2011

At every step of the reaction, the spliceosome adopts a specific conformation or state. We thus distinguish the complexes E, A, pre-B, B, B^{act}, B*, C, C*, P, and ILS

(Fig. 2.1b). The constitutive components of the spliceosome are composed of small nuclear RNA (snRNA) and snRBP that interact to form snRNP. The main snRNPs participating in splicing are U1, U2, U4/U6, and U5. First, the E (*Early*) spliceosome is formed by U1-binding of the 5'-SS through base-pair interactions, while the 65 and 35 subunits of splicing factor U2AF (U2AF35 and U2AF65, respectively) bind the BPS and the polypyrimidine track through the interaction with the protein SF1/BBP. Next, the A (*Activated*) spliceosome is formed as U2 is recruited by U2AF65/35 and binds the BPS, displacing the SF1/BBP protein. A pre-assembled U4/U6-U5 snRNP binds the pre-mRNA to form an catalytically inactive form of the B complex. To reach a catalytically active form, U1 and U4 interactions are destabilized, and the NineTeen Complex (NTC) is recruited, forming the B* complex. This form catalyzes the first splicing reaction and forms the C complex. The spliceosome undergoes further RNP rearrangements to form the catalytically active C* complex and perform the second splicing reaction. This forms the P (*Post-catalytical*) complex, which contains the spliced junction and the intron lariat. Finally, the spliceosome is released from the spliced mRNA as the Intron Lariat Spliceosome (ILS), whose components will be recycled for further splicing reactions (Wahl, Will, and Lührmann 2009).

The protein-protein and protein-RNA interactions in each state are necessary for the proper splicing of each exon junction of a transcript, because a) the spliceosome needs to be re-assembled for each splicing event, and b) there is no evidence that suggests the existence a complete pre-assembled spliceosome. However, the consensus sequence of 5'-SS, 3'-SS and BPS is extremely short, the elements around them are poorly conserved, and the snRNA-preRNA interactions are rather weak (Wahl, Will, and Lührmann 2009). In addition, introns are on average 10 times longer than exons (G. Singh, Pratt, et al. 2015). Thus, the spliceosome needs to specifically distinguish these sites from a multitude of similar sequences to catalyze splicing with high fidelity. This task is accomplished through a step-wise compositional switch that involves core small nucleolar RNPs (snRNP) and a variety of mRBPs that aid in exon and intron definition.

Exon definition results form the interplay between sequence elements and RBPs. Several *cis*-acting elements are known to influence splicing, and are classified depending on their location (exonic or intronic) and their effect (enhancers and silencers). Thus, we distinguish exonic splicing enhancers (ESE) and silencers (ESS), and intronic splicing enhancers (ISE) and silencers (ISS) (Y. Wang et al. 2012). These elements are bound by different splicing factors, that either compete or act in synergy. For instance, SR proteins recognize ESE elements via their RRM and disordered arginine-serine regions, and recruit with splice site recognition components: U1 at the 5'-SS, and U2AF65 at the 3'-SS (Graveley, Hertel, and Maniatis 1998; Long and Caceres 2009). Conversely, members of the heterogenous nuclear RNP (hnRNP) family promote exon exclusion by interacting with ISS elements. However, there is evidence that suggests that the relative position of the binding site determines whether hnRNP binding suppresses or enhances the inclusion of an upstream exon (Dvinge 2018). hnRNPA/B form oligomers on the binding sites of SR proteins, resulting in exon exclusion (Fu and Ares 2014). hnRNP I, or PTB, has affinity for the polypyrimidine sequences recognized by U2AF65, blocking 3'-SS recognition directly. Thus, the balance between enhancers and silencers, ultimately defines the recognition of splice

sites, and thus the inclusion of specific exons. It is unclear whether the composition of the spliceosome varies between different junctions.

RNA editing

RNA editing consists in the conversion of single nucleotide bases of a particular transcript. The most prevalent form of RNA editing is the deamination of A (adenosine) to I (inosine, Glisovic et al. 2008). The main RBPs that catalyze RNA editing are members of the ADAR (adenosine deaminases acting on RNA) family (Valente and Nishikura 2005). These RBPs recognize imperfect double strand loops in pre-mRNA and catalyze the A-to-I conversion. Although most editing events have been found in non-coding regions of transcripts, A-to-I conversion can impact the coding sequence and the function of the resulting protein. Moreover, ADAR1 interacts with HuR (human antigen R), an RBP that binds AU-rich elements (AREs) and increases mRNA stability. Recently, RNA-mediated interactions between ADAR1 and DROSHA, ILF2 and ILF3 have been found to influence editing and stability of miRNA (Quinones-Valdez et al. 2019). Threfore, RBPs can regulate transcript stability through RNA editing and protein-protein interactions.

Poly-adenylation

Poly-adenylation (PA) is a crucial step in mRNA maturation, and has an impact in nuclear export, translation efficiency, and stability of virtually all eukaryotic mRNAs (Dassi 2017). RBPs of the CPSF complex recognize the highly conserved AAUAAA motif, and together with the nuclear poly-A binding protein 1 (PABN1) activate the poly-A polymerase and the synthesis of around 200 nucleotides. It is estimated that 50% of human genes contain APA (alternative poly-adenylation) sites. This can impact gene expression not only by generating protein isoforms (if the APA event occurs in the open reading frame), but also by varying the length of the 3'-UTR and thus the effect of other RBPs (Fatscher, Boehm, and Gehring 2015). The choice of PA site can be influenced by the abundance of PA factors. For instance, overexpression of Ctf64 causes the usage of proximal PA sites over distal sites. Moreover, splicing factors such as Nova2, PTB, hnRNP H, U2AF65, and SRm160, impact PA site definition by either recruiting or competing with cleavage factors. Thus, RBPs effect mRNA regulation through APA.

Nuclear export

Pre-mRNA maturation produces 5'-capped, spliced, polyadenylated, and possibly base-edited transcript dressed with multiple RBPs along three distinct domains: 5', internal, and 3' (G. Singh, Pratt, et al. 2015). The mRNP that results from mRNA processing is then ready to continue its journey in gene expression. The first necessary step is to export mature mRNPs from the nucleus into the cytoplasm.

The best characterized mRNA nuclear export pathway involves the heterodimer

NXF1-NXT1. These factors directly bind phenylalanine-glycine rich nucleoporines (FG-Nups) at the NPC (nuclear pore complex), docking the mRNP and facilitating export (Carmody and Wente 2009). Multiple RBPs present in the packaged mRNP are able to recruit NXF1 and NXT1, including ALYREF, UAP56-interacting factor (UIF), SR proteins SRSF1, SRSF3 and SRSF7, Dbp5, and CPSF6. The redundant capacity of several mRNP components to recruit NXF1-NXT1 may enhance export efficiency. It may also regulate nuclear export of specific sets of mRNPs targeted by a particular NXF1-NXT1 recruiting factor. For example, ALREX (alternative mRNA export) elements bind mRNAs encoding the signal peptides for endoplasmic or mitochondrial localization and recruit NXF1-NXT1 (Cenik et al. 2011). Moreover, some genes follow alternatives to NXF1-NXT1 pathway. For instance, HuR binds AU-rich elements and recruits the exportin CRM1 instead of NXF1, promoting mRNP export (Brennan, Gallouzi, and Steitz 2000). Hence, the composition of mRNPs regulates its translocation to the cytoplasm.

To avoid being shuttled back to the nucleus, cytoplasmic mRNPs must undergo major remodeling. The best known effectors of this process are the helicase DDX19 and Gle1 (Carmody and Wente 2009). DDX19 is locked onto RNA during mRNP processing in the nucleus, its closed conformation stabilized by ATP binding. In turn, DDX19 stabilizes export factor binding to mRNA. Interaction with nucleoporins through the NPC and with Gle1 changes DDX19 to an open, ADP-bound conformation. This destabilizes mRNA binding of NXF1-NXT1 and displaces them from the mRNP. The detailed mechanism behind the displacement, and whether there are DDX19-independent remodeling pathways, is still unknown.

mRNP localization

The ability to place transcripts in particular sub-cellular location allows spatial regulation of protein production (Gáspár and Ephrussi 2017). Localization can be achieved with passive diffusion of mRNPs, or through active transportation via the cytoskeleton by molecular motors. While there are several modes of mRNP localization, we will only present briefly a few examples.

In the unicellular organism *Saccaromyces cerevisiae*, repression of mating type switching in the daughter cell requires the localization of the ASH1 mRNA to the budding place (Bobola et al. 1996). Localization signals in the ASH1 coding sequence and 3'-UTR are recognized by She2 protein dimers, which in turn recruit She3 (Böhl et al. 2000). This complex interacts with myosine and promotes the active transport via actin fibers of the cytoskeleton.

In multi-cellular organisms, mRNP localization is necessary for cellular polarization, particularly in differentiation and embryonic development. In fibroblasts, beta-actin mRNA contains ACACCC repeats known as the *zipcode*. This element is recognized by zipcode-binding proteins (ZBPs) and promote transport towards *lamellipodia*. Translation of localized beta-actin transcripts guarantees cytoskeleton-mediated cell motility (K. C. Martin and Ephrussi 2009). Similarly, hnRNP A2 mediates MBP (myelin basic protein) localization in oligodendrocytes.

In *Drosophila melanogaster*, protein concentration gradient is required for the definition of the anterior-posterior axis of oocytes. Proper oocyte morphology is crucial for embryonic development. Staufen proteins bind *bicoid* mRNA and promote its active transport to the anterior pole via the microtubule structures, ensuring correct anterior development of the embryo (Johnstone and Lasko 2001). p75 binding is necessary for *nos* mRNA posterior localization, which leads to abdominal development; ectopic localization of *nos* leads to the development of mirrored abdomens instead of the proper head-abdomen phenotype. Finally, localization of *oskar* mRNA to the posterior pole is necessary for oocyte polarization.

RBP-mediated localization is thus crucial for differentiation and developmental processes. Incorrect mRNP assembly has detrimental consequences for the cell and the organism. Overall, fine-tuning the RBP interactions with mRNA targets and other factors allows spatial regulation of gene expression.

Translation

Translation of mRNA constitute another level of regulation, which can also be mediated by RBPs (Sonenberg and Hinnebusch 2009). Evidence of differential translation rates has been elucidated with high-throughput ribosome profiling (Legrand and Tuorto 2020). Translation is a sequential reaction where the ribosome, using mRNA as a template, catalyzes the formation of peptide bonds between aminoacids. Aminoacids are carried by transfer RNA (tRNA) forming aminoacyl-tRNA molecules. In eukaryotes, it starts with the interaction between eukaryotic initiation factors eIF1, eIF1A, eIF3, eIF5, and the complex eIF2-TC, with the 40S subunit of the ribosome (Majumdar, Bandyopadhyay, and Maitra 2003). This results in the formation of the 43S-PIC (43-S pre-initiation complex). In parallel, the nuclear cap-binding proteins are displaced by initiation factors eIF4E, eIF4A and eIF4G, which bind the 5'-cap and form the eIF4F complex (Jackson, Hellen, and Pestova 2010). Additionally, cytoplasmic poly-A binding proteins (PABPC) displace their nuclear counterparts (PABPN) after mRNP nuclear export (G. Singh, Pratt, et al. 2015). Next, the 43S-PIC is recruited to the 5'-UTR of the mRNA and scans the transcript until it reaches the initiation codon. Recognition of the initiation codon triggers the recruitment of the 60S subunit of the ribosome, which displaces the factors eIF1, eIF1A, eIF2, and eIF5.

Translation elongation consists in the sequential recruitment of tRNAs whose anti-codon sequence has perfect complementarity with the codons read by the ribosome. This is determined by the interaction between eEF1A and aminoacyl-tRNAs. After peptide bond formation, eEF2 mediates the ribosome translocation to the next codon and the process reiterates. Translation finishes once the ribosome reaches a stop codon, a trinucleotide sequence that is not recognized by any aminoacyl-tRNA under normal conditions. The release the synthesized peptide is mediated by the termination factors eRF1 and eRF3. Ribosome subunits are released from mRNA through the activity of ABDE1, which displaces eRF1 and eRF3 and allows the formation of a new 43S-PIC unit (Jackson, Hellen, and Pestova 2012).

This perfectly coordinated process is not exclusively under the control of translation factors. Indeed, RBPs play important roles in protein synthesis regulation.

The 43S-PIC is only capable of binding the 5'-UTR in the absence of secondary structures. The RNA helicase eIF4A, along with the factors eIF4G and eIFB/H, linearizes 5'-UTR allowing 43S-PIC recruitment. eIF4G remodeling is also necessary for 43S-PIC scanning prior to initiation codon recognition. PABPC interact with the eIF4F at the 5'-end forming a closed-loop conformation (Jacobson 1996). This model explains PABP translation enhancement observed in development dynamics. Conversely, translation can be prevented through mRNA deadenylation. Translation arrest is necessary during oocyte maturation and guarantees proper embryo development. Deadenylation is considered to be mediated by PARN (poly-A specific RNase) and EDEN-BP (embryonic deadenylation element binding protein), which recognizes specific motifs in the 3'-UTR of specific transcripts (Johnstone and Lasko 2001). Translation of mRNAs can also be prevented by *masking*. This strategy consists in maintaining mRNPs in a conformation that inhibit recognition by translation initiation factors. The Y-box protein p50, a component of mRNP in mammals, inhibits translation at high concentrations (Davydova et al. 1997). These examples illustrate the central role of RBPs in protein synthesis. The translation of specific mRNAs is under the control of RBPs in specific cellular contexts, defining the fate of the cell and the organism.

Transcript stability and turnover

Although an individual mRNA molecule can be translated several times, its existence is not infinite. The average half-life of a mammalian mRNA is approximately 8h hours, but it ranges from a few minutes to over 24 hours (G. Singh, Pratt, et al. 2015). Several pathways exist to control the abundance of mRNA: deadenylation-dependent (executed by the CCR4-NOT complex); deadenylation-independent (triggered by Edc3 and Rps28B), and endonuclease-mediated (e.g. IRE1, PMR1, or MRP) (Garneau, J. Wilusz, and C. J. Wilusz 2007). mRNAs are ultimately degraded by the exosome complex (from the 3'-end after deadenylation or endonuclease cleavage), or the exonuclease XRN1 (from the 5'-end after decapping by DCP1-DCP2, or endonuclease cleavage). Altogether, these mechanisms regulate the stability of mRNA, which ultimately defines the amount of protein produced in the cell. mRNA decay is therefore a highly regulated process.

Decay is also a quality control strategy to eliminate incorrect transcripts (Conti and Izaurralde 2005). Shifts in the open reading frame can generate premature stop codons, causing the ribosome to halt mid-translation. Nonsense-mediated decay (NMD) is a mechanism that allows the cell to detect these events and degrade the offset transcript. Conversely, transcripts lacking stop codons undergo non-stop decay (NSD, Hoof et al. 2002; Frischmeyer et al. 2002), where the Ski7 protein triggers exosome degradation (if it interacts with a stalled ribosome), or XRN1 degradation (if PABP and Ski7 are absent). Additionally, transcripts that form strong secondary structures that hamper ribosome translocation are targeted by no-go decay (NGD). In yeast, Dom34 and Hbs1 trigger endonuclease cleavage of the mRNA at the stalling site, which is followed by XNR1 and exosome degradation (Doma and Parker 2006). Although they are known to guarantee the elimination of faulty mRNAs, the exact mechanisms behind each target vary across species and are not fully understood.

AU-rich elements (ARE) are widely studied *cis*-acting regulators of mRNA stability. They are defined by the presence of one or more copies of the AUUUA sequence. The rapid decay of ARE-containing transcripts happens in two steps: first, shortening of the poly-A tail, and second, digestion from both ends by 5'- and 3' exonucleases (Schoenberg and Lynne E. Maquat 2012). Destabilizing ARE binding proteins (ARE-BP) such as TTP, BRF1, BRF2, KSRP, and hnRNP D, recruit deadenylases and nucleases to favor mRNA decay. Conversely, HuR binds ARE and counteracts the action of destabilizing ARE-BP by preventing the recruitment of decay factors. ARE-containing transcripts represent 9% of cellular mRNAs, including proto-oncogenes and inflammation response genes in immune system cells. The RBP interplay in mRNA decay is therefore central to understand cellular homeostasis in health and disease.

2.1.4 The study of PTGR is a network problem

We have given an overview of the stages of PTGR where RBPs intervene. They shape the life of mRNPs through RNA- and protein-interactions in a coordinated fashion that involves several different factors. Ultimately the cooperative or antagonistic relationship of these factors define the fate of mRNPs and gene expression. The knowledge accumulated over the decades comes from studies that analyze the interplay between a handful of RBPs and their known targets. Yet, the function of over one third of RBPs is still unknown (Gerstberger, Hafner, and Tuschl 2014). Moreover, recent studies have revealed the existence of non-conventional RBPs with previously unknown RNA-related functions (Hentze et al. 2018). The high diversity of RBPs and the complexity of their interactions make the study of RBPs a systems biology problem that needs systems biology tools to be addressed.

2.2 The Exon Junction Complex

Among the vast list of PTGR actors, the Exon Junction Complex (EJC) is a central player in many stages of mRNA life. Discovered through the study of nonsense mediated decay (NMD), subsequent insights revealed its role in pre-mRNA processing, mRNP packaging, nuclear export, sub-cellular localization, and translation enhancement. It is thus not surprising that haploinsufficiency of the core components of this complex is responsible for several developmental syndromes. In this section, we will first present the discovery and the nature of this multi-protein complex. We will then overview its multiple functions throughout the life of mRNAs.

2.2.1 The discovery of the EJC

Nonsense mediated decay insures the elimination of incorrect transcripts that would otherwise generate truncated proteins. Premature termination codons (PTC, also known as premature nonsense codons) can arise from DNA rearrangement or mutations, or from pervasive RNA transcription, splicing or editing. The first obser-

vations in mammals revealed lower abundance of mRNA containing a PTC (Losson and Lacroute 1979; Lynne E. Maquat et al. 1981, see Fig. 2.2a). Inhibiting translation or *masking* the PTC from the ribosome (through frame-shifting or mutant tRNA addition) increases abundance of PTC-containing transcripts, which shows that NMD is translation-dependent (Fig. 2.2b).

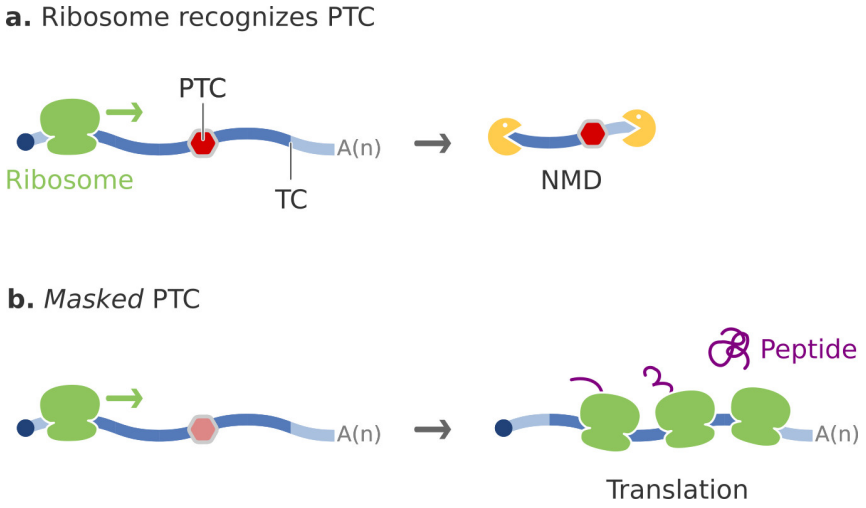

Figure 2.2: **a.** PTC-containing transcripts are targeted by NMD, resulting in low abundance. **b.** *Masking* PTC from the ribosome prevents PTC-containing transcripts to be targeted by NMD. *PTC*: premature termination codon. *TC*: termination codon. *NMD*: nonsense mediated decay. *A(n)*: poly-A tail.

Studies in mammals revealed that transcripts containing a PTC close to the 5'-end were more likely to be degraded than those containing a PTC close to the 3'-end. Further evidence showed that efficient NMD requires the presence of an intron downstream of the PTC, on the condition that the PTC is placed at least 50 to 55 nucleotides away from the exon junction. This suggested that NMD was also dependent on mRNA splicing (Fig. 2.3). It was surprising that a nuclear process had an effect only observable in the cytoplasm. Thus, it was contemplated that the splicing process imprinted mRNAs with a molecular mark that, once detected downstream of a nonsense codon by the ribosome, would trigger NMD.

To bring this molecular mark to light, Le Hir and colleagues performed *in vitro* splicing assays followed by RNase H protection assays (Le Hir, Izaurralde, et al. 2000). *In vitro* splicing is a convenient strategy to assess the splicing reaction under controlled conditions (Mayeda and Krainer 2012). Typically, nuclear extracts are incubated with short pre-mRNA reporters, and the splicing products are visualized on polyacrylamide gel electrophoresis (Fig. 2.4). This allows for the spliceosome to act on the pre-mRNA substrate in the presence of nuclear factors, thus simulating *in vivo* splicing.

RNase H digestion assays are a powerful tool to map protein-RNA interactions (Günzl and Bindereif 1999). Incubating complementary DNA oligos with mRNA forms DNA-RNA hybrids, which are targeted and cleaved by RNase H (Fig. 2.5). If a protein binds the mRNA where it is complementary to the oligo, the DNA-RNA hybrid does not form and RNase cleavage is prevented.

The authors incubated several mini-gene pre-mRNAs with either HeLa nuclear extracts, or were injected to *Xenopus laevis* oocytes. The splicing products were

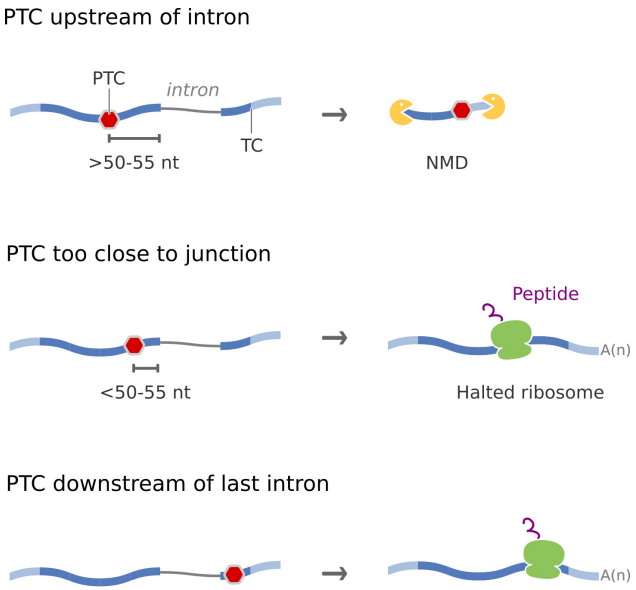

Figure 2.3: PTC triggers NMD depending on its relative position to the intron. If it is too close to the exon junction (less than 50 nucleotides), or if it is downstream of the last intron, NMD is not triggered. *TC*: termination codon. *NMD*: nonsense mediated decay. *A(n)*: poly-A tail.

then incubated with 12-nucleotide long oligos targeting different locations along the exon junction. They found that the window centered 24 nucleotides upstream (-24) of the exon junction was consistently protected from digestion. Protection of the region happened regardless of the gene, and both in nuclear extracts and *X. laevis* oocytes. This suggests that the factors deposited by the spliceosome was sequence independent and evolutionarily conserved. In the study, selective immunoprecipitation (IP) of the proteins bound to the -24 region identified SRm160, DEK, RNPS1, Y14, and REF. Posterior studies of the proteins associated to this region unveiled the components of what we know today as the EJC.

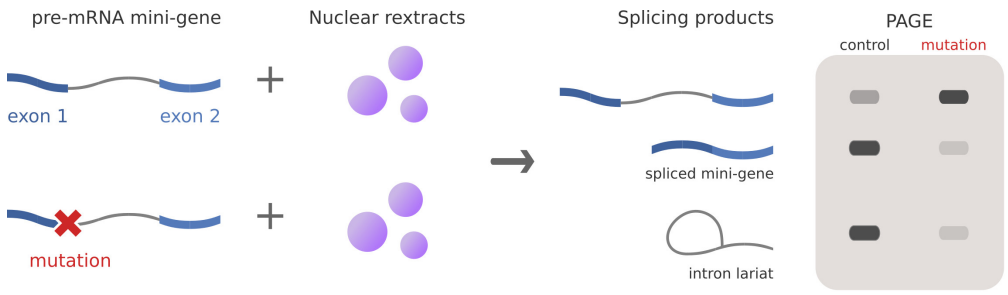

Figure 2.4: In this example, a pre-mRNA mini-gene containing a 5'-splice site mutation is compared to a control pre-mRNA in an *in vitro* splicing assay. The impact on splicing product abundance is assessed using electrophoresis. *PAGE*: polyacrylamide gel electrophoresis

Figure 2.5: **a.** RNase H is an endonuclease that recognizes and cleaves DNA-RNA hybrids. **b.** Protected RNA stretches prevent DNA hybridization and thus the action of RNase H.

2.2.2 Elucidating the structure of the EJC

The core of the EJC is composed of four proteins: MAGOH (*mago nashi* homologue), Y14 (or RNA-binding motif 8A, RBM8A), CASC3 (or metastatic lymph node 51, MLN51), and eIF4A3 (eukaryotic initiation factor 4A3). These *minimal* EJC core was inferred by Tange and colleagues through alternating IPs of each component, which revealed that they were capable of forming a stable tetrameric complex (Tange et al. 2005; Ballut et al. 2005). Structural studies revealed how the interaction between the components result in a stable, sequence-independent grip on single stranded RNA (Bono, Ebert, et al. 2006; Andersen et al. 2006), conserved across different species. Here, we will summarize the experimental evidence of each component, as well as their link between structure and function.

Hand in hand: the MAGOH/Y14 heterodimer

The *mago nashi* gene was first described to be essential for *D. melanogaster* oocyte mRNA localization through deleterious mutation assays (Newmark and Boswell 1994). It was later discovered that *mago* deposition on mRNA was splicing dependent (Le Hir, Gatfield, Braun, et al. 2001). Soon, *mago* homologues were described in several species, including human, and were designated MAGOH, for *mago* homologue. Yeast two-

hybrid and IP assays demonstrated that MAGOH specifically binds Y14 (Zhao et al. 2000). Y14 was later identified as a component of mRNPs, specifically deposited at the -24 region in a splicing-dependent manner, which would remain on mRNA in the cytoplasm (Kataoka et al. 2000). Their overlapping functional impact and splicing-dependent effect suggested that they were part of the EJC.

D. melanogaster and human crystal structure of the MAGOH/Y14 heterodimer revealed their close interaction (Shi and R.-M. Xu 2003; Lau et al. 2003). In the heterodimer, MAGOH forms a flat β-sheet stacked against two α-helices (Fig. 2.6a). Y14 contains two well-conserved motifs (RNP1 and RNP2) that form RNA-binding domain (RBD), which was thought to be responsible of the EJC anchor to RNA (Fig. 2.6b). Yet, the structure shows that the RBD interacts directly with MAGOH's helices to form a highly stable complex (Fig. 2.6c). The structure further reveals exposed surfaces on MAGOH that could serve as interaction platforms with other RBPs. Thus, the tight structural relationship between the proteins explains their functional overlap.

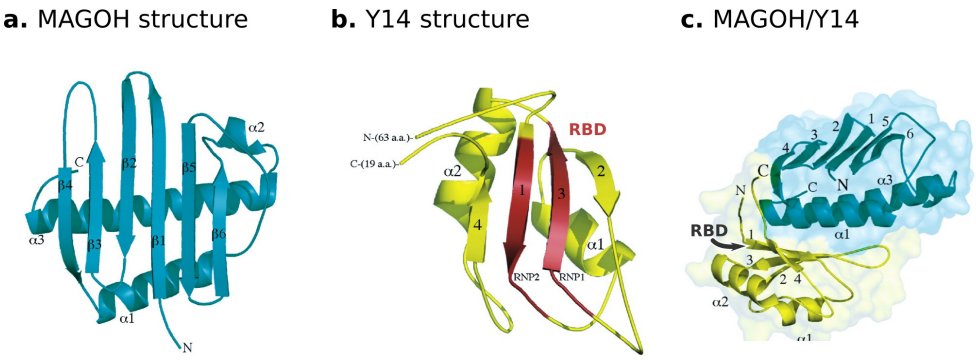

Figure 2.6: **a.** Crystal structure of free MAGOH. **b.** Crystal structure of free Y14. Two anti-parallel β-sheets (in red) form the RBD. **c.** Crystal structure of the MAGOH/Y14 heterodimer. The Y14 RBD is covered by the MAGOH α-helices. RBD: RNA-binding domain. Adapted from Lau et al. 2003.

A stable clamp: the eIF4A3 helicase

The roles of the *mago*/Y14, and the CASC3 homologue Barentz (*Btz*) were clear in the context of mRNA localization in *D. melanogaster* oocyte maturation. It was proven that *mago*/Y14 were necessary for *Btz* assembly on *oskar* mRNA and subsequent localization to the posterior pole of the oocyte. Although they are all essential for embryo development, there was no evidence that these proteins interacted *in vivo* or *in vitro*. Yeast two-hybrid assays of *Btz* revealed the DEAD-box helicase eIF4A3 as an interaction partner (Palacios et al. 2004). eIF4A3 knock-down in *D. melanogaster* oocytes resulted in the same phenotype as mutants of the other components, showing its functional link to the complex. Co-IP assays revealed that eIF4A3 interacts with the *mago*/Y14 heterodimer as well, proving that it was the missing link between *mago*/Y14 and *Btz*. *In vitro* splicing followed by IP confirmed the splicing-dependent incorporation of eIF4A3 to the EJC in human cells (Shibuya, Tange, Sonenberg, et al. 2004). Moreover, the presence of eIF4A3 was necessary to trigger NMD, confirming its functional role in the complex.

eIF4A3 is a member of the DEAD-box helicase family. Crosslinking assays

proved that spliced RNA is directly bound by eIF4A3, and not by the MAGOH/Y14 heterodimer (confirming the observations from its crystal structure 2.6c). The eIF4A3 RNA-binding activity is ATP-dependent, and is stabilized by the inhibition of its hydrolisis by MAGOH/Y14 (Ballut et al. 2005). The crystal structure of the free form of eIF4A3 shows two globular RecA-like domains joined by a 10-residue linker (Andersen et al. 2006; Bono, Ebert, et al. 2006). In its free form, eIF4A3 is in an *open* conformation where the two RecA domains are apart from each other (Fig. 2.7a). In the presence of ATP and RNA, it adopts a *closed* conformation where the ATP binds the linker domain (Fig. 2.7b). The RecA domains wraps the RNA backbone, interacting with the 2'-OH groups, and covering between 8 to 9 nucleotides. The irrelevance of the nucleotide bases to the interaction explains the sequence-independent binding of the EJC. The crystal structure of the tetrameric EJC core, shows how MAGOH/Y14 surround the eIF4A3 linker region and approaches the ATP-binding site (Fig. 2.7c). Interestingly, the MAGOH residue Ile146 directly interacts with ATP molecule bound to eIF4A3 (Fig. 2.7d). This offers structural proof of the MAGOH/Y14-dependent inhibition of eIF4A3 opening after ATP hydrolysis (Nielsen et al. 2009), and their role in the EJC stabilization.

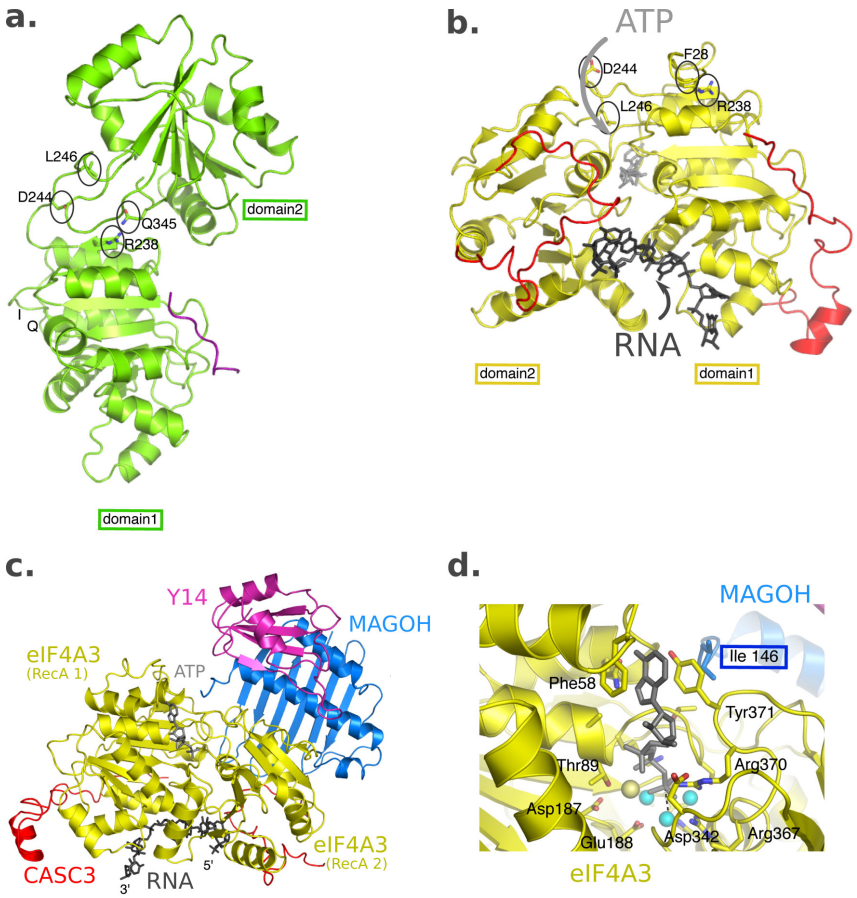

Figure 2.7: **a.** Crystal structure of free eIF4A3 in its open conformation. **b.** Crystal structure of eIF4A3 (yellow) as part of the EJC in its closed conformation, in the presence RNA (in dark gray), and ATP (in light gray). **c.** Crystal structure of the EJC showing the interactions between the MAGOH (in blue) and Y14 (in magenta) heterodimer; CASC3 is shown in red. **d.** Detailed interaction between eIF4A3 (in yellow) and ATP (in gray), and the involvement of MAGOH the Ile146 residue (in blue). Adapted from Bono, Ebert, et al. 2006

Wrap it up: CASC3

Human CASC3 was first identified in breast cancer tissues, hence the name metastatic lymph node 51 (MLN51, Degot, Régnier, et al. 2002). Its co-localization in nuclear speckles with MAGOH/Y14, along with its specific binding to spliced mRNA suggested a functional link with the EJC (Degot, Le Hir, et al. 2004). As discussed above, its role in *D. melanogaster* oocyte development also suggested a role in the EJC. *In vitro* EJC reconstitution with recombinant proteins, showed that CASC3 was necessary to stabilize the MAGOH/Y14 interaction with eIF4A3. Truncated forms of the protein have shown to impair mRNA localization in *D. melanogaster*, thus proving its functional role in the complex.

CASC3 has no globular structural folding. It contains two highly conserved domains in its N-terminal and C-terminal ends, linked by a sequence with no ordered structure (Fig. 2.8a). The N-terminal domain directly interacts with the RecA domain 1 of eIF4A3, while the C-terminal domain interacts with the RecA domain 2. Moreover, CASC3 residue GLu190 forms a cluster of interaction with eIF4A3 and MAGOH, which is necessary to maintain EJC function in *D. melanogaster* oocytes. Additionally, CASC3 interacts with one RNA nucleotide (Fig. 2.8b), which has been shown to increase EJC RNA-binding efficiency. Altogether, these findings demonstrate the role of CASC3 in stabilizing the EJC core structure. However, there is some evidence that shows that CASC3 is not an essential component of the EJC, which suggests that the core composition of the EJC can vary among different mRNPs (Gehring, Lamprinaki, Hentze, et al. 2009; Mabin et al. 2018).

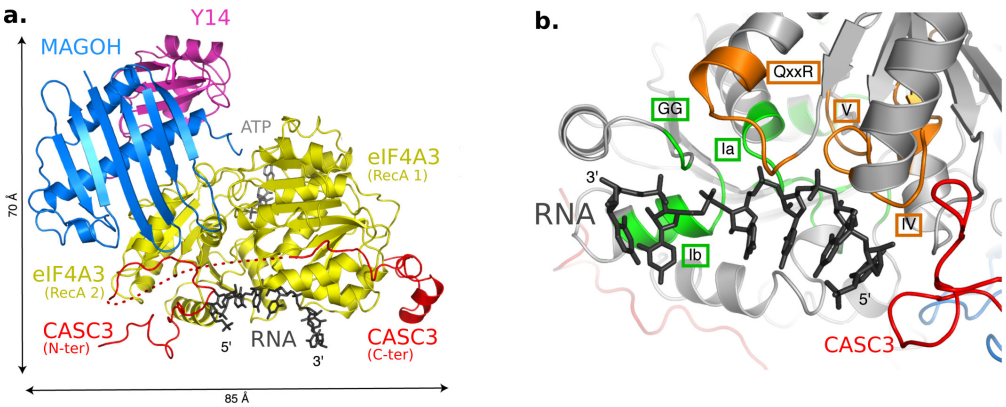

Figure 2.8: **a.** Crystal structure of the EJC revealing the CASC3 (in red) interaction with eIF4A3 RecA domains (in yellow). **b.** Detailed interaction between eIF4A3 and RNA showing the involvement of CASC3 (in red). Adapted from Bono, Ebert, et al. 2006

2.2.3 The life cycle of the EJC

From its assembly by the spliceosome, through its journey to the cytoplasm within the mRNP, until its disassembly, the core components of the EJC interact with multiple partners designated as EJC peripheral factors. In this section, I will describe the current knowledge of the EJC assembly by the spliceosome, the dynamics of EJC peripheral factors, and finally the disassembly and recycling of its core components.

Putting the pieces together: spliceosome-dependent assembly

Although the role of the spliceosome has been proved for EJC assembly and function, the exact assembly mechanism has only been partially elucidated. IP of eIF4A3 nuclear extracts, followed by Mass-spectrometry (MS), revealed the splicing factor CWC22 (Complexed With Cef1 22) as a strong partner (Barbosa et al. 2012). CWC22 is composed of two distinct domains: MIF4G, near the MIF4G, and MA3, towards the middle of the peptide (Buchwald et al. 2013). Published crystal structures of several spliceosome conformations show CWC22 as part of splicesomes B^{act}, C, C*, and P (Haselbach et al. 2018; X. Zhang et al. 2017). Prior to the first splicing reaction, the MA3 domain binds the 5'-exon while the MIF4G appears placed on the opposite side of the exon canal. Pull-down and MS experiments revealed CWC27, another splicing factor, as an important partner of CWC22 (Busetto et al. 2020). CWC27 also appears in an early form of B^{act}, but not in the later B^{act} form, or the C splicesome.

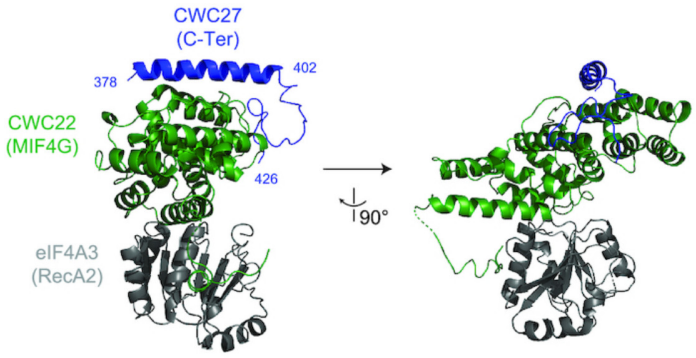

Figure 2.9: Crystal structure of the CWC22/CWC27/eIF4A3 trimer in two orientations. Adapted from Busetto et al. 2020

Recently, a crystal structure of recombinant CWC22, CWC27 and eIF4A3 shows how the three proteins are able to interact (Busetto et al. 2020). The C-terminal end of CWC27 contacts the MIF4G domain of CWC22, which in turn is in touch with the RecA 2 domain of eIF4A3 (Fig. 2.9). The current model of EJC assembly was inferred from the spliceosomes crystal structures and the CWC22/CWC27/eIF4A3 trimer structure. It proposes CWC27 as a binding partner of CWC22 prior to recruitment to the spliceosome, and in the early B^{act} (Fig. 2.11). The CWC22/CWC27/eIF4A3 trimer exists in an intermediate, or *mature* B^{act}, which is structurally compatible with the published B^{act} crystal structure. CWC27 is then released from the late B^{act}, while CWC22 remains associated with eIF4A3 via the MIF4G domain until the spliceosome P conformation. How the other EJC components are assembled onto mRNA between the late B^{act} spliceosome and the C spliceosome is still unknown. Whether additional factors contribute to the recruitment of eIF4A3, and the other EJC components to the spliceosome remains an open question.

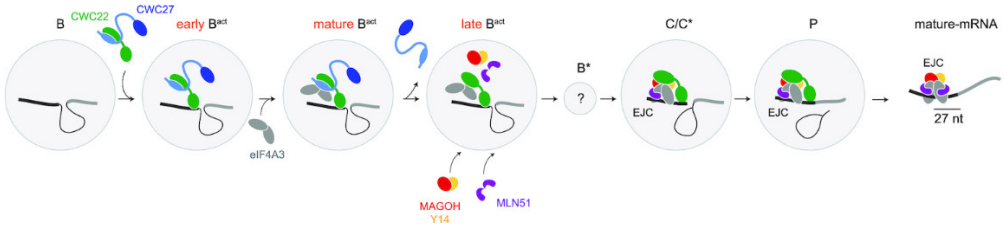

Figure 2.10: A schematic representation of the splicing-dependent assembly of the EJC. The steps are inferred from the most plausible structural compatibility between the CWC22/CWC27/eIF4A3 crystal structure and published spliceosome structures. Adapted from Busetto et al. 2020

Preparing the package: nuclear peripheral factors and mRNP packaging

Following EJC assembly and RNA processing, several interactions between RBPs take place during what is known as *mRNP packaging*. The first known peripheral factors to come into contact with the EJC during mRNP packaging, are likely present prior to its assembly. They include splicing factors RNPS1 (RNA-binding protein with Ser-rich domain 1), ACINUS, PININ, and SAP18 (Le Hir, Saulière, and Z. Wang 2016). RNPS1 and SAP18 are known splicing enhancers that form 2 different complexes known as ASAP and PSAP, depending on whether they are bound to ACINUS or PININ, respectively. Although they have been identified as EJC partners, the exact protein domains involved in the interaction are still unknown.

Other pre-assembly factors include mRNA nuclear export factors ALYREF and UAP56. They are known components of the TREX (transcription-export) complex, necessary for export to the cytoplasm. The low affinity between UAP56 and the EJC suggests that their interaction is transient, and that UAP56 release is necessary for the recruitment of other peripheral factors such as NXT1 and NXF1. As discussed in section 2.1.3, the latter form a heterodimer that directly interacts with NPC subunits and favors mRNP nuclear export.

Because some peripheral factors are restricted to the nucleoplasm, the composition of the EJC is remodeled in the cytoplasm. Among the cytoplasmic peripheral factors, we find the homologues UPF3A and UPF3B, UPF2 and SMG6. ACINUS and PININ nuclear localization implies that they must be released from the EJC. Since RNPS1 and SAP18 are still part of the mRNP in the cytoplasm, it is hypothesized that other RBPs stabilize their interaction. Such may be the case of UPF3A and UPF3B, which can physically interact with RNPS1. Interestingly, several peripheral factors are mutually exclusive: UPF3A/UPF3B/SMG6, and ACINUS isoforms/PININ. This indicates that EJC composition is not only dynamic throughout its life, but it is also variable across different mRNPs (Z. Wang, Ballut, et al. 2018). However, the mechanisms that determine the composition of individual EJCs and their functional implications are still unknown.

Finally, some studies suggest that members of the SR protein family interact with the EJC as peripheral factors: SRSF1, SRSF3, and SRSF7. Physical interactions between seem to result in higher-order organization of mRNPs (Metkar et al. 2018). Recently, it has been shown that mRNP composition can affect NMD sensitivity (Mabin et al. 2018). Whether this is an effect of the presence of specific factors,

or the mRNP higher-order organization is yet to be determined.

The separation: EJC disassembly in the cytoplasm

The EJC exists as a complex within mRNPs until they start being translated. Pull-down of spliced mRNPs using antibodies against cap-binding proteins co-purified EJC components when using antibodies against nuclear cap-binding protein CBP80, but not with cytoplasmic cap-binding initiation factor eIF4E (Lejeune et al. 2002). A different approach showed that mRNPs bound by a single ribosome (monosomes) contained Y14, while those bound by multiple ribosomes (polysomes) did not (Dostie and Dreyfuss 2002). These results indicated that EJC is not present in actively translated transcripts. Thus, they proved that EJC disassembly is translation-dependent and that it occurs after the first round of translation.

It was later shown that EJC disassembly was ribosome-dependent, and supported by the cytoplasmic protein PYM (partner of Y14 and MAGOH). The N-terminal section of PYM is able to interact with MAGOH/Y14 at the interface with eIF4A3. This prevents EJC re-assembly by hindering MAGOH/Y14 interaction with eIF4A3, thus keeping the latter in an open conformation (Gehring, Lamprinaki, Kulozik, et al. 2009). In mammalian systems, the C-terminal section of PYM is able to associate with the ribosome. This offers a molecular link between translation and EJC disassembly. However, in *D. melanogaster*, PYM does not bind the ribosome, suggesting a different translation-dependent mechanism for EJC disassembly (Ghosh, Obrdlik, et al. 2014).

Although the CASC3/eIF4A3 heterodimer localizes in the cytoplasm, it is still unknown if it has an impact on mRNP regulation outside of the EJC (Le Hir, Saulière, and Z. Wang 2016). Their recycling mechanism has not been characterized. MAGOH/Y14, on the other hand, are localized in the nucleus and their recycling mechanism is well described. Importin 13 (Imp13) is a member of the karyopherin family that binds MAGOH/Y14 forming a ring structure around the heterodimer (Bono, Cook, et al. 2010). This interaction destabilizes the association with PYM. Once in the nucleoplasm, Imp13 binds RanGTP, which releases MAGOH/Y14 (Mingot et al. 2001). RanGTP then escorts Imp13 back to the cytoplasm. The free MAGOH/Y14 can then be incorporated to new EJCs by the spliceosome.

In summary, we have presented how the EJC components are anchored on mRNA by the spliceosome with a stable grip. Hence, this complex is the molecular *message* left on spliced transcripts in the nucleus that can be read by cytoplasmic factors, such as other RBPs and the ribosome. The EJC accompanies mRNAs throughout their journey from the nucleus to the cytoplasm, thus marking transcripts that have not yet undergone translation.

2.2.4 A versatile actor: the EJC roles in the PTGR network

In the previous section, we presented the journey of the EJC from the nucleus to the cytoplasm. This matches the structural features of the complex that maintain

a *locked* conformation on mRNA. This allows the EJC to act as a platform for a succession of dynamic interactions with several RBPs, from pre-mRNA processing, to mRNP export, until mRNA translation and decay. The EJC is thus able to intervene in almost every step of mRNP regulation.

Spliced junctions impact splicing

The splicing reaction is concurrent with transcription elongation (Herzel et al. 2017). The fact that exon definition depends on numerous *cis*- and *trans*-acting factors, shows that splicing sites hold varying degrees of *strength* within the same gene (Graveley, Hertel, and Maniatis 1998; Fu and Ares 2014; Fontrodona et al. 2019). Thus, as nascent pre-mRNA is synthesized, its introns are spliced in an asynchronous manner, rather than in an consecutive 5'-to-3' fashion. Surprisingly, *trans*-acting factors from concluded splicing events, such as the EJC, can impact subsequent splicing of different junctions.

The RAS/MAPK signaling pathway has a central role in cellular proliferation, differentiation, and survival. Screening potential *trans*-acting regulators of the RAS/MAPK pathway in *D. melanogaster* cells, revealed a significant effect of core components *eIF4AIII*, Y14 (*tsu*), and *mago* (Ashton-Beaucage et al. 2010). Mutation of *eIF4AIII* and *mago* caused both reduction of *mapk* transcript levels, and exon skipping events resulting in different variants lacking exons 2 to 4, 2 to 3, or 2 to 5. *In vivo*, these EJC mutants displayed wing and eye morphology abnormalities. Moreover, transcriptome-wide data, showed that lack of EJC causes several intron retention events in long-intron containing genes. This indicates a specialized EJC role in splicing of a specific type of intron.

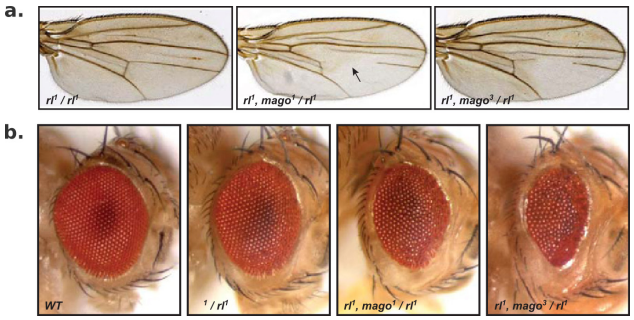

Figure 2.11: **a.** Atrophied wing phenotype in rl^1/rl^1 mutants (left panel) is aggravated in *mago* heterozygous mutants (center and right panels); vein formation is hindered in *mago* mutants (indicated by the arrow). **b.** Wild-type eye phenotype (left panel); rough eye phenotype in rl^1/rl^1 (second panel); aggravated rough eye in *mago* mutants (third and fourth panels), showing a smaller eye. Adapted from Ashton-Beaucage et al. 2010.

A similar study revealed the EJC-dependent regulation of the Piwi-interacting RNA (piRNA) pathway. The Piwi protein has an essential and highly conserved role in transposon silencing during germline development (Theurkauf et al. 2006). Inhibiting the piRNA pathway results in the impairment of the axes polarization in *D. melanogaster* oocytes, similar to the effect of EJC core component mutations. Analysis of the *piwi* transcript in mutated EJC components *eIF4AIII*, *tsu*, *mago*, and *RnpS1*, revealed intron 4 retention and reduction of Piwi protein (Hayashi et al. 2014; Malone et al. 2014). Removal of introns 3 and 5 of the *piwi* transcript resulted

in intron 4 retention, even in the presence of the EJC. Furthermore, studying the sequence of intron 4 revealed a *weak* polypyrimidine tract, which is sufficient for intron retention in the absence of EJC. Taken together, these results suggest the role of EJC in promoting the splicing of adjacent junctions that are particularly hard to detect, either because of intron length or weak splice sites.

Evidence suggests the EJC impacts splicing in human cells. HeLa cells treated with siRNA against EJC core components followed by mRNA-seq revealed splicing event alteration, including skipping of constitutive exons (Z. Wang, Murigneux, and Le Hir 2014). Silencing of either peripheral factors ACINUS or PININ caused different splicing alterations depending on whether the EJC was associated with ASAP or PSAP complexes (Z. Wang, Ballut, et al. 2018). This suggests EJC composition has distinct regulation roles of splicing events. Additionally, the EJC silences cryptic splicing sites that originate after the splicing reaction (Blazquez et al. 2018; Boehm et al. 2018). Knock-down of EJC core components resulted in aberrant splicing caused by recursive splicing. These observations highlight the importance of EJC assembly to guarantee transcript integrity. However, the effect of the EJC has exclusively been inferred from knock-down studies. Direct evidence of EJC binding sites near the aberrant splicing events is still absent.

Export to the cytoplasm

The effect of splicing on mRNA nuclear export was first tested in *X. laevis* oocytes (Luo and Reed 1999). An intron-containing reporter injected to the nucleus was more efficiently exported than the intron-less counterpart with identical sequence. The EJC peripheral factor ALYREF is recruited to mRNPs as part of the TREX complex, and promotes nuclear export through NXF1/NXT1. It has been shown that ALYREF-mediated export is cap- and EJC-core dependent (Gromadzka et al. 2016). Mutations of a short motif in an unstructured domain of ALYREF prevents eIF4A3 binding and impairs mRNP export to the cytoplasm. An alternative export pathway involves SR proteins (such as SRSF1, SRSF3, and SRSF7), which are able to interact with NXF1/NXT1 as well (Le Hir, Saulière, and Z. Wang 2016). The EJC may play a role in this pathway by binding SR proteins to stabilize mRNP higher-order organization (G. Singh, Kucukural, et al. 2012). However, the EJC effect on mRNP export has been observed for short transcripts, suggesting longer transcripts are bound by several factors with redundant export-enhancing effects.

Transcript localization

As discussed in section 2.1.3, mRNA localization is crucial for *D. melanogaster* oocyte polarization and proper embryonic development. One example is the posterior localization of the three-intron *oskar* transcript. Intron-less transgenes revealed that splicing of only intron 1 and EJC deposition was necessary for *oskar* localization and correct oocyte development (Hachet and Ephrussi 2004). Replacing *oskar* sequence stretches with *lacZ* sequences, revealed the essential role of a loop-forming localization element near the exon 1 junction (Ghosh, Marchand, et al. 2012, see Fig. 2.12).

Interestingly, the integrity of the stem loop is not necessary for EJC deposition. This led to two hypotheses: either the secondary structure promotes recruitment of EJC stabilization and localization factors, or the EJC itself or peripheral factors stabilize stem loop structure to ensure mRNP localization. The *oskar* transcript constitutes the only example with experimental evidence for EJC-dependent localization. A direct impact of the EJC on sub-cellular localization of other invertebrate or mammalian transcripts is still unknown.

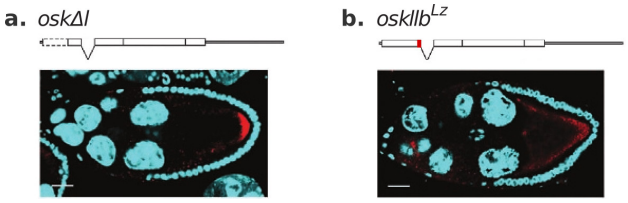

Figure 2.12: **a.** Truncated *oskar* construct missing endogenous 5'-end of exon 1. Truncation does not have a negative effect in *oskar* localization to posterior pole (FISH probes in red). **b.** *oskar-lacZ* hybrid construct where endogenous exon 1 junction sequence is replaced by *lacZ* sequence (in red). Hybrid transcript fails to localize to the posterior pole. Oocyte visualization with DAPI staining in cyan. Adapted from Ghosh, Marchand, et al. 2012.

The EJC in translation

Spliced transcript expression is enhanced at the protein level. Higher protein levels for intron-containing reporters relative to intron-less counter-parts was observed in plant (Callis, Fromm, and Walbot 1987), mice (Palmiter et al. 1991), and human (NOTT, MEISLIN, and MOORE 2003) cells. Tethering experiments showed that EJC-bound transcripts are more efficiently translated, suggesting an EJC-mediated mechanism of translation enhancement. One possible mechanism is through CASC3-ribosome interaction. Expression of CASC3 correlates with global protein synthesis in an EJC-dependent manner (Chazal et al. 2013). Purification of polysome fractions, and in vitro reconstitution shows that CASC3 directly interacts with eIF3 components (responsible for 43S-PIC formation). Interestingly, eIF4A3 and CASC3 are detected in heavy polysome fractions, but not MAGOH/Y14, indicating eIF4A3/CASC3 interaction with actively translated transcripts. This suggests eIF4A3/CASC3 enhance translation outside of the EJC core.

Another translation enhancement mechanism involves eIF4A3 and the mTOR signaling pathway (Ma et al. 2008). Following mTOR activation, the S6K1 kinase promotes translation through riboprotein and initiation factor phosphorylation. It was shown that the SKAR protein (S6K1 Aly/REF-like substrate) is able to bind to eIF4A3 and recruit S6K1, which in turn is able to phosphorylate its targets. This mTOR-dependent mechanism suggests regulation of specific transcripts rather than a global EJC effect on translation. However, additional evidence of eIF4A3/SKAR interaction is limited (Le Hir, Saulière, and Z. Wang 2016).

Till death do us part: EJC-dependent NMD

As discussed in section 2.2.1, NMD is a quality control mechanism that degrades PTC-containing transcripts, crucial for homeostasis and cell survival (L. E. Maquat 1995;

Lynne E. Maquat 2005). Following the discovery of the EJC, several studies demonstrated NMD was dependent on correct EJC assembly (Le Hir, Gatfield, Izaurralde, et al. 2001; Lykke-Andersen, Shu, and Steitz 2001; Palacios et al. 2004; Gehring, Kunz, et al. 2005; Shibuya, Tange, Stroupe, et al. 2006; K. K. Singh et al. 2013). The increasing evidence was coherent with the minimum 50-nucleotide distance between the PTC and the exon junction. A smaller distance to the EJC would indeed induce its disassembly and the inability to trigger NMD (Le Hir, Saulière, and Z. Wang 2016).

How does an assembled EJC downstream of a PTC trigger NDM? (Fig. 2.13). The current model proposes that a stalling ribosome binds to the helicase UPF1 (up-frameshift 1), an essential factor of NMD. UPF1 binds ribosome release factors eRF1 and eRF3, and SMG1, forming the SURF complex (SMG1/UPF1/eRF1/eRF3). UPF1 then interacts with EJC peripheral factors UPF2 and UPF3 to form the DECID complex (DECay InDucing). UPF1 is then phosphorylated and recruits SMG6, an endonuclease that cleaves mRNA, and SMG5 and SM7, which recruit general decay factors.

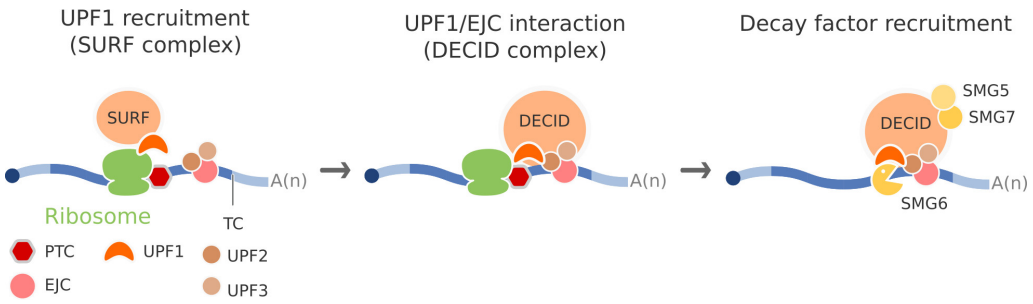

Figure 2.13: Simplified representation of the EJC-dependent NMD mechanism. PTC: premature termination codon. TC: termination codon. EJC: Exon Junction Complex. SURF: SMG1/UPF1/eRF1/eRF3 complex. DECID: decay inducing complex. Theoretical source: Le Hir, Saulière, and Z. Wang 2016.

In humans, EJC-mediated NMD regulates the expression of PTC-free transcripts as well. In mature neurons, the translation of *ARC* (activity-regulated cytoskeleton-associated) takes place at the synapse. The *ARC* gene contains two introns downstream of the stop codon. Once in the synapse, the *ARC* mRNP engages in a few rounds of translation before being targeted by EJC-mediated NMD (McMahon, Miller, and D. L. Silver 2016). This exemplifies how physiological gene regulation can take place through NMD.

The role of the EJC in NMD is, however, not universal. Alternative NMD pathways exist in plants, fungi (such as *S. cerevisiae*, and *S. pombe*), invertebrates (such as *D. melanogaster*, and *C. elegans*), and mammals (**'wen'splicing-dependent'2010**; Bühler et al. 2006; Gatfield et al. 2003). For instance, extended 3'-UTRs can trigger NMD in mammalian cells, although less efficiently than the EJC (Amrani et al. 2004; Brogna and Wen 2009). Furthermore, not all intron-containing 3'-UTRs genes are sensitive to NMD. These discrepancies on the EJC role in NMD across eukaryotes may be explained in part by alternative pathways. Another possibility is differential deposition of the EJC in specific junctions to regulate gene NMD-sensitivity and protein synthesis.

2.2.5 When the EJC fails: involvement in physiological disorders

Due to the multiple processes involving the EJC, it is not surprising that altered expression of its components has a major physiological impact. As discussed previously, EJC-mediated *oskar* localization is essential in developing *D. melanogaster* oocytes. Absence of EJC core components and failure to localize *oskar* mRNP to the posterior pole results in impaired abdominal patterning of the embryo (Kim-Ha, J. L. Smith, and Macdonald 1991; Hachet and Ephrussi 2004). The EJC regulation of *mapk* splicing is necessary for photoreceptor cell development, which is mediated by the EGF (epidermal growth factor) signaling pathway (Roignant and Treisman 2010). Finally, EJC-dependent splicing of the *piwi* transcript is required for proper transposon silencing during germline development, which is crucial for gonad development and fertility (Malone et al. 2014). These examples indicated the central role of the EJC in invertebrate embryonic development.

The EJC is also essential for mammalian development. In mice, knock-out of EJC components is lethal for embryos . Haploinsufficiency of MAGOH causes impairment of neural precursor mitosis, which results in microcephaly (Debra L. Silver et al. 2010). Mutants with decreased Y14 display similar phenotypes. Silencing of core component eIF4A3 is associated with affected neural stem cell (NSC) mitosis and increased apoptosis (Bartkowska et al. 2018). Similarly, depletion of EJC peripheral factors associated with NMD (UPF2 and UPF3) results inhibition of NSC proliferation. These observations are evidence of the central role of the EJC in mouse neurogenesis.

The EJC involvement is not limited to the nervous system development, but also extends to post-mitotic neurons (McMahon, Miller, and D. L. Silver 2016). Depletion of peripheral factor UPF2 induces accumulation of Robo3.2 receptor at the axon, causing axonal growth inhibition. The EJC-mediated NMD regulation of *ARC* dosage at the synapse is necessary for synaptic plasticity, which is associated with learning and memory in rats. Interestingly, the EJC components have an impact in animal behavior. Over-expression of Y14 correlates to increased synaptic activity, which results in anxiety and autism-like behavior in adult mice (Alachkar et al. 2013). The molecular mechanisms behind the EJC-dependent brain functions and their impact on behavior remain obscure.

The study of clinical syndromes has revealed the EJC role in human development (McMahon, Miller, and D. L. Silver 2016). Deletions in the chromosomal region that includes the Y14 gene are associated with an array of intellectual disabilities, autism, epilepsy, schizophrenia, and aberrant brain morphology. Combined with this chromosomal deletion, a point mutation in the Y14 gene causes the thrombocytopenia with absent radius (TAR) syndrome, which affects blood composition, limb morphology and the nervous system (Albers et al. 2012). Additionally, copy number expansion in the eIF4A3 gene causes the Richieri-Costa-Pereira (RCP) syndrome (Favaro et al. 2014). RCP patients display cranio-facial and limb malformations, as well as learning disabilities.

Mutations in CWC27 are associated with an array of phenotypes shared among splicing factor mutations, known as spliceosomopathies (Busetto et al. 2020). CWC27

deficiency causes a spectrum of disorders with varying degrees of severity, ranging from retinitis pigmentosa, short stature and skeletal development syndromes, craniofacial abnormalities, and neurological impairments (M. Xu et al. 2017). These phenotypes are similar to the ones we have mentioned for the EJC core deficiency. It is yet unknown whether splicing impairment causes EJC assembly failure resulting in similar disorders as direct reduction of EJC components. Alternatively, EJC assembly may be necessary for the correct splicing of particular transcripts in specific cells. Whether the EJC is assembled on specific junctions remains an open question.

Altogether, the evidence underpins the importance of the EJC at the cellular and the organism level. Its role in development, neurogenesis and cell differentiation suggests that the EJC is crucial in particular cellular contexts. Yet, further research is necessary to elucidate the EJC-dependent mechanisms at play.

Chapter 3

Mapping RNA-binding proteins with CLIP

The discovery of transcriptome-wide binding sites is highly valuable to elucidate the interplay between RBPs in PTGR networks. crosslinking and IP (CLIP) protocols have opened the door to study RBP targets and dynamics at the transcriptome level. Based on the pull-down principle of RNA IP (RIP), the covalent link created with UV radiation between proteins and RNA allows stringent purification conditions, resulting in noise reduction compared to prior protocols. In parallel to CLIP development, high-throughput sequencing technologies were on the rise. Coupled with sequencing, CLIP became a protocol to obtain a transcriptome-wide snapshot of the binding sites of a particular RBP. Soon, members of the PTGR community contributed to improve the efficiency and specificity of CLIP. Thus, the following decade witnessed the birth of several variations of the protocol, and the development of dedicated data analysis tools and pipelines. However, as the excitement sparked the creation of protocols and tools, the community also discovered the complications and shortcomings of CLIP, both at the experimental and the data analysis levels. On one hand, the efficiency of crosslink and IP are the major limiting steps to obtain high-quality data. On the other hand, there is yet no consensus to assess the quality and reproducibility of binding site detection.

In this chapter, we will first summarize the principle of CLIP and the different protocol variants. We will highlight the main steps of data analysis and outline the available tools. We will then present an overview of how CLIP data is actually used in the literature, then focus on how the community assesses its reproducibility. Finally, we will conclude this chapter with a summary of the knowledge obtained with CLIP protocols regarding the EJC.

3.1 The essential steps of CLIP

Although multiple experimental variations of CLIP exist, 4 main steps define the essence of the protocol:

1. Irradiating live cells (or tissues) with ultra-violet (UV) light to specifically create covalent links between proteins and RNA.
2. Lysing cells and purifying the RBP of interest with a specific antibody.
3. Digesting proteins with proteinase K and reverse transcribing RNA fragments to cDNA.
4. Preparing the sequencing library (by PCR amplification and size selection).

There are two main challenges in CLIP: a) the crosslinking efficiency, and b) the IP efficiency. On one hand, 1 to 5% of protein molecules are crosslinked to RNA with UV-light radiation (Darnell 2010). For RBPs targeting mRNA, obtaining CLIP libraries is much more challenging, as mRNAs comprise between 1 and 5% of the total RNA in a cell. As for IP, antibodies need both high specificity and affinity for the protein of interest. These characteristics are crucial to separate the desired protein-RNA interactions from the vast molecular entanglement in the lysate. The

success of a CLIP experiment is thus defined by: the interaction between the protein and its target (one that facilitates crosslink), the abundance of the target, and the quality and availability of antibodies.

3.2 A brief history of CLIP

In the early 2000's, the first attempts to identify protein-RNA interactions in vivo were taking place. The first transcriptome-wide protocol was published in the year 2000: RNA IP coupled with microarrays, RIP-Chip (Tenenbaum et al. 2000). It consisted in performing IP against the RBP of interest in native conditions, followed by protein digestion, purification of RNA fragments, reverse transcription (RT), and PCR amplification (Fig. 3.1a). The authors claimed that skipping crosslink and performing native was sufficient to obtain a library of binding sites, while avoiding sequence bias and background noise introduced by crosslink (Keene, Komisarow, and Friedersdorf 2006). However, subsequent studies soon revealed that background noise was higher in RIP-Chip and -seq experiments than in libraries obtained with CLIP (Darnell 2010). In native conditions, protein-RNA interaction must be sufficiently strong to endure extensive washes. Moreover, artificial interactions may spontaneously happen *in vitro* after cell lysis, promoting the purification of non-specific interactions (Lee and Ule 2018; Fig. (Fig. 3.1b). The focus of the community then shifted towards the application and improvement of the CLIP protocol.

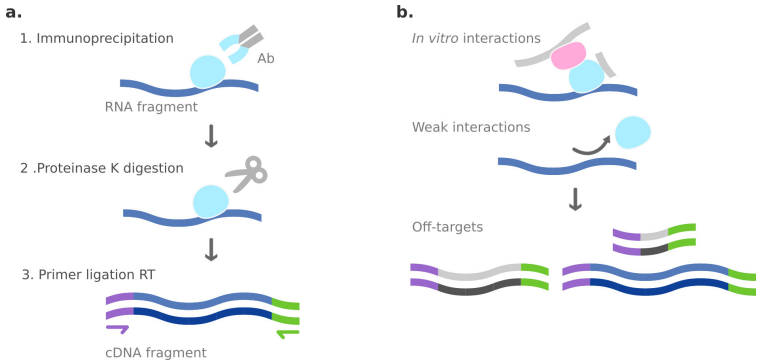

Figure 3.1: **a.** Main steps of the RIP protocol. **b.** Main caveats of the RIP protocol due to native IP. RT: reverse transcription. Ab: antibody. Theoretical source: Keene, Komisarow, and Friedersdorf 2006

The first attempts of *in vivo* crosslink to identify protein and nucleic acid (NA) interaction involved the use of formaldehyde (Ule, K. Jensen, et al. 2005). In addition to inducing protein-NA covalent bonds, formaldehyde creates protein-protein bonds. This forms macromolecular complexes involving the protein of interest, their NA targets, and interacting protein factors that may in turn bind other NA elements. As a result, the signal-to-noise ratio in the final library is low, which is particularly inconvenient for low-abundance RBPs. Irradiation with ultra-violet (UV) light overcomes this caveat by inducing covalent bonds between proteins in direct contact with RNA. At an irradiation of 254 nm, nucleotide bases (especially pyrimidines C and U) are photoreactive and link mainly to cystein, lysine, phenylalanine, tryptophan, and tyrosine protein residues. Using UV-light irradiation avoids protein-protein crosslinking, increasing signal-to-noise ratio. Thus, it offers a snapshot of

direct protein-RNA interactions *in vivo*. However, the efficiency of UV crosslinking is highly dependent on how the protein interacts with RNA, and the distance and position between crosslinkable protein residues and RNA bases. This results in highly variable and unpredictable crosslink efficiency.

Although the first published CLIP protocol provided functional insight into the protein Nova in the mouse brain, the power of sequencing was still limited to a few hundred reads (Ule, K. B. Jensen, et al. 2003). High-throughput protocols flourished during the second half of the 2000's and the early 2010's. In the first version, designated HITS-CLIP (high-throughput sequencing and CLIP) cell culture was irradiated with UV-B light to create covalent bonds between protein and RNA (Darnell 2010). crosslinking has two main advantages. First, proper RNase digestion conditions allow to obtain relatively short RNA fragments (approximately 100 nt). Second, stringent purification conditions remove the majority of non-specific interactions. Next, SDS-PAGE migration and nitrocellulose protein-RNA transfer allow to separate covalently bound protein-RNA complexes from non-crosslinked RNA, which reduces background noise; non-specific RNA may thus correspond to fragments crosslinked to other proteins. Subsequent steps consist in RNA radiolabeling to visualize protein-RNA complexes, proteinase K digestion, primer ligation, and RT (Fig. 3.2).

However, crosslink efficiency is estimated to be around 1 to 5%. The authors thus proposed PCR amplification as compensation to obtain enough material for sequencing. Although this strategy does increase the amount of material, it does not increase the complexity of the library—i.e. the representation of unique RNA fragments in the library. Moreover, abundant small non-coding RNAs highly increase the background noise in CLIP libraries, and represent many of the reads after sequencing Eric L. Van Nostrand et al. 2016.

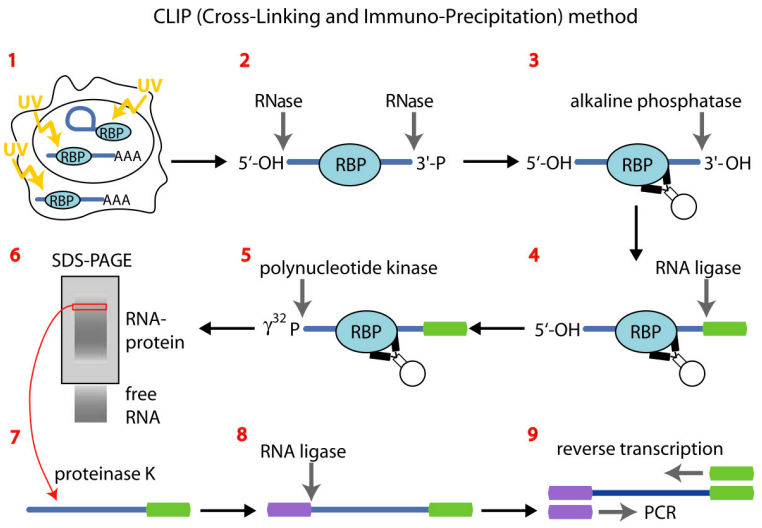

Figure 3.2: **1** UV light irradiation on live cells. **2** RNase digestion. RNA fragments are protected by the bound protein. **3** & **4** IP, dephosphorylation and 3'-primer ligation. **5** 5' radiolabeling for protein-RNA complex visualization. **6** Denaturing gel separation, nitrocellulose transfer and excision of protein-RNA complexes. **7** Proteinase K digestion. **8** 3'-primer ligation. **9** RT and PCR amplification. Adapted from: Ule, K. Jensen, et al. 2005

An attempt to increase crosslinking efficiency was introduced with Photoactivatable Ribonucleoside-enhanced CLIP (PAR-CLIP) (Danan, Manickavel, and Hafner

2016). Newly transcribed RNA is labeled by incubating the starting cell culture with ribonucleoside analog 4-thiouridine (4SU) or 6-thioguanine (6SG). Upon irradiation with UV-A or UV-B light, the labeled transcripts form covalent bonds with RBPs more efficiently than the unlabeled counterparts. Additionally, RT introduces a distinctive nucleotide transition that can be detected during data analysis (which will be detailed in the next section). Subsequent library preparation steps are essentially the same as HITS-CLIP. However, nucleoside analog concentration needs to be optimized to minimize cellular toxicity, and an additional step is required to assess the efficiency of incorporation. Moreover, PAR-CLIP, as does HITS-CLIP, overlooks a step that significantly decreases library yield.

Prior to RT, proteinase K digestion leaves a lingering fragment of protein at the crosslinking site. This creates a steric impediment for the RTase, resulting in *truncated* cDNA fragments lacking the template for the 5'-PCR primer. Thus, only the fragments that are read through the crosslinking site (designated *read-through* fragments) can be amplified and sequenced. It was estimated that this caused the loss of the majority of cDNA fragments (Fig. 3.3a, G. Martin and Zavolan 2016). Introducing individual-nucleotide resolution CLIP (iCLIP) was therefore a major tipping point in CLIP protocols (König et al. 2010). Instead of ligating primers at each end of RNA fragments, only one ligation is performed at the 3'-end (3.3b). The ligated primer is a composite of both 5'- and 3'-end PCR adaptors. Following RT, both truncated and read-through cDNAs are circularized, then cleaved at the composite primer to obtain a construct with primers at both ends. This strategy not only rescues truncated cDNAs, but offers the location of the exact crosslinking site for a large fraction of the fragments. iCLIP was thus the precedent for what are now considered single-nucleotide resolution CLIP protocols.

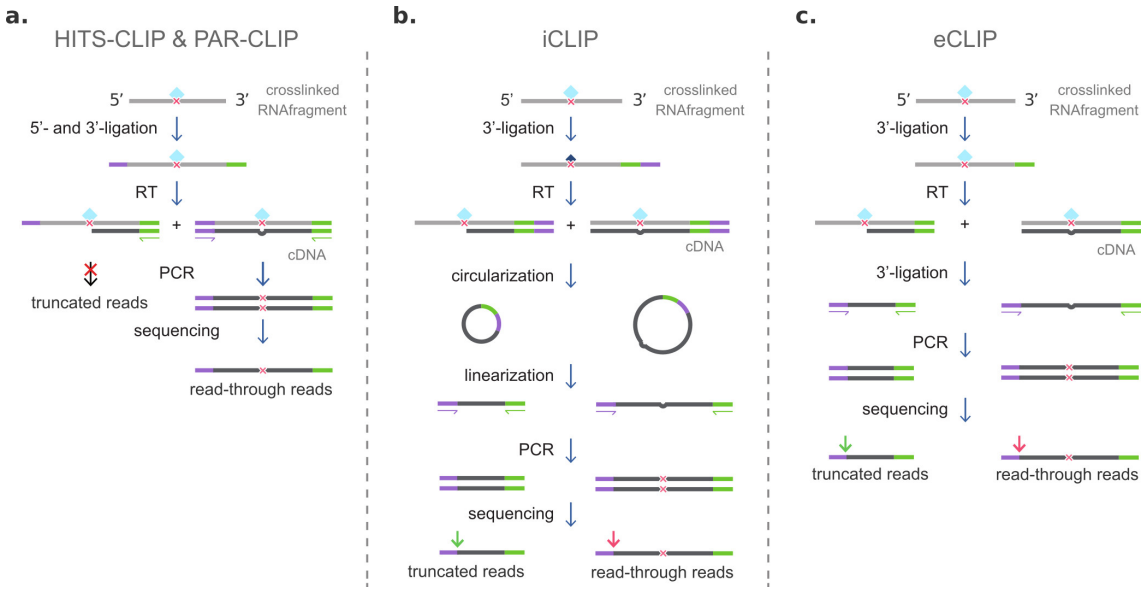

Figure 3.3: **a.** For HITS-CLIP/PAR-CLIP, adaptors (purple and green) are ligated at both extremities of the crosslinked RNA fragments. Only read-through cDNAs can be amplified by PCR using primers complementary to adaptors generating read-through reads. **b.** For iCLIP, a single bipartite adaptor is ligated at the 3' extremity of the crosslinked RNA fragments. Full-length or truncated cDNAs are circularized and then linearized leading to the presence of adaptors at both extremities. PCR amplifies both truncated and read-through reads. **c.** For eCLIP, a single adaptor (green) is ligated at the 3' extremity of the crosslinked RNA fragments. After RT, a second adaptor (purple) is ligated to the 3' extremity of the cDNAs. PCR amplifies both truncated and read-through reads. The green arrows indicate the position of the crosslinking site. The red arrows indicate the 3' extremity of the cDNA, upstream the crosslinking site. Adapted from: Hocq et al. 2018

Subsequent variations aimed at further optimizing the iCLIP protocol. BrdU-CLIP incorporates the nucleotide analog Br-dUTP at the RT step to perform an additional IP and remove potential background noise (Weyn-Vanhentenryck et al. 2014). Infrared iCLIP (irCLIP, Zarnegar et al. 2016) replaces radioactive isotopes with infrared tags to visualize protein-RNA complexes after crosslink. These two protocols share the circularization approach of iCLIP to capture truncated cDNAs, which in turn reduces the overall yield of material (Hocq et al. 2018).

With enhanced-CLIP (eCLIP), Van Nostrand and colleagues introduced a library preparation strategy to avoid circularization while rescuing truncated cDNAs (Eric L. Van Nostrand et al. 2016). A first ligation adds the 3' PCR primer to perform RT. The 5' PCR primer is ligated to the resulting cDNA, allowing amplification of both truncated and read-through cDNAs (3.3c). The authors introduced sized-matched (SM)-input as an important control to detect non-specific signal in eCLIP data sets, which proved to be more sensitive than other IP controls such as IgG, or plain transcript abundance obtained with RNA-seq. We will detail how SM-input is considered in the following section. Importantly, they performed eCLIP over hundreds of RBPs, establishing the first ENCODE resource for protein-RNA interactions.

Overall, we have observed a true evolution of CLIP over the years. From efficiency optimization to increase of resolution, the effort of the community has produced multiple CLIP variants. Summarizing all of the existing protocols is beyond the scope of this work, but have been exhaustively reviewed by Lee and Ule 2018. With the diversity of protocols, the choice depends on the desired objectives and the tools available to the researcher interested in capturing transcriptome-wide RBP interactions.

3.3 Mining CLIP-seq data

The advent of different library construction methods was matched with a diversity of data analysis tools. Despite their availability, the community soon faced the challenge of CLIP data analysis. Naively compared to its DNA counterpart, ChIP-seq, CLIP data proved to be more complex in terms of statistically significant binding site detection. In response, several bench-marking articles and reviews appeared in the literature, especially in the second half of the 2010's. In this section, we will present an outline of the data analysis steps, followed by an overview of peak detection methods. Then we will present a few examples of how these are used in the literature, and comment how reproducibility is addressed. Finally, we will address the knowledge obtained from EJC CLIP-seq data prior to this work.

3.3.1 Main steps of data analysis

Over the past decade, several reviews have listed the main aspects of data analysis and overview the available tools best adapted to the CLIP variant of choice (Reyes-Herrera and Ficarra 2014; Liu et al. 2015; Bottini, Hamouda-Tekaya, et al. 2017; De 2018; Ule, Hwang, and Darnell 2018; Chakrabarti et al. 2018). Often, recommendations list the following main

steps: aligning to the reference genome; identifying and removing PCR duplicates; drawing a meta-analysis plot and/or detecting binding sites, and finally discover sequence motifs.

Data pre-processing in a nutshell

CLIP-seq data needs to be pre-processed for quality assessment and bias minimization before binding site detection and analysis. Assessing raw sequencing data with FastQC (Andrews 2010) is a widespread practice in the community. We will therefore focus on PCR duplicate removal as an important step in data pre-processing.

PCR duplicate removal can be performed using mapping coordinates alone or with a unique molecular identifier (UMI), and should in either case minimize PCR amplification bias. PCR duplicate detection is particularly useful to assess the complexity of a CLIP library. Plotting the number of non-duplicate reads obtained from increasing fractions of uniquely mapped reads yields a library complexity curve (Fig. 3.4). Similar to a saturation curve, a library complexity curve shows the minimal amount of reads necessary to obtain most non-PCR duplicate reads from the library. A library that reaches a plateau with a few uniquely mapped reads is less complex than a library that has not reached a plateau with several millions of uniquely mapped reads.

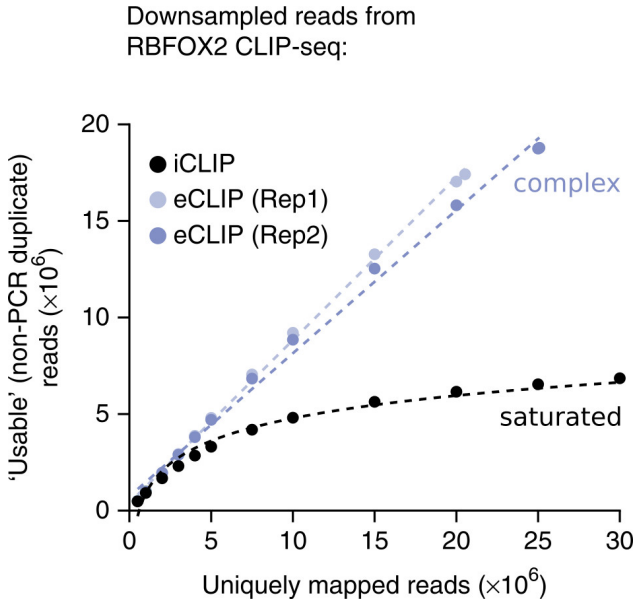

Figure 3.4: A comparison of PCR duplication levels of one iCLIP library and two eCLIP libraries of the RBFOX2 protein. Several fractions of each library are randomly sampled and undergo PCR removal. Thus each uniquely mapped value yields a corresponding "usable" read count (non-PCR duplicates). Adapted from: Eric L. Van Nostrand et al. 2016

A landscape in a single graph: meta-analysis plots

Meta-analysis plots are, primarily, a quality assessment tool. They represent the distribution of read counts relative to a particular region of the gene (for instance

the exon-intron junction). Although they require prior knowledge of the protein of interest, it gives important information about the signal in the expected binding site of the RBP, in terms of enrichment and precision.

A use case of meta-analysis plots is presented in a study on the impact of read length on binding site assignment (Hauer, Curk, et al. 2015). The authors represent the density of iCLIP signal from several RBPs relative to the exon-exon, or exon-intron junctions (see an example in Fig. 3.5a). They showed that the 5'-end signal of longer reads was shifted upstream, which could potentially bias the precision of binding site assignment. A posterior study argued that optimizing RNase digestion conditions minimizes the cDNA length bias (Haberman et al. 2017).

Another use case of meta-analyses consist in assessing the distribution of peaks relative to a known binding site, instead of read counts (Krakau, Richard, and Marsico 2017; Chakrabarti et al. 2018; Yee et al. 2018). These representations are useful to assess the quality of peak detection, as the majority of peaks should locate around the expected binding site (Fig. 3.5b).

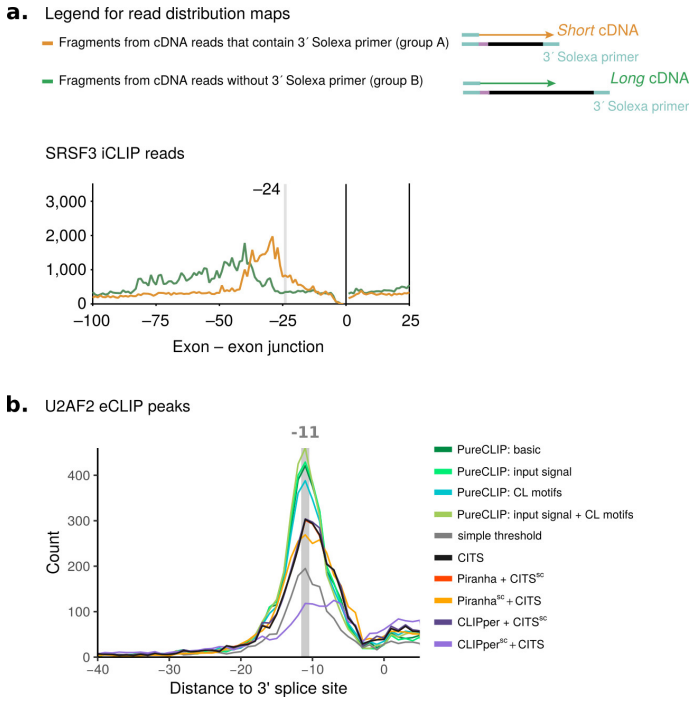

Figure 3.5: **a.** Comparison between the distribution of two different populations of SRSF3 iCLIP reads relative to the exon junction. **b.** Comparison of U2AF2 eCLIP peaks obtained with different peak detection strategies relative to the 3'-splice site. Adapted from: Hauer, Curk, et al. 2015 (a) and Krakau, Richard, and Marsico 2017 (b).

In summary, meta-analysis plots are a valuable representation of CLIP data. They are a qualitative tool to compare libraries of the same protein obtained with different strategies, or to assess the performance of different peak detection strategies. However, it should be noted that they paint a global picture of the data. Because they aggregate the whole transcriptome signal, they do not provide individual binding-site information.

Spell it out: motif discovery

Motif discovery is performed after binding site detection (discussed in section **??**). It consists in identifying short sequences enriched near or within the binding sites of the protein of interest. Algorithms such as DREME, from the MEME suite (Bailey 2011), and HOMER (Heinz et al. 2010) are widely applied on CLIP peaks, despite being originally developed for chromatin IP (ChIP)-seq data. These tools generally compare the sequences within or in neighboring regions of a peak to a background, and report the statistically significant words or motifs (Chakrabarti et al. 2018). In early HITS-CLIP experiments, this strategy resulted in the discovery of previously unknown binding motifs of several RBPs (Darnell 2010). Another application of characterizing and mapping binding motifs is representing the distribution of reads relative to the motif, known as a meta-analysis plot (T. Wang, Xie, and Xiao 2014; Krakau, Richard, and Marsico 2017). Conversely, the motif density relative to the peaks assesses the spatial distribution of enriched sequences relative to binding sites (T. Wang, Xie, and Xiao 2014; Krakau, Richard, and Marsico 2017; Bottini, Hamouda-Tekaya, et al. 2017; Haberman et al. 2017). Interestingly, a large-scale study using binding sites of 78 RBPs discovered that binding motifs tend to be repetitive low-complexity sequences (Dominguez et al. 2018). However, the functional diversity of RBPs requires different binding modes that range from sequence specific, to context-dependent, to sequence-independent (Hentze et al. 2018). Thus, the relevance of motif discovery is determined by prior knowledge of the RBP of interest, and whether its binding is sequence-dependent or influenced by other RBPs. As the EJC binds RNA in a sequence-independent manner, motif discovery around EJC peaks may reveal binding sites of peripheral factors or other interacting RBPs.

3.3.2 CLIP peak discovery

From the data analysis point of view, the goal of the CLIP protocol is to obtain genomic regions that are significantly enriched with read signal. These regions are designated as peaks. This description may lead to an attempt to compare CLIP-seq peaks to ChIP-seq peaks. After all, widespread tools with robust underlying statistical frameworks, and extensive consortium guidelines exist for the latter (Y. Zhang et al. 2008; Bailey 2011; Heinz et al. 2010; Landt et al. 2012; Nguyen et al. 2018), tempting the data analyst to apply these tools on CLIP-seq data. However, dealing with transcriptome-generated data drastically changes the framework of data analysis. Direct extrapolation of ChIP peak detection tools is prevented by the heterogeneous background signal related to variable transcript abundance, as well as the discontinuous gene coverage due to splicing. This explains the development of dedicated tools for CLIP that took place almost in parallel to the emergence of CLIP protocols.

Several peak detection tools have been reviewed over the years by several authors from different teams (Reyes-herrera and Ficarra 2014; Liu et al. 2015; T. Wang, Xiao, et al. 2015; Uhl et al. 2017). Yet, a recent review by Chakrabarti and colleagues summarizes with detail the currently existing tools for peak detection (Chakrabarti et al. 2018). We can divide peak callers according to the resolution of peak detection, i.e. the size of the reported binding sites. Further distinctions may be applied to separate peak

callers according to how the read coverage are used, or the statistical framework to assess significance. Here, we will focus first on the resolution of binding sites.

Broad binding site detection

Several peak calling methods detect broader enrichment regions, therefore providing lower-resolution results. These methods are often coverage-based, using the total length of reads and ignoring truncation or mutation sites. In a sense, they can be applied on any data set independently of library construction protocol. However, applying them on single-nucleotide resolution libraries implies loss of binding-site assignment precision.

Piranha divides the genome into bins of user-defined size, then computes a p-value for the number of reads in each bin using a zero-truncated negative binomial distribution (Uren et al. 2012); consecutive significant bins are merged into significantly enriched regions. The method used by dCLIP is also based on a binning approach to analyze read signal, but it is designed to compare it across different biological or experimental conditions (T. Wang, Xie, and Xiao 2014). It implements MA (M: logarithmic ratio, A: average value) transformation—an approached originally designed for micro-array signal—to scale and compare the signal from different data sets. Then a hidden Markov model (HMM) is used to infer differentially enriched regions. Finally, CLIPper assesses coverage at individual nucleotide locations (Lovci et al. 2013). Similar to the CTK and iCounts methods, it establishes an empirical null distribution by shuffling read positions within a gene then computes a FDR value of the observed counts per position. Contiguous locations with significant read coverage are merged and broad binding sites are reported. It should be noted that CLIPper was designed for eCLIP data analysis, but does not provide single-nucleotide resolution.

High resolution peak callers

To detect crosslinking sites with high resolution, peak callers detect either truncations or mutations. Therefore, these methods can only be applied on CLIP protocols that: a) contains crosslinking induced mutations—HITS-CLIP, and PAR-CLIP, or b) include cDNA fragments generated by RT truncation at the crosslinking site: iCLIP, irCLIP, BrdU-CLIP.

Because PAR-CLIP induces a particular T to C transition, dedicated methods have been developed for crosslinking site detection. After clustering overlapping reads, PARalyzer uses kernel density smoothing to estimate two separate events: T to C transitions and non-transitions (Corcoran et al. 2011). Locations where T to C density is higher than non-transitions, as well as above a minimum number of reads, are reported as significant crosslinking sites. Since defining an arbitrary threshold of minimum reads may reduce sensitivity, wavClusteR (Sievers et al. 2012), implements a Bayesian network that computes the probability of observing a particular count of T to C transitions assuming they are caused by crosslink. They then compare this to the probability of observing the same counts assuming they were not crosslinking

induced and report positions more likely to be crosslinking induced than otherwise. Another Bayesian-based method, PAR-CLIP HMM, combines coverage information with T to C transition occurrence to report the most likely crosslinking sites Yun, T. Wang, and Xiao 2014.

PIPE-CLIP (Chen et al. 2014), is a hybrid method that combines broad peak detection with a mutation-based approach. First, a zero-truncated negative binomial regression model is fitted to detect significantly enriched clusters of reads. Then crosslinking sites within significant clusters are detected by computing the number of mutations and assessing their significance with a binomial test. The authors claim that the occurrence of truncations can also be computed to detect crosslinking sites, making PIPE-CLIP suitable for iCLIP data analysis. Similarly CIMS (crosslinking induced mutations) and CITS (crosslinking induced truncations), from the CLIP data toolkit CTK (Shah et al. 2017), compute the number of mutations and truncations at particular genomic locations. To test their significance, they compute an empirical null distribution by shuffling read positions within a gene and computing the number of mutations in the shuffled distribution. A similar approach is implemented by the iCounts tool (Chakrabarti et al. 2018), where the shuffled distribution is established in a user-defined region rather than the whole gene. Thus, these truncation-based methods propose an empirical null distribution to test the significance of crosslinking-sites.

On the other hand, PureCLIP proposes a Bayesian framework to detect crosslinking induced truncation events (Krakau, Richard, and Marsico 2017) in iCLIP, eCLIP, and derivatives. The number of read 5'-ends at a particular genomic position is combined with the coverage information of the surrounding region—which is modeled with kernel density estimates. Then, PureCLIP infers the probability of observing a particular number of truncations and surrounding read coverage for four possible states: 1) the position is neither enriched or crosslinked; 2) the position is not enriched but it is crosslinked; 3) the position is enriched but not crosslinked, and 4) the position is enriched and crosslinked (Fig. 3.6a). Only single-nucleotide locations with the highest probability of being in state 4 are reported. Optionally, neighboring positions can be merged if they are within a user-defined distance. In addition, SM-input controls and regions prone to non-specific crosslinking can be included as co-variates to report specific crosslinking sites (Fig. 3.6b-c). Thus, PureCLIP models the CLIP signal and the truncation events to report single-nucleotide crosslinking sites.

In summary, the choice of peak detection tool will be highly influenced by the library construction protocol. In turn, this depends on the ultimate objective of the study. If precise location of the binding sites is preferred, single-nucleotide protocols and crosslink detectors should be prioritized. If, on the other hand, resolution is secondary to the study, broad-peak callers may be sufficient.

3.3.3 CLIP data in the literature

The outcomes of CLIP-seq data analysis tend to adhere to the data analysis steps described in the previous section. First, CLIP signal is aggregated and represented

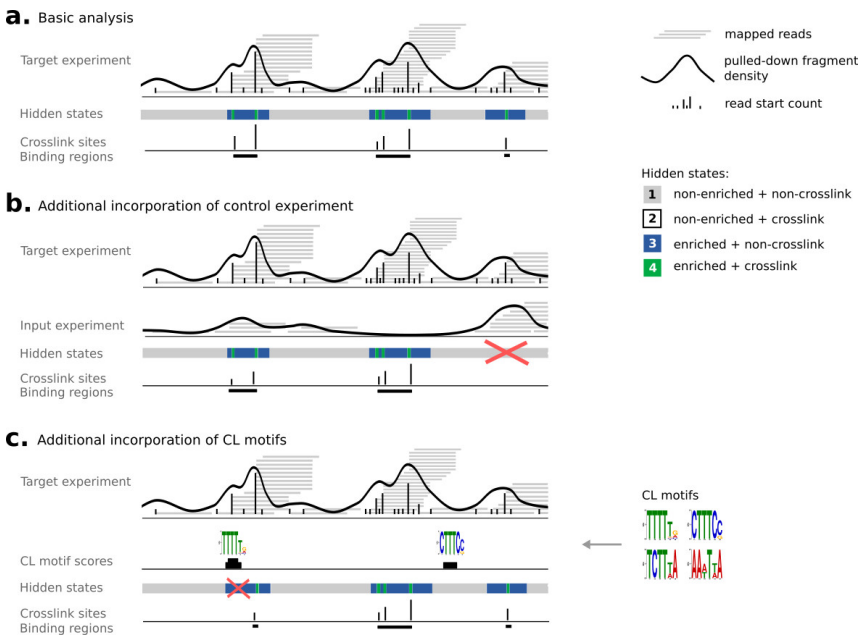

Figure 3.6: **a.** Detection of crosslinking sites without co-variates. **b.** Detection of crosslinking sites using SM-input signal information as co-variate. **c.** Detection of crosslinking sites using non-specific crosslink motifs as co-variate. Adapted from Krakau, Richard, and Marsico 2017

as a meta-analysis plot. Next, binding sites are detected, and some individual examples are shown as genome browser tracks. Finally, if it is relevant for the RBP under study, motif enrichment results are depicted as Logo representations. In this section I will present some examples of how CLIP data has been utilized for the study of RBPs in the literature.

Binding site detection is essential for RBP target discovery. Alternative splicing events regulated by the Nova splicing factor were discovered with iCLIP peak detection (Ule, K. B. Jensen, et al. 2003; Ule, Stefani, et al. 2006). Similarly, the role of hnRNP particles in alternative splicing was revealed by crosslinking site enrichment in excluded exons (König et al. 2010). Ago2 CLIP peaks revealed biologically relevant miRNA targets validating computationally predicted targets (Ule, Hwang, and Darnell 2018; Bottini, Pratella, et al. 2018). iCLIP experiments of two SR proteins (SRSF3 and SRSF4) revealed subsets of mRNA targeted by each protein; interestingly, SRSF3-mediated splicing events resulted in mRNA down-regulation through NMD (Änkö et al. 2012). Hence, individual binding sites can reveal important roles in the regulation of particular genes. Notably, the high-resolution data allows to correlate binding site positioning to specific functional effects, such as exon exclusion or miRNA targeting.

The massive eCLIP experiments data sets by the ENCODE consortium is a valuable resource for the characterization of a diverse array of RBPs. For instance, binding sites of over 120 RBPs overlapping miR loci revealed RBP regulation of miRNA transcripts (Nussbacher and Yeo 2018). Another study analyzed transcriptome-wide binding sites of over 150 RBPs, giving major insights on RNA processing (Eric L Van Nostrand et al. 2019). Due to the complexity of the data, the authors opted to use peak density meta-analyses to present their discoveries. For instance, they represented the RBP binding profiles as the aggregation of binding sites along all detected genes. RBPs were then clustered according to their binding profile along

the transcript. The distinct mRNA processing functions were then identified within each cluster.

The last two examples show that aggregating binding site information across different eCLIP experiments may reveal RBP regulation modalities. It should be noted that these studies used broad binding site information aggregated into meta-analyses representations. Despite their value, they do not take into account the reproducibility of individual binding sites, and do not take advantage of the high-resolution provided by eCLIP data.

3.4 Assessing reproducibility of CLIP-seq data

Reproducibility is a fundamental principle of experimental sciences, regardless of the discipline. Consider a simple experiment where we aim to determine whether a coin is not fair. We expect that it is equally likely to obtain heads or tails after tossing it. If the coin were tossed once and we were to obtain heads, we cannot state that the coin is biased towards heads (Fig. 3.7a). Thus, we need to toss the coin a certain number of times (N) and counting the fraction of heads observed. Yet, because tossing the coin is a stochastic process, we may obtain a fraction of heads that is not exactly 0.50, even if the coin is indeed fair (Fig. 3.7b). To claim that the observed fraction is due to a coin bias and not to stochastic noise, we ought to repeat the coin tossing N times. If after several iterations we consistently obtain fractions of heads that are far from the expected 0.50, we can conclude with confidence that the coin is biased (Fig. 3.7c). This would not be possible if we did not *reproducibly* observed skewed fractions when repeating the experiment several times.

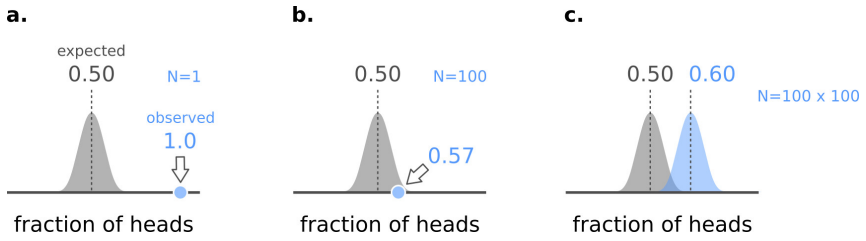

Figure 3.7: **a.** Tossing the coin one time (N=1) and observing *heads* does not prove that the coin is biased. **b.** Tossing the coin multiple times (N=100) yields an observed fraction of heads close to the expected value; we cannot conclude whether the coin is biased because of statistical noise. **c.** Repeating the experiment in **b** multiple times (100 tosses repeated 100 times) yields a distribution of observed values centered around 0.60, indicating that the coin is indeed biased. In blue: observed fraction of heads from the total times the coin was tossed; in gray: distribution of expected values centered around the expected value of 0.50.

For this reason, technical and biological replicates are essential to distinguish biologically relevant events from randomly detected noise and natural variability. In the context of high-throughput sequencing, and more specifically in RNA-seq differential expression studies, signal is typically aggregated within genes or exons, followed by RPKM (reads per kilobase per million) or FPKM (fragments per kilobase per million) normalization (Mortazavi et al. 2008; SEQC 2014). The reproducibility of the signal is then assessed as the read count correlation between replicates (Fig. 3.8). With current sequencing technologies, correlation values generally approach 1.0 in successful RNA-seq libraries. This indicates that the abundances of transcripts can be reproducibly estimated with RNA-seq data.

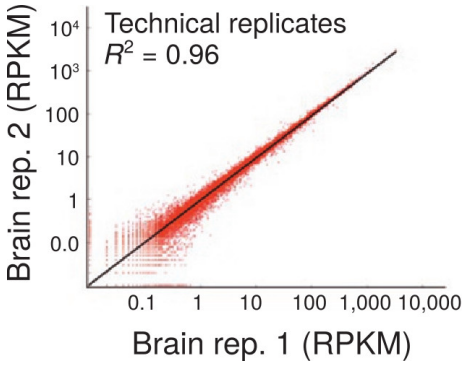

Figure 3.8: Scatter plot of RPKM values of genes in two different technical replicates of mouse brain mRNA-seq experiments. Adapted from Mortazavi et al. 2008

Sound inference of RBP roles and mechanisms relies on reproducible observations. Thus, reproducibility is also crucial in CLIP-seq data analysis. However, reproducibility assessment is not as straightforward as for other NGS data. This is especially true when the goal is to detect relatively small enriched regions, rather than quantifying the signal over hundreds of base pairs. Therefore, the RBP community interested in CLIP has hardly reached consensus on how to assess CLIP-seq data reproducibility.

3.4.1 The recommendations from the community

The ENCODE consortium adopted data quality guidelines for CLIP-seq experiments, highly inspired by ChIP-seq data quality standards (https://www.encodeproject.org/eclip/). These guidelines incorporate the irreproducibility discovery rate (IDR) as a valuable reproducibility assessment tool. It is a statistical approach to evaluate the consistency of detection in high-throughput experiments (Li et al. 2011). Binding sites detected in two separate replicates are ranked either by p-value or the score given by the peak detection tool of choice. The authors distinguish two classes of binding sites: reproducible and irreproducible, according to the distribution of their rank correspondence (Fig. 3.9a). The IDR value is the probability of a pair of peaks belonging to the irreproducible class. It is computed by modeling the marginal distributions of peak ranks with mixture models. Similar to selection of significant peaks, it suffices to define an irreproducibility tolerance threshold α to select reproducible peaks (IDR $< \alpha$). The authors of IDR present the comparison of ChIP-seq peak callers as a use case (Fig. 3.9b).

In practice, flatter IDR curves indicate a higher number of reproducible peaks, hence a better performance of the tool. As part of the ENCODE project, Van Nostrand and colleagues use IDR curves to compare peak reproducibility between iCLIP and eCLIP data (Eric L. Van Nostrand et al. 2016). Nevertheless, they reported the analysis on only two of 73 data sets. The use of IDR is recommended as well by Chakrabarti and colleagues to assess binding site reproducibility (Chakrabarti et al. 2018). However, the use of the IDR is not widely spread in the RBP community, and is rarely present in CLIP publications.

Finally, a general recommendation when dealing with RBP binding site re-

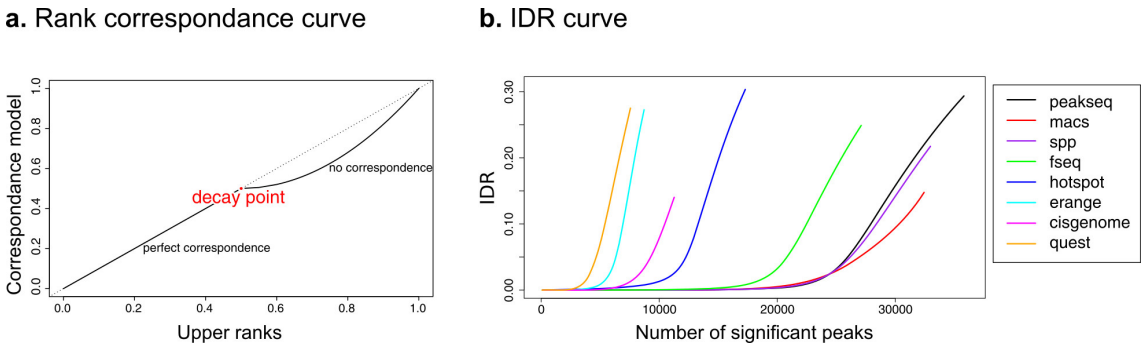

Figure 3.9: Scatter plot of RPKM values of genes in two different technical replicates of mouse brain mRNA-seq experiments. Adapted from Li et al. 2011

producibility consists in union or intersection of technical replicates (Ule, Hwang, and Darnell 2018; Chakrabarti et al. 2018). The choice will depend on whether we favor sensitivity (the number of discovered binding sites), or the specificity (the reliability of binding sites). For higher sensitivity, some authors recommend the union of binding site sets from different replicates, whereas for specificity, the intersection of binding site sets are recommended (Fig. 3.10).

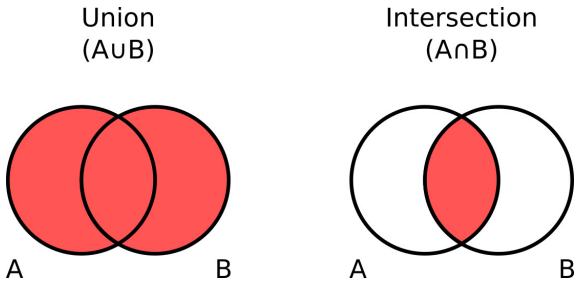

Figure 3.10: Graphical representation of the union (left), and the intersection (right) of two sets of data.

The Jaccard index

The Jaccard index (Equation 10.1) indicates the level of overlap between two different sets, where 0 corresponds to 0 common elements, and 1 indicates a perfect overlap:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{3.1}$$

Equation 3.1: The Jaccard index (J) between two sets (A and B) corresponds to the ratio between the size of the intersection ($|A \cap B|$) over the size of the union ($|A \cup B|$).

This metric can be applied to quantify the level of overlap between the peaks detected between different replicates.

3.4.2 More examples from the literature

Other strategies have been proposed to estimate the reproducibility of CLIP-seq studies. König and colleagues computed the distance between the nearest read

truncation sites in iCLIP replicates (König et al. 2010). The distribution of these distances revealed that reads were frequently located at the exact genomic location in both replicates (Fig. 3.11a). While this assesses the similarity of "raw" iCLIP signal, it disregards whether the same binding sites are reproducibly detected across replicates.

A posterior publication compared the reads within crosslinking sites of two RBPs (hnRNP C and U2AF65) detected on two replicates each (Zarnack et al. 2013). Although the reproducibility is higher for peaks with the strongest signal, Spearman coefficient values, ranging from 0.33 to 0.48, show a moderate global reproducibility (Fig. 3.11b). For subsequent analyses, the authors merged replicates claiming a high Jaccard index of detected peaks: 0.78 for hnRNP C, and 0.59 for U2AF65 (from a total of 438,360 and 518,794, respectively).

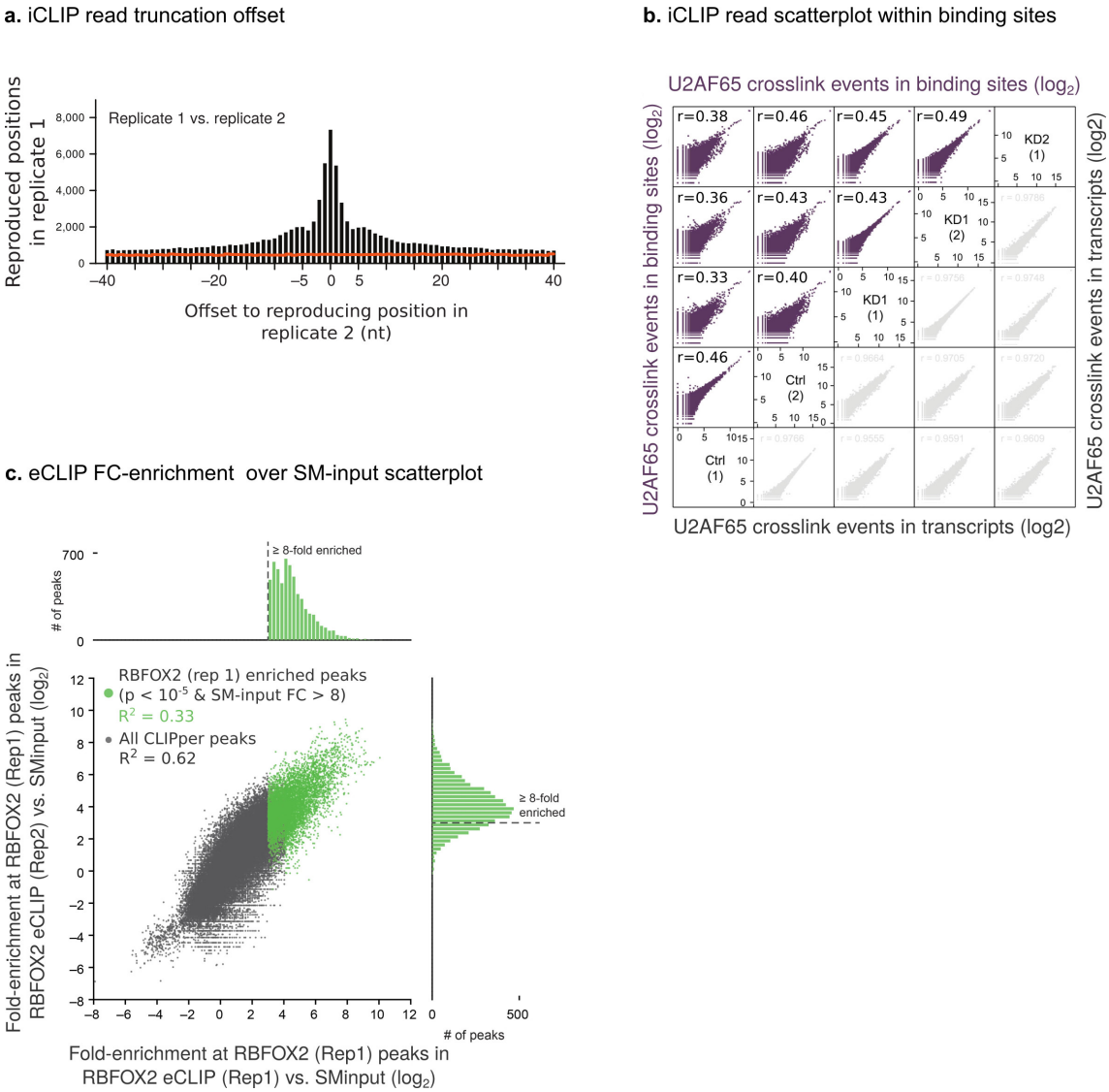

Figure 3.11: **a.** The distribution of the distance between the 5'-end of reads one replicate and the nearest 5'-end in the second replicate (designated offset of reproducing position by the authros). **b.** Number of read 5'-ends within iCLIP binding sites (purple) or within the whole transcript (gray). **c.** Scatter plot of SM-input enrichment of replicates 1 and 2 within the binding sites detected in replicate 1; in green: binding sites with SM-input enrichment ¿ 8. Adapted from König et al. 2010; Zarnack et al. 2013; Eric L. Van Nostrand et al. 2016

The authors of eCLIP proposed a strategy where instead of comparing read

truncations within binding sites, they compute fold-change enrichment of the signal within binding sites over the SM-input (Eric L. Van Nostrand et al. 2016). Yet, to compute SM-input enrichment, the authors use the coordinates of the peaks detected only in one replicate (rep. 1), then use the signal of the other replicate (rep. 2) within these binding sites (Fig. 3.11c). Moreover, in their scatter plot they present not significantly enriched (in grey) along the significantly enriched binding sites (in green), thus creating the illusion of highly reproducible data. Yet, coefficient correlation values of significantly enriched peaks show a rather moderate correlation. Technically, this strategy assesses the reproducibility of the signal itself rather than the binding site detection, as only peaks from one replicate are employed.

The PureCLIP authors assess the reproducibility of peak detection results with a ranking strategy (Krakau, Richard, and Marsico 2017). First, only peaks with identical genomic locations in two replicates are considered as commonly detected. Then, they define the top x percentile by ranking peaks from highest to lowest scores. Next, they compute the percentage of peaks in the top percentile of one replicate that are also detected in the top percentile of the second replicate. They called this value the percentage of agreement between two replicates. This value allows to assess the effect of the peak score on reproducibility, by computing the replicate agreement for several percentiles. Interestingly, the agreement for PUM2, RBFOX2, and U2AF2 eCLIP peaks reached maximum levels around 35%, 25%, and 30% respectively. However, when the whole set of peaks was considered, the agreement curves plateau at approximately 20% for PUM2 and RBFOX2, and 30% for U2AF2 (Fig. 3.12). This indicates that high-resolution binding sites show a relatively low level of reproducibility.

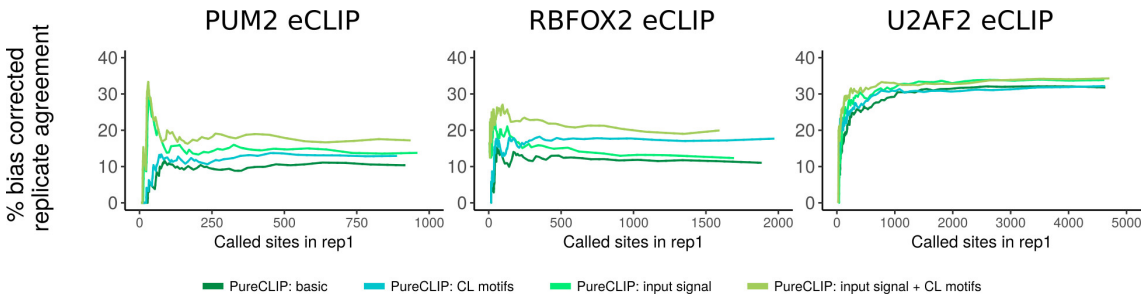

Figure 3.12: Agreement of called sites between replicates. For each eCLIP data set, the authors report for each given number of called sites x in replicate 1 (corresponding to a certain p value or score threshold), the percentage that were also called within the top x ranking sites in replicate 2, after correcting for crosslinking bias. The leftmost point of each curve corresponds to the number of calls associated with the lowest p value or highest score the strategy can report. rep1 replicate 1. Adapted from Krakau, Richard, and Marsico 2017

Overall, there are currently multiple strategies to assess the reproducibility of CLIP binding sites. However, there is yet no consensus that allows researchers to accurately assess the reproducibility of CLIP-seq binding sites, especially from single-nucleotide resolution protocols.

3.5 Learning from EJC binding site data

Thus far, I have presented CLIP protocols as a strategy to study the transcriptome-wide binding modalities of RBPs, and highlighted the main approaches and chal-

lenges in data analysis. In this section I will summarize: a) the knowledge of the EJC that has been provided by transcriptome-wide assays, and b) their limitations.

3.5.1 Prior to high-throughput: studying individual junctions

The effect of the EJC on specific splicing events observed in *D. melanogaster* prompted the study of individual junctions (Saulière, Haque, et al. 2010). Coupling luciferase reporters to naturally intron-containing 3'UTRs allows to detect introns capable of triggering EJC-dependent NMD. Inhibiting NMD factors causes an increase of luciferase activity relative to control only for NMD-targeted reporters. This assay revealed that only in a subset of the tested introns were sensitive to NMD, which were confirmed to be EJC-bound by pull-down experiments. Replacing exonic sequences of NMD-sensitive constructs with those from NMD-oblivious, caused a decrease in NMD sensitivity. These observations indicate a possible role of *cis*-acting elements on EJC deposition or stability. In consequence, a regulated EJC assembly implies specific targeting of splicing junctions and transcripts in *D. melanogaster*. To determine whether this is a global phenomenon, it is crucial to elucidate the binding sites of the EJC.

3.5.2 The first HITS-CLIP hints a differential loading

Prior to the early 2010's, there were no published *in vivo* transcriptome-wide studies of the EJC. Saulière and colleagues successfully obtained HITS-CLIP libraries using anti-eIF4A3 antibodies to pull down the transcriptome-wide binding sites of eIF4A3 in human embryonic kidney (HEK) cells (Saulière, Murigneux, et al. 2012), setting the precedent for known EJC binding modalities *in vivo*. The meta-analysis plot of the distribution of the center of reads relative to the exon junction confirmed the canonical binding site of the EJC *in vivo* (3.13a), which had only been observed in *in vitro* assays. Peak calling was performed with the *FindPeaks* tool, originally developed for ChIP-seq data (Fejes et al. 2008). Interestingly, around 50% of these peaks were found outside of the canonical binding site, even at high peak signal thresholds (3.13b). Moreover, around 0.1% of detected mRNAs contained only *non-canonical* EJC peaks; some of these *all non-canonical EJC* transcripts were later validated by immunoprecitipation followed by PCR. These results suggested for the first time that the EJC is frequently deposited away from its typically known binding site.

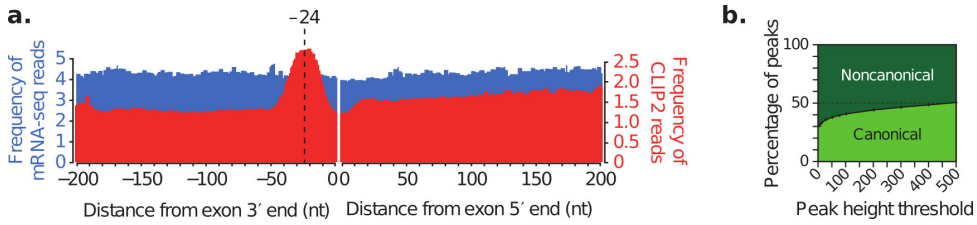

Figure 3.13: **a.** Meta-analysis plot of HITS-CLIP (red) and RNA-seq (blue) read distribution relative to the exon junction. The canonical EJC binding position is indicated with a dotted line at the -24 nt position. **b.** Proportion of canonical and non-canonical peaks according to different peak height thresholds. Adapted from Saulière, Murigneux, et al. 2012.

It was observed in pull-down experiments in *D. melanogaster* that the EJC is not loaded on all junctions of a gene (Saulière, Haque, et al. 2010). In addition to non-canonical bindings sites, the results in human cells revealed the absence of peaks in several exons of the same gene (Fig. 3.14a). The authors computed the percentage of exons that had at least one peak (canonical or non-canonical), and divided transcripts according to their abundance. They observed that most transcripts were between 60% to 70% loaded with EJC, reaching an average of 80% for highly abundant transcripts (Fig. 3.14b). These results indicated that the EJC does not occupy all the exons of human genes, and that the observed loading rate depends on the abundance of the transcript.

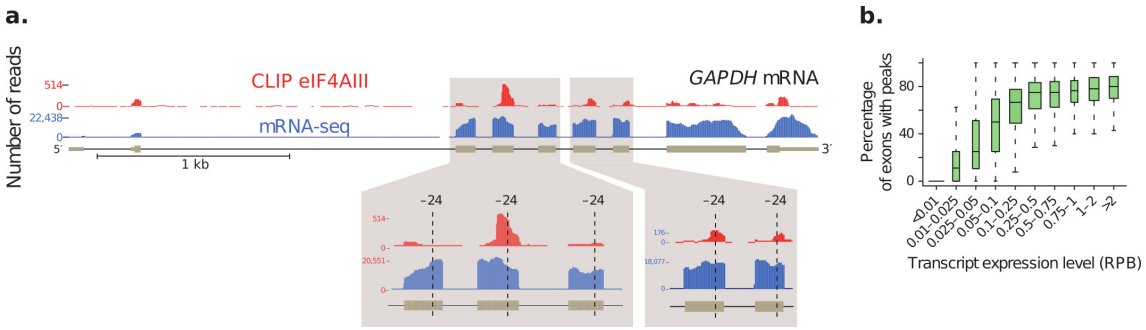

Figure 3.14: **a.** An example of differential EJC loading in the GAPDH mRNA. CLIP read coverage is represented in red, while RNA-seq coverage is represented in blue. GAPDH exons are drawn as beige boxes while introns are drawn as black lines. Exons 3 to 7 are enlarged. **b.** Percentage of exons within a gene with at least one canonical or non-canonical peak according to transcript abundance. RPB: reads per base. Adapted from Saulière, Murigneux, et al. 2012

Despite being the first transcriptome-wide landscape of the EJC binding sites, this study presents two major limitations. First, the reproducibility of CLIP signal is assessed by comparing the read counts at the gene level, then computing the Pearson's correlation coefficient between the two HITS-CLIP replicates (Fig. 3.15). As with early iCLIP experiments, this strategy assesses the reproducibility of the signal, but disregards the reproducibility of specific binding sites. This is aggravated by using the signal within a whole gene, rather than within the peaks where a specific signal enrichment is expected. Due to the relatively high correlation value, the authors proceeded to merge the two CLIP replicates prior to peak detection. This decision hinders the proper assessment of binding site reproducibility, thus preventing the distinction of stochastic or non-specific detection from the true EJC signal.

The second limitation is the lack of an input control to discard non-specific signal. In the study, CLIP signal is systematically compared to RNA-seq signal. While this controls for transcript abundance, it does not control for signal from non-specific IP. At the time of publication, the notion of SM-input had not yet been introduced. Therefore, the analyses do not consider the influence of potential technical artifacts on the detection of canonical and non-canonical peaks.

Overall, despite the major contributions of this work, the data analysis holds limitations in terms of reproducibility and specificity of the signal.

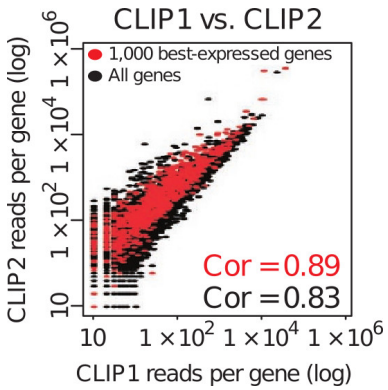

Figure 3.15: Log-transformed read counts per gene of CLIP signal. In red, the values for the top 1000 most abundant genes. *Cor*: Pearson's correlation coefficient of gene-level read counts for all genes (in black), and the top 1000 most abundant genes (in red). Adapted from Saulière, Murigneux, et al. 2012

3.5.3 mRNP footprints that suggest particle packaging

A study published in the same year as the eIF4A3 CLIP-seq, performed *in tandem* RIP-seq (RIPiT-Seq) of tagged versions of MAGOH and eIF4A3 in HEK cells (G. Singh, Kucukural, et al. 2012). Libraries were obtained by first purifying FLAG-tagged EJC components, followed by a second IP using either anti-eIF4A3 or anti-MAGOH antibodies. Instead of crosslinked RNA fragments, this strategy sequences the RNase-protected fragments bound by the proteins of interest. Radio-labeling and gel separation of purified fragments revealed two distinct size footprints: fragments longer than 30 nt (long footprints), and fragments 10 to 15 nt long (short footprints, see Fig. 3.16a). Mass spectrometry of the proteins binding long footprints revealed EJC partners spanning up to 150 nt, which were mainly members of the SR and SR-like protein families. This suggested a that mRNPs were packaged due to higher-order protein-protein interactions (Fig. 3.16b).

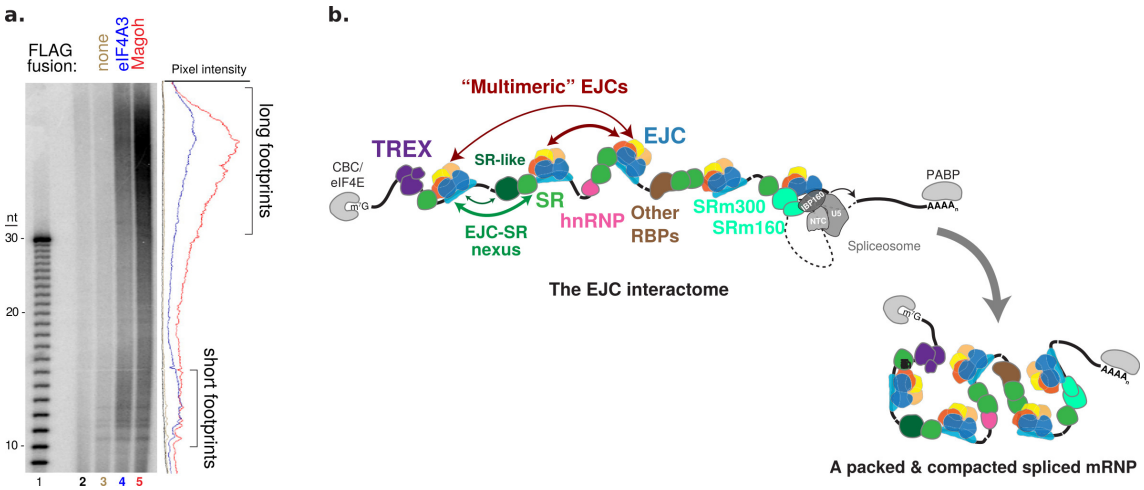

Figure 3.16: **a.** Length distribution of RNase I-resistant EJC footprints. Base-hydrolyzed synthetic polyU 30 RNA (lane 1) or purified RNA fragments from RIPiTs indicated at top (lanes 2-5) were 5'-end ^{32}P labeled and separated by denaturing PAGE. Auto-radiography pixel intensity profiles of lanes 2–5 are on the right. **b.** The EJC interactome and a new view of mRNP structure. Exonic RNA (solid black line); a generic intron (dashed black line); proteins enriched ¿10-fold in the EJC proteome (color ovals); undetected proteins known to bind to mRNA ends (gray ovals); bridging protein-protein interactions (green spheres). Adapted from G. Singh, Kucukural, et al. 2012

Short footprint fragments from both anti-eIF4A3 (FLAG-eIF4A3) and anti-

MAGOH (FLAG-MAGOH) pull-downs were sequenced to map transcriptome-wide interactions. Additionally, long footprint fragments (FLAG-EIF4A3 long), RNase protected fragments prior to IP (nuclear mRNP protection), and total RNA were sequenced as well. The meta-analysis plots of short fragment libraries confirmed the signal enrichment around the canonical region, whereas long fragments showed an upstream-shifted distribution (Fig. 3.17a). The authors found a higher variation of read signal inside the canonical regions within a gene for short fragment libraries, suggesting a differential EJC loading (Fig. 3.17b).

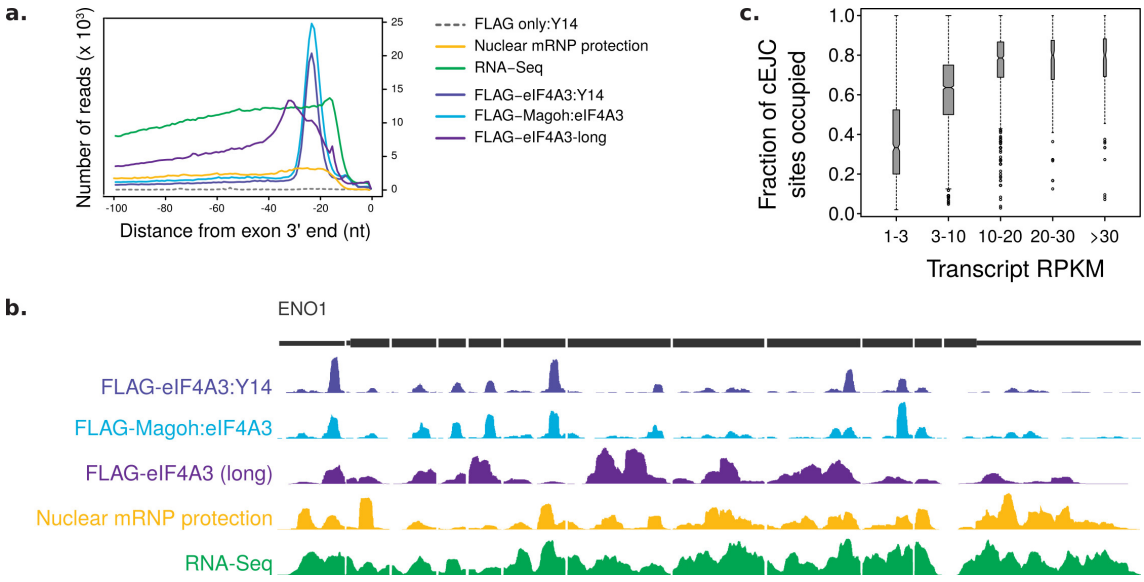

Figure 3.17: **a.** Meta-analysis plots of short EJC footprints (FLAG-eIF4A3:Y14 and FLAG-MAGOH:eIF4A3), long EJC footprints (FLAG-eIF4A3 long), nuclear mRNP footprints, or RNA-seq reads; distances to the 5'-exon junction were computed using the center of reads. **b.** Distribution of reads from the different libraries along the spliced representation of the *ENO1* gene. **c.** Fraction of EJC-occupied exons according to transcript abundance determined by *RPKM* from RNA-seq data. Adapted from G. Singh, Kucukural, et al. 2012

To evaluate the differential loading globally, peaks were detected with a custom algorithm, and the exon occupancy according to gene expression was computed, considering canonical peaks exclusively. The authors found a stable value of 80% for highly abundant genes, whereas EJC occupancy was more coverage-dependent in lower abundance genes (Fig. 3.17c). Altogether, these results suggested again that the EJC was not present in all junctions of a gene, despite originating from a crosslink-free technique.

Interestingly, the authors developed a dedicated algorithm to detect EJC peaks. Their strategy consisted on using a Poisson distribution to find significant read counts at individual positions. The parameter λ of the Poisson distribution was fitted using maximum likelihood with the signal *outside* the canonical region (-31 to -15 nt upstream the exon junction). Consecutive positions with significant counts ($P < 0.01$) were merged into the same peak. However, a detailed assessment of binding site detection reproducibility is not reported. Technical replicates from each IP, either FLAG-eIF4A3 or FLAG-MAGOH, were merged prior to peak calling. This decision is justified with a high peak correlation between the replicates, but the method to assess this is not specified. Reproducible peaks were selected by intersecting the peaks detected on each IP library, but the fraction of reproducible peaks over the total detected peaks is not indicated. Thus, the reproducibility of this binding site

detection strategy is not reported, despite targeting the canonical signal of the EJC. Finally, similar to the 2012 HITS-CLIP study, the EJC footprint data also suggested a non-canonical deposition. However, motif analysis of the regions neighboring non-canonical peaks revealed significantly enriched ESE sequences bound by SR-proteins. This suggests that non-canonical signal enrichment is likely due to non-specific co-purification of EJC partners.

3.5.4 A high-resolution approach yields a sharp signal

By 2016, the first single-nucleotide resolution CLIP libraries of the EJC came to light. In a 2016 study, iCLIP data sets of core components eIF4A3 and CASC3, and peripheral factors RNPS1 and UPF3B were obtained from HeLa cells (Hauer, Sieber, et al. 2016); PTB (polypyrimidine tract binding protein) iCLIP, and RNA-seq data was used as control. First, meta-exon representations of EJC components show the typical signal enrichment near the 3'-end of exons, while RNA-seq and PTB do not (Fig. 3.18a). However, the choice of representing read positioning as percentages of the total exon length does not show the precise nucleotide where signal is sharper. After crosslinking site detection with iCounts, sequence motif analysis was performed on all data sets. The most enriched motif corresponded to the splice donor site, which was highly enriched approximately 25 nt downstream of binding sites (Fig. 3.18b). These results confirm that iCLIP data of EJC components is enriched near the canonical binding region.

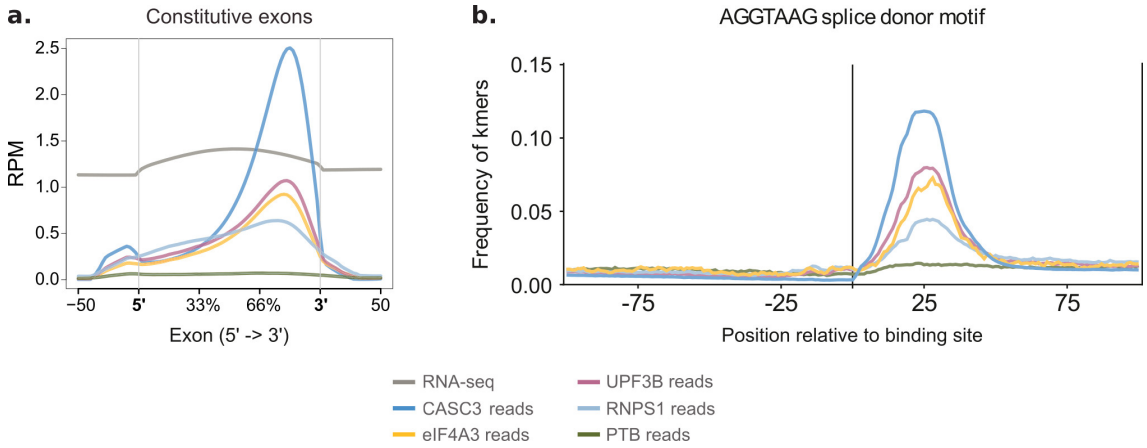

Figure 3.18: **a.** Meta-analysis plot of read 5'-end distribution within an exon; positions are represented as percentage of the total exon length. **b.** Distribution of splice donor motif relative to binding sites of EJC components and PTB. *RPM*: reads per million. Adapted from Hauer, Sieber, et al. 2016

To elucidate whether the composition of the EJC was homogeneous across all binding sites, peaks from different EJC components were intersected, and only those corresponding to at least two of the four components were kept. Interestingly, the majority of compound peaks had CASC3 as the common component. CASC3 was thus considered the component that provided the most *bona fide* binding sites. A higher enrichment was observed in peaks containing CASC3 than in the CASC3-less counterparts (Fig. 3.19a-b). According to these peak distribution profiles, it was hypothesized that the EJC is assembled on the canonical region when CASC3 is part of the complex, whereas non-canonical positions corresponded to CASC3-less EJC

(Fig. 3.19c). These results suggested that the EJC has a heterogeneous composition that may affect the site of deposition.

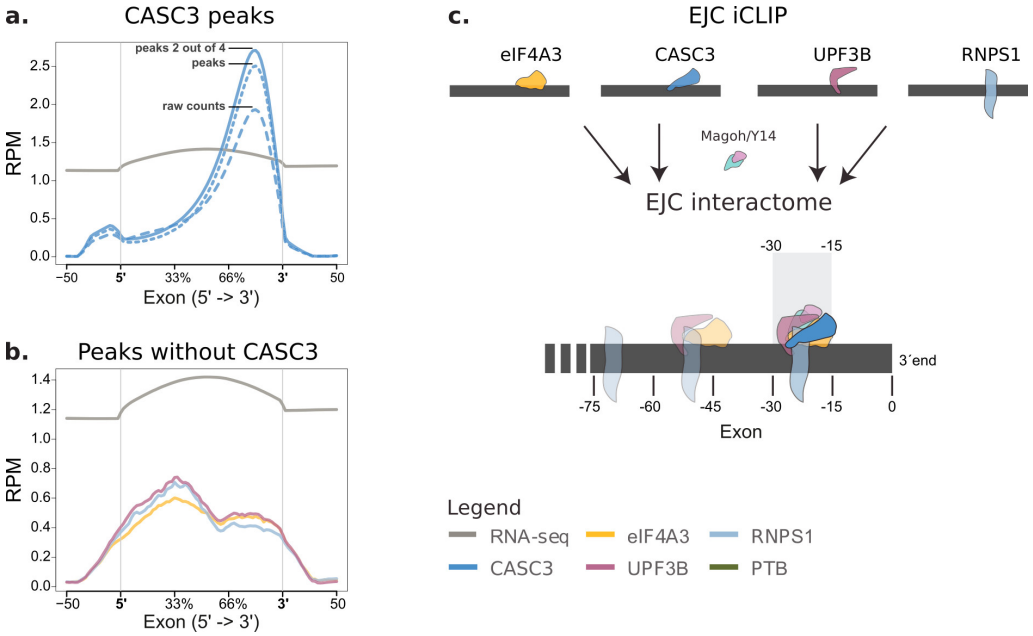

Figure 3.19: Meta-analysis plot of peak distributions within an exon for **a.** CASC3-containing peaks: shared with other components (2 out of 4), all CASC3 peaks (peaks), or read counts; **b.** peaks not containing CASC3. **c.** Interpretation of the differential EJC composition revealed by peak intersection. *RPM*: reads per million. Adapted from Hauer, Sieber, et al. 2016

Gene ontology analyses revealed that the functions of CASC3-loaded genes were enriched in RNA processing, cell cycle and chromosome organization. Previous studies had reported that genes involved in these processes were prone to alternative splicing events. Thus, genes were ranked according to their EJC occupancy and expression, revealing that alternative exons were enriched in highly abundant and highly occupied exons (Fig. 3.20a-b). Conversely, poorly occupied and highly abundant exons were mostly found in genes encoding ribosomal proteins (Fig. 3.20c). This differential occupancy suggests that the EJC has a *trans*-acting regulating role in RNA processing.

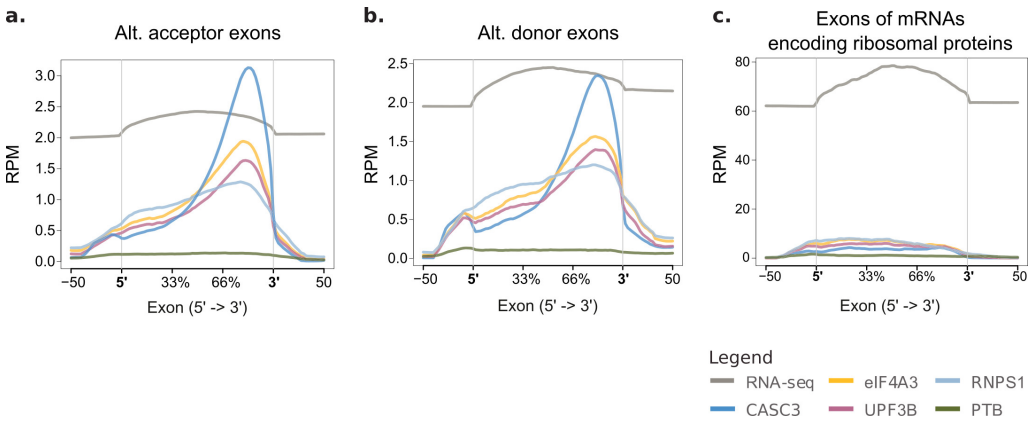

Figure 3.20: Meta-analysis plot of read 5'-ends of **a.** alternative splice acceptor exons; **b.** alternative splice donor exons; **c.** exons from genes encoding ribosomal proteins. *RPM*: reads per million. Adapted from Hauer, Sieber, et al. 2016

Regarding the reproducibility of the data, this study followed the same strategies as the previous EJC CLIP studies. On one hand, the reproducibility of the

iCLIP signal is assessed by comparing gene-level read counts across replicates, then by computing the associated coefficient of correlation (Fig. 3.21). Again, this a) ignores local read enrichment, and b) assesses the correlation of the signal rather than the reproducibility of the binding sites. Furthermore, technical replicates were merged prior to peak detection to increase the coverage of the data sets. Thus, despite the contributions presented by this study, the reproducibility of EJC binding sites is not addressed.

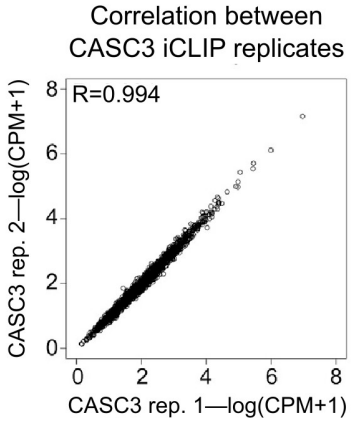

Figure 3.21: Scatter plot of log-transformed read counts per gene across CASC3 iCLIP replicates. *CPM*: counts per million. Adapted from Hauer, Sieber, et al. 2016

As discussed in section 3.3.1, meta-analysis plots were employed in a 2015 study to show a bias the effect of cDNA length on the crosslinking site positioning of iCLIP data (Hauer, Curk, et al. 2015). However, a subsequent study proved that precise crosslinking sites can be recapitulated with proper RNase conditions that result in cDNAs with high size variation (Haberman et al. 2017). In the latter study, three eIF4A3 iCLIP libraries were analyzed to assess the effect of RNase treatment in crosslinking site positioning. To do so, the 5'-end of read distribution is represented in a meta-analysis plot (Fig. 3.22). The signal of eIF4A3-iCLIP1 reads showed a shift upstream of the canonical binding site up to the -50 nt (similar to low resolution HITS-CLIP data); this library contained cDNAs with highly constrained sizes. Conversely, both eIF4A3-iCLIP2 and eIF4A3-iCLIP3 presented a sharper signal enrichment near the canonical region, in spite of an upstream shift of a few nucleotides. Although the aim of the study was to show the proper usage of 5'-ends in iCLIP libraries, it showed for the first time the exact positioning of eIF4A3 signal with single-nucleotide resolution. Yet, no peak detection for eIF4A3 was reported.

3.5.5 Transcriptome-wide in the fly confirms the main EJC roles

A recent study in *D. melanogaster* cells explore the transcriptome-wide landscape of EJC binding sites (Obrdlik et al. 2019). An alternative to UV crosslink was employed to stabilize EJC-RNA interaction. Instead of inducing protein-RNA crosslink, the authors use the crosslinking agent dithio(bis-) succinimidylpropionate (DSP) to generate covalent bonds between aminoacids in close proximity (Fig. 3.23). Stabilizing the interactions between EJC components in the close RNA-binding conformation allows the use of stringent purification conditions. An IP input is used to control

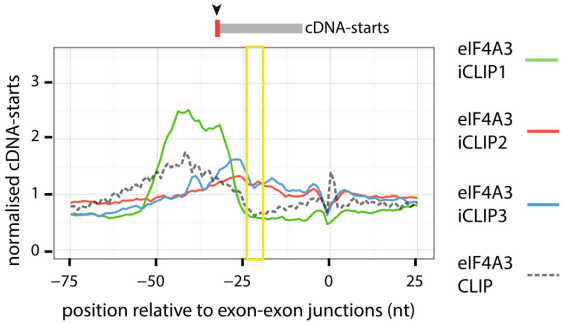

Figure 3.22: Meta-exon representation 5'-ends of reads from three iCLIP and one HITS-CLIP libraries of eIF4A3. The yellow box represents the canonical region of the EJC. Adapted from Haberman et al. 2017

for non-specific purification, also referred to as the total mRBP footprint. This approach is designated as ipaRt: isolation of protein complexes and associated RNA targets.

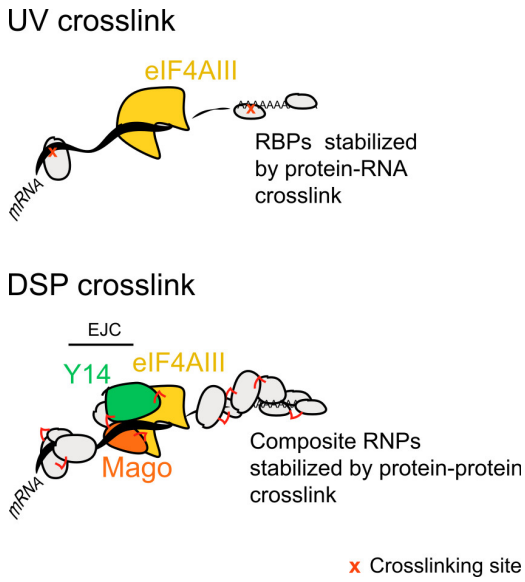

Figure 3.23: Representation of the effect of crosslinking with UV compared to DSP. UV irradiation leads to stabilization of direct protein-RNA in- teractions. DSP treatment results in efficient retention of proteins associated with RNA by stabilization of polypeptide interactions either within an RBP or between an RBP and other moieties within a complex. Adapted from Obrdlik et al. 2019

The global EJC footprint of ipaRt reads is represented with a meta-analysis plot of the total read coverage around the exon junction. The typical enrichment near the canonical binding site is observed (Fig. 3.24). Yet, as the signal of whole reads is used, this data does not position the EJC binding site with high resolution.

To detect individual binding sites, a simple thresholding strategy was used. Regions were read coverage was above 2-fold the average transcript coverage were considered as peaks, whose location corresponded to the position of the signal maximum. To distinguish specific EJC signal enrichment from non-specific IP artifacts, a window of 20 nt surrounding peak locations was defined to compute the $\log_2$-fold change (log2-FC) between EJC ipaRt and the input signals. Only peaks with log2-FC >1 were kept.

To assess the detection of non-canonical EJC binding sites, peaks were divided

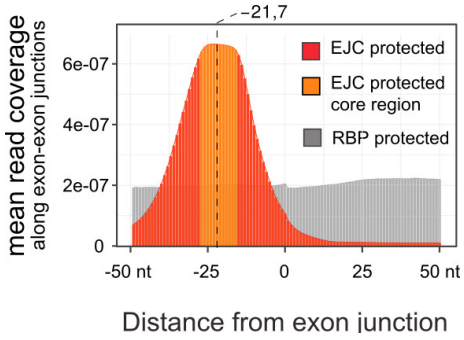

Figure 3.24: Meta-analysis plot of EJC (red) or total mRBP (gray) ipaRt reads around the exon junction; the maximal EJC protection region is highlighted in orange between the -27 and -15 position, with a median positioning around the -21 nt. Adapted from Obrdlik et al. 2019

into proximal ($\leq$ 50 nt from the exon junction), and remote ($>$ 50 nt). Proximal peaks were more numerous (around 95.5% of all peaks), and had stronger read coverage than remote peaks (Fig. 3.25a). Furthermore, the fraction of remote peaks was negligible when considering the top 25% most covered peaks (Fig. 3.25b). These results indicate that the EJC binding sites retrieved by ipaRt are in the canonical region, and question the existence of non-canonical EJC binding in *D. melanogaster*.

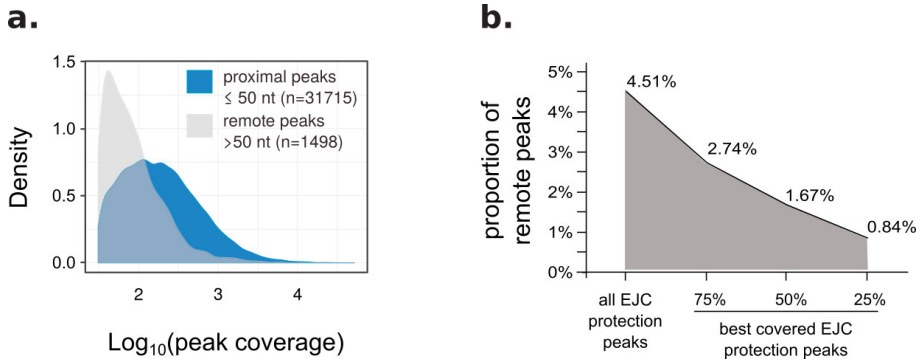

Figure 3.25: **a.** Read coverage distribution inside peaks within the canonical regional (proximal or located less than 50 nt from the exon junction, indicated in blue), and outside the canonical region (remote or located more than 50 nt from the exon junction, indicated in gray). **b.** Percentage of non-canonical (remote) peaks according to coverage thresholds selecting the most covered peaks. Adapted from Obrdlik et al. 2019

To detect EJC enriched genes relative to the input, the differential expression analysis tool DESeq2 was used. DESeq2 computes the gene-level log2-FC between two conditions and tests the statistical significance using a negative binomial distribution, while taking into account the variance between replicates (Love, Huber, and Anders 2014). Gene Ontology analysis was performed on the resulting list of significantly EJC enriched genes to assess whether specific processes were under EJC regulation. Genes involved in developmental and differentiation processes were particularly enriched in EJC reads (Fig. 3.26a). Interestingly, genes undergoing specific sub-cellular localization represented an important fraction of EJC enriched genes. These results confirmed the specific role of the EJC in the spatio-temporal regulation of gene expression.

Differential analysis between EJC ipaRt and input control allowed the separation of genes between two classes: EJC-enriched and RBP enriched. To determine which gene and exon features correlate with EJC enrichment, a decision tree model

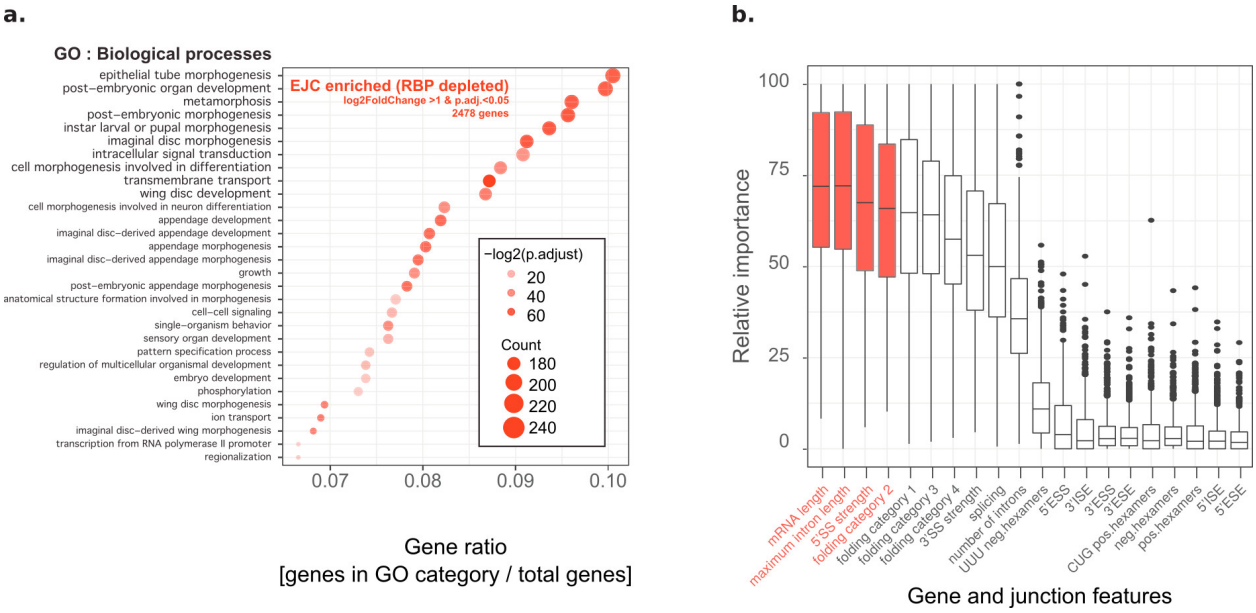

Figure 3.26: **a.** Read coverage distribution inside peaks within the canonical regional (proximal or located less than 50 nt from the exon junction, indicated in blue), and outside the canonical region (remote or located more than 50 nt from the exon junction, indicated in gray). **b.** Percentage of non-canonical (remote) peaks according to coverage thresholds selecting the most covered peaks. Adapted from Obrdlik et al. 2019

was trained to distinguish EJC enriched from RBP enriched genes. Computing the predictive power of gene and exon features allowed to assess their importance to explain EJC enrichment. Transcript length, maximum intron length, 5'-splice site strength, and secondary structure (designated as folding categories) were among the most relevant features (Fig. 3.26b). This statistical model confirms the previously observed roles of the EJC in long intron and weak intron splicing regulation, as well as the effect of RNA secondary structure on EJC positioning.

Despite these contributions, this study bypasses the reproducibility of binding sites. As with the studies analyzed in previous sections, biological triplicates are merged prior to binding site detection. Moreover, the correlation between the signal inside the peaks across replicates is not reported. Importantly, differential analyses are performed using the gene-level signal rather than individual binding sites. This highlights the existing shortcomings in assessing binding site reproducibility, which is crucial for differential binding studies.

3.5.6 An interesting complex that is hard to pin down

The effort of several teams to obtain a transcriptome-wide map of the EJC highlights the importance of its binding site localization to better understand its role in gene expression regulation. Yet, the current strategies to analyze CLIP data limit the new insights of the EJC binding modalities and their functional consequences. As we have presented in this section, a common limitation to EJC CLIP studies is the reproducibility of binding sites, often bypassed by assessing gene-level signal reproducibility. However, overcoming this limitation is essential to determine *bona fide* binding sites. Only reproducible binding sites hold the potential to yield significant results when performing comparisons in different biological conditions or

cellular contexts.

3.6 Aims of the project: the complexity of the EJC is a transcriptome-wide study

Thus far, we have presented the variety of cellular functions of the EJC and their physiological role at the organism level (Chapter 2). Despite the extensive knowledge gathered over almost two decades of research on the EJC, many questions remain to be addressed. Biochemistry and molecular biology approaches have been crucial to characterize its structure and to discover many of its functions. Yet, these approaches limit the study of the multiple ramifications of the EJC within PTGR.

Reporter gene assays in *D. melanogaster* suggest the complex is deposited differentially along the junctions, rather than homogeneously. Importantly, some genes are not pulled down by EJC component IP, indicating a targeted EJC-regulation. However, the low throughput of these assays yields evidence for only a handful of genes. Moreover, although EJC deposition impacts specific splicing events in human cells, there is no evidence for whether assembly is targeted on specific junctions.

Several EJC partners have been determined using biochemical and molecular biology assays. Recombinant protein incubation determines whether the EJC is able to physically interact with other proteins, but does not offer proof of in vivo interactions. Pull-down experiments have revealed some EJC partners and their functional implication. Yet, detection of these interactions often requires prior knowledge of the partners under study, as it often involves gel separation and immunostaining. Although mass-spectrometry detects potential EJC partners, the most highly enriched peptides are associated with its assembly rather than the fine-tuned regulation of specific processes.

EJC assembly or absence may have a major impact on gene expression regulation. Whether EJC deposition follows a particular pattern along the transcript can influence mRNP packaging, with potential functional implications, is not known. Notably, regulating the assembly on junctions downstream of a stop codon can impact protein levels through NMD. However, the factors that determine the EJC deposition on individual exon junctions, as well as its consequences in expression regulation are still unknown.

We have claimed that studying PTGR is a systems biology problem. This is especially true for the EJC, due to its central placement in PTGR networks. To untangle the code behind the deposition and subsequent effects of the EJC, it is crucial to identify its individual binding sites across the transcriptome.

This work was carried out in the Expression of Eukaryotic Messenger RNA team at the Institute of Biology of the École Normale Supérieure of Paris. The research of the team is centered around the molecular characteristics of the EJC and its impact in eukaryotic gene expression. Studies range from spliceosome-dependent assembly of the complex, to sub-cellular mRNP localization, to mechanism of NMD. In par-

allel to the invention and the evolution of CLIP and high-throughput technologies, multiple attempts to obtain EJC CLIP libraries have been performed by the team. The ultimate goal is to unravel the transcriptome-wide binding landscape of the EJC and its functional implication in PTGR. Nevertheless, both the experimental and data processing aspects of CLIP are challenging. On one hand, published EJC HITS-CLIP data is limited by low binding site resolution and uncertain reproducibility. On the other hand, single-nucleotide resolution data has been employed for meta-analyses to assess CLIP quality, and not to study the binding modalities of the EJC. In this context, the main aims of this work were:

1. **At the technical front:** To develop a data analysis strategy to obtain highly reproducible binding sites from single-nucleotide resolution CLIP data, and thus establish a transcriptome-wide map of the EJC.
2. **At the EJC knowledge front:**
   (a) To infer from the binding site map whether an EJC is differentially loaded along the exons of the same gene. Reproducible binding sites are therefore crucial to confidently determine the presence, or absence, of the EJC.
   (b) To correlate EJC deposition with gene structural factors, provided that the EJC is differentially assembled. This would allow us to gain insight into the rules that determine EJC deposition on specific junctions.

The results of this work are divided in two parts. First, we will present a dedicated analysis pipeline to mine EJC-specific high resolution CLIP data. Next, we will present our contribution to EJC knowledge, notably by comparison to existing knowledge obtained form CLIP data.

PART II

RESULTS

Chapter 4

Acquiring single-nucleotide EJC CLIP data

In the previous chapter, I presented an overview of different CLIP protocols that have emerged in the past decade. In this chapter, we will focus in our work on EJC data acquisition. We will first summarize our published results of the development of an eCLIP variant that distinguishes crosslinking-induced truncations from read-through events. We will next give an overview on quality check and pre-processing of CLIP data, then present our EJC data sets.

4.1 Detecting crosslinking sites with monitored eCLIP

Despite major advancements, the protocols presented in section 3.2 do not distinguish between crosslinking-induced truncation and read-through events at the RT step. Yet, single-nucleotide resolution protocols rely on the signal from truncation events to precisely identify the crosslinking site. Read-through events can be estimated during data analysis by computing the rate of crosslinking-induced mutations, (Sugimoto et al. 2012). However, mutation rates can be variable among different RBPs, and are rare within the cDNAs resulting from read-through events (C. Zhang and Darnell 2011; Sugimoto et al. 2012). We introduced monitored eCLIP (meCLIP) to precisely assess the rate of read-through events. We obtained 8 eIF4A3 meCLIP libraries and assessed the effect of read-through reads in binding site definition of the EJC.

The use of polyclonal antibodies can decrease the specificity of IP. This can be overcome by transfecting cells with plasmids containing tagged versions of the protein of interest. The tagged protein is over-expressed and an anti-tag antibody is used to purify it. However, increasing the amount of protein in the cell may generate technical artifacts. To increase the efficiency and the specificity of IP, while avoiding plasmid transfection, endogenous proteins are tagged with FLAG (Hopp et al. 1988) and HA epitopes using CRISPR as a genome editing tool. After RNA and protein digestion, a 13 nt long linker of a known sequence is ligated at the 5'-end of RNA fragments. As RT takes place in the 3' to 5' sense, truncated cDNAs will not include the 5'-linker, whereas cDNAs resulting from reading through the crosslinking site will include the 5'-linker (Fig. 4.1a).

The presence, or absence, of the 5'-linker in reads separates the signal from truncation events from the read-through events. This allows the precise quantification of the read-through rate of different reverse transcriptases (RTase) (Fig. 4.1b). We observed rates as low as 3% and can reach up to 25%, with most RTases producing 10 to 15% of read-through reads. On one hand, we confirmed that a high percentage of reads originate from RTase halting at the crosslinking site (truncated reads). On the other hand, ligating a 5'-linker allows to precisely quantify the percentage of read-through events in a CLIP library, which varies depending on the RT enzyme. On the experimental front, tagging the endogenous protein with CRISPR allows a more efficient IP, with less starting material and using less antibodies.

Single-nucleotide CLIP methods use the 5'-end of reads to locate the protein-RNA crosslinking site. In the case of read-through reads, the 5'-end may be distant from the crosslinking site depending on the length of the RNA fragment. As only the first 75 nucleotides of the fragments are sequenced, this may be aggravated for

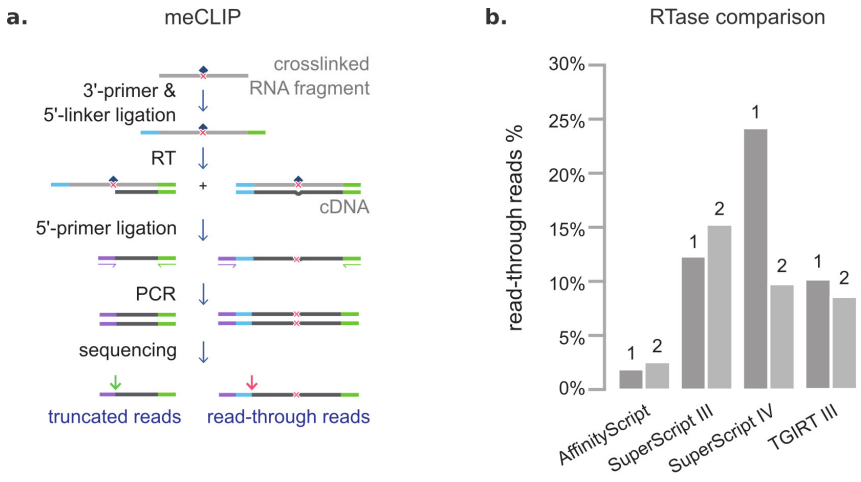

Figure 4.1: *a.* The main steps of the meCLIP protocol. The crosslinking site (represented with a red X) induces RTase halt and produces truncated reads. RTase read-through detects the 5'-linker (in blue), allowing separation of truncated and read-through reads. **b.** The percentage of read-through reads in libraries obtained with different reverse transcriptases (RTase).

cDNAs of hundreds of nucleotides of length. Thus, not distinguishing read-through events may lead to imprecise crosslinking site assignment. To assess the effect of read-through reads on binding site assignment, we ran peak detection with CITS (Shah et al. 2017) on EJC meCLIP data sets both before and after read-through read filtering. We designated each data set (*all reads* and *truncated reads* respectively. As revealed in genome browser examples, not removing read-through reads leads to detection of imprecise binding sites, generally shifted upstream (Fig. 4.2a-c). Next, we computed the percentage of peaks detected only on *all read* data sets (Fig. 4.2d). We found the percentage of *all read* peaks was correlated to the percentage of read-through reads in a library (Fig. 4.2e), but not to the total number of reads (Fig. 4.2f). Thus, we observed a direct effect of read-through events in imprecise binding site assignment.

Next, we compared the read-through and truncated read distribution of EJC libraries relative to the exon junction. Most read-through read signal is located upstream of the canonical EJC binding site, while truncated reads display a sharp enrichment around the -27th nucleotide (Fig. 4.3a). Thus, separating read-through events increases the precision of binding site definition for the EJC. We then compared meCLIP to EJC data sets obtained with other CLIP protocols (4.3b). We observe the increase in signal resolution from HITS-CLIP to iCLIP and eCLIP. Yet, meCLIP shows both higher signal enrichment and signal-to-noise ratio compared to other protocols. This high enrichment in the canonical region raises doubts about the EJC binding to non-canonical positions.

Altogether, we proposed a modification of the eCLIP protocol that increases the precision of binding site assignment. Further experimental and data analysis details can be found in Annex I (Chapter 9). Subsequent efforts aimed at increasing the number of EJC replicates to ensure the reproducibility of our results. However, obtaining high quality EJC libraries resulted to be a challenging task.

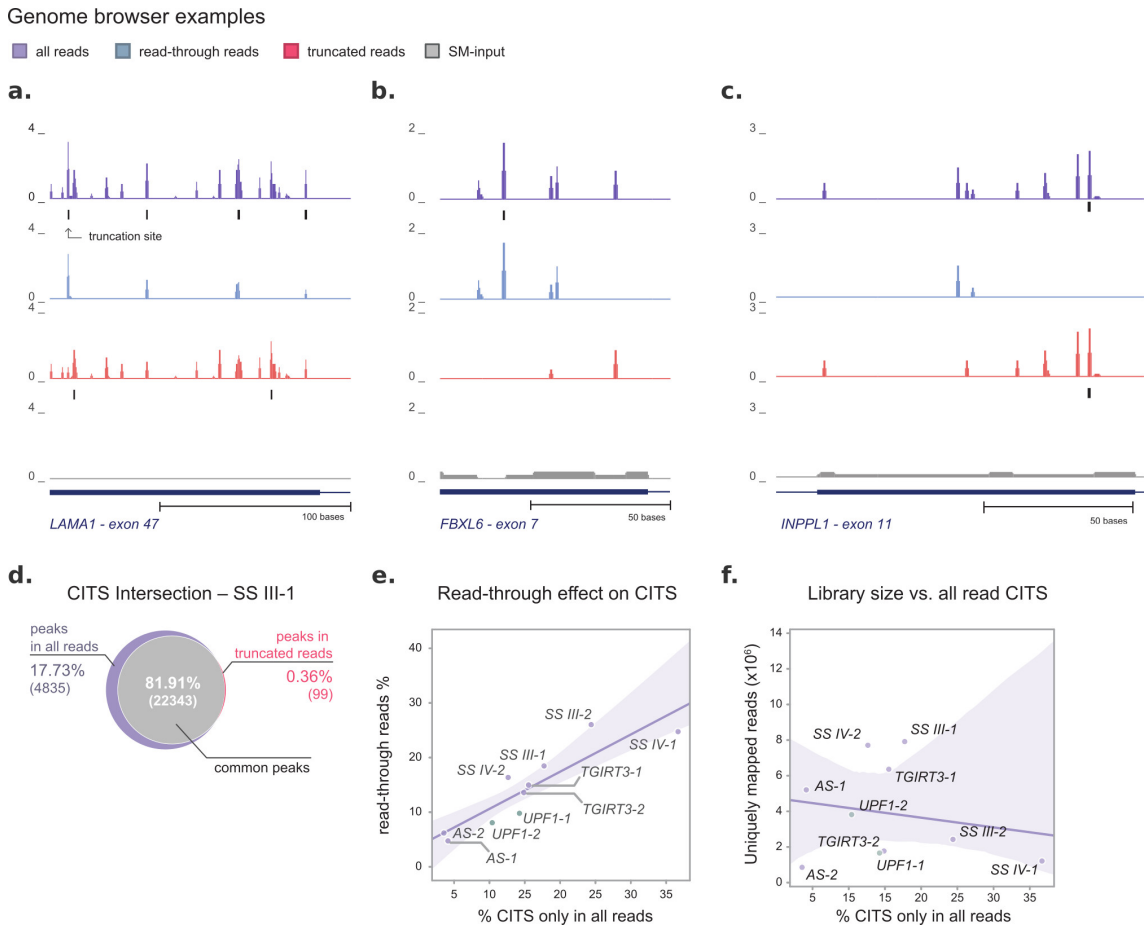

Figure 4.2: CITS detection is biased by read-through reads. **a-c** meCLIP reads mapped on three examples of exons. Each black underline corresponds to a CITS detected with CTK. Read coverage is in Reads Per Million (RPM). **d.** Venn diagram representing the intersection of peaks detected in the unfiltered (all reads) and filtered (truncated reads) data sets from SuperScript-IV replicate 1. **e.** Read-through reads percentages are plotted against the percentage of truncation sites (CITS) detected exclusively in all reads (purple in d). **f.** Number of uniquely mapped reads versus the percentage of CITS detected exclusively in all reads; the shade around the line indicates the confidence interval (95%) of the linear regression. AS: AffinityScript, SSIII: SuperScript III, SSIV: SuperScript IV.

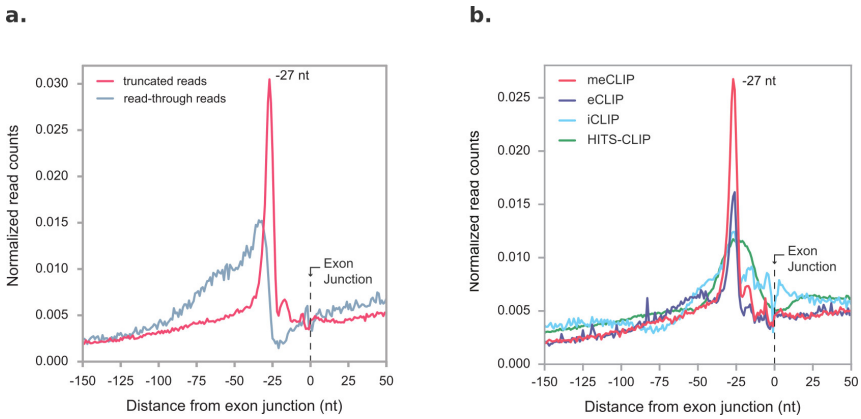

Figure 4.3: Positioning of 5'-ends of meCLIP reads relative to the exon junction. **a.** Distribution of eIF4A3 meCLIP reads: truncated reads (red) and read-through reads (blue). **b.** Distributions of eIF4A3 reads obtained with the meCLIP, eCLIP, iCLIP (Haberman et al. 2017), and HITS-CLIP (Saulière, Murigneux, et al. 2012) procedures. meCLIP signal corresponds to truncated reads, and is normalized using the number of uniquely mapped truncated reads.

4.2 Quality control and data pre-processing

Data pre-processing is crucial to assess the quality of the data and to prepare it for downstream analyses. In this section I will present the shortcomings of our

prior pre-processing pipeline, as well as our strategy to overcome them and obtain higher-quality data.

CLIP-seq data pre-processing is similar to other second-generation sequencing data. Adaptor trimming and base calling quality check are the essential steps prior to read alignment to the reference genome. In the library construction protocol, a unique molecular identifier (UMI) system was incorporated to 1) sequence different libraries in the same run, and 2) to identify PCR duplicates. Thus, CLIP data pre-processing involves the additional steps of library de-multiplexing, and PCR duplicate removal.

Over the past years, the lab has generated several CLIP libraries of the EJC core protein eIF4A3. Protocols such as HITS-CLIP, iCLIP, eCLIP and meCLIP have been applied on both wild-type and tagged forms of the protein (Saulière, Murigneux, et al. 2012; Haberman et al. 2017; Hocq et al. 2018). In this work, we focus on the analysis of eIF4A3 meCLIP and eCLIP data, with the goal of obtaining a high resolution map of EJC binding sites. In early stages, we exclusively had 8 sequenced meCLIP libraries from HeLa cells at our disposal (Hocq et al. 2018). However, a close examination revealed that these data had a high PCR duplication rate after pre-processing.

Before sequencing, libraries are amplified through PCR to obtain a sufficient amount of cDNA. Although it is a necessary step, it can bias read quantification and peak detection in downstream analyses (Fig. 4.4a). Prior to this work, PCR duplicate removal was performed by discarding reads whose entire sequence was an exact match. This approach ignores potential PCR amplification or sequencing errors that introduce mismatches between duplicates, thus underestimating the PCR duplication rate (Fig. 4.4b). To overcome this problem, we incorporated UMI tools in the pre-processing pipeline (T. Smith, Heger, and Sudbery 2017). It computes edit distances between unique molecular identifiers (UMIs) of reads mapping to identical genomic coordinates, then tags and clusters reads below a UMI distance threshold as duplicates. This approach takes into account both sequence variation between duplicates and genomic location (Fig. 4.4c). After applying UMI tools, the number of usable reads in each meCLIP data set decreased between 3 and 12 times compared to our previous strategy. Because of the limited coverage of individual meCLIP data sets, we decided to merge two separate sets into two pseudo-replicates of approximately 600,000 reads each (see table 4.1).

Despite merging, the coverage of meCLIP data sets remained limited. Using published RNA-seq data from HeLa cells (Z. Wang, Murigneux, and Le Hir 2014; see table 4.1), we estimated the number of expressed genes to be around 9000. Assuming that the average number of exon junctions per gene is 8, this tells us that there are on average 11 reads per junction. This relatively low coverage motivated parallel efforts both in the experimental and data quality assessment of CLIP.

In the experimental front, the team opted for the eCLIP protocol over the meCLIP protocol, favoring RNA yield over read-through event detection. Additionally, they tweaked experimental conditions and tested purification alternatives to maximize library complexity. In the data analysis front, we set up a pipeline to estimate and predict the library complexity resulting from the experiments. In col-

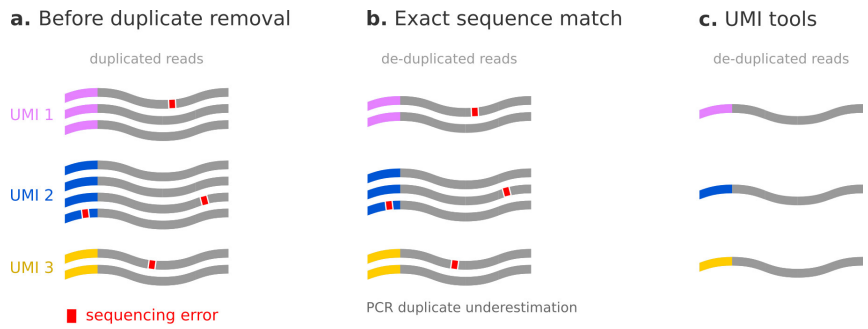

Figure 4.4: **a.** Representation of reads prior to PCR removal; the colored segments represent different unique molecular identifiers (UMI) ligated to cDNA fragments prior to sequencing; the red squares represent nucleotide mismatches between otherwise identical reads. **b.** PCR duplicate removal comparing the entire read sequence; only exact sequence matches are considered as PCR duplicates, which leads to PCR duplicate underestimation when there are sequencing errors. **c.** PCR duplicate removal with UMI tools; it compares only the UMI sequence of reads with identical genomic coordinates, taking into account sequencing errors, which results in a better estimation and removal of PCR duplicates. UMI: unique molecular identifier.

Table 4.1: Data set summary

Library	**Protocol**	**Date**	**Usable reads (coding exons)**
meCLIP-1	meCLIP	10/2016	638636
meCLIP-2	meCLIP	10/2016	600934
eCLIP1-1	eCLIP	10/2019	2031071
eCLIP2-1	eCLIP	10/2019	8201373
eCLIP1-2	eCLIP	11/2019	2014508
eCLIP2-2	eCLIP	11/2019	8253491
input-1	eCLIP	11/2019	1182290
input-2	eCLIP	11/2019	2006825
RNA-1	RNA-seq	11/2014	45278330
RNA-2	RNA-seq	11/2014	43735087

laboration with the sequencing platform of the Cellular Integrative Biology Institute (I2BC), we incorporated the analysis of sequencing pre-runs. A sample of the cDNA libraries is sequenced to obtain a small number of reads from which PCR duplication rate and complexity can be estimated. This prior analysis informs on whether a deeper and more costly sequencing run is worth pursuing. This collaborative effort accelerated library acquisition, and produced two new eIF4A3 eCLIP libraries (that were both sequenced twice in separate runs), and two SM-input control data sets (see table 4.1). Due to their exceptional coverage, we randomly sub-sampled each eCLIP2 data set to the exact number of reads in eCLIP1 data sets; we performed two sub-samplings to take into account the random variation. Thus, for subsequent analyses, we used the merged meCLIP pseudo-replicates (meCLIP-1 and meCLIP-2), the two sequencing runs of eCLIP1 (eCLIP1-1 and eCLIP1-2), the sub-sampled data sets of eCLIP2 (eCLIP2-S1, eCLIP2-S2, eCLIP2-S3, and eCLIP2-S4), the SM-input data sets (input-1 and input-2), and two RNA-seq data sets previously obtained from HeLa cells (RNA-1, RNA-2, published in Z. Wang, Murigneux, and Le Hir 2014).

To summarize, our efforts to improve the pre-processing pipeline gave us insight into CLIP-seq data quality check. This allowed us to obtain and assess the quality of CLIP libraries efficiently, and resulted in a higher-quality data set for downstream

analyses. Moreover, the CLIP technical replicates as well as the sequencing run replicates allow us to distinguish reproducible signal from technical noise. In the following section we will discuss the results of peak calling and its limitations.

Chapter 5

Peak calling with high resolution: a tale of reproducibility

5.1 Sensitivity and specificity of peak detection

As we presented in the previous chapter, peak callers are applied on CLIP-seq data to detect significant signal enrichment at precise genomic locations. Among available peak callers, PureCLIP (Krakau, Richard, and Marsico 2017) and CITS (Shah et al. 2017) yield single-nucleotide resolution binding sites in iCLIP, eCLIP and meCLIP data. For this reason, these tools had the potential to provide us with a high-resolution map of EJC targets.

We thus ran both peak callers in our CLIP data sets, excluding the whole eCLIP2 library and using its sub-sampled data sets. Independently of the peak caller, the distribution of peaks relative to the exon junction showed the specific EJC enrichment around the upstream 27th position (Fig. 5.1a-b). This revealed that, from a global point of view, peaks were frequently detected in the expected EJC binding site.

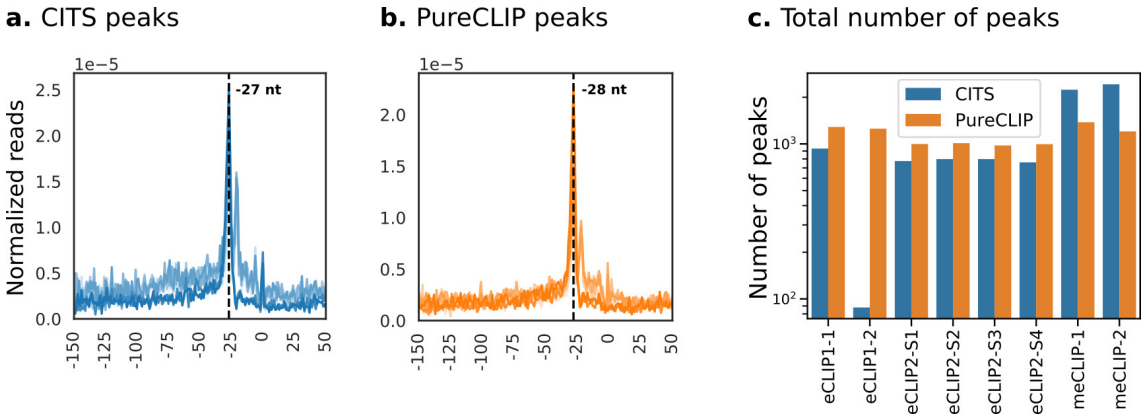

Figure 5.1: Normalized peak counts around the exon junction of **a.** peaks detected with CITS (blue), and **b.** peaks detected with PureCLIP (orange). Note the shift from -27th to -28th in PureCLIP results. This may be explained by the merging of several crosslink sites into binding regions of slightly lower resolution. **c.** Number of peaks detected by PureCLIP and CITS in CLIP data sets.

To assess the sensitivity of each method, we counted peaks inside a 10-nucleotide window around the 27th position; we will refer to these peaks as **canonical peaks**. Overall, we obtained from 977 to 1380, and from 88 to 2424 binding sites with PureCLIP and CITS respectively (Fig. 5.1c). We observed a consistently higher number of peaks in PureCLIP results, except for the meCLIP pseudo-replicates. These results suggest that PureCLIP detects EJC signal with higher sensitivity than CITS.

To assess the strength of the signal at the peak level, we selected counted the number of reads within canonical peaks. We sorted them by decreasing number of reads and selected the top 1000 for each peak caller (Fig. 5.2). We observed that a) PureCLIP signal is consistently stronger than CITS signal, and b) CITS detects less than 1000 canonical peaks for several CLIP data sets. Notably, PureCLIP signal was indistinguishable to CITS signal in meCLIP pseudo-replicates despite having a lower number of detected binding sites than CITS. Overall, peak detection results showed that PureCLIP offered both higher specificity and higher sensitivity than CITS for EJC signal detection. Thus, in following analyses we will focus on PureCLIP results when referring to peak detection.

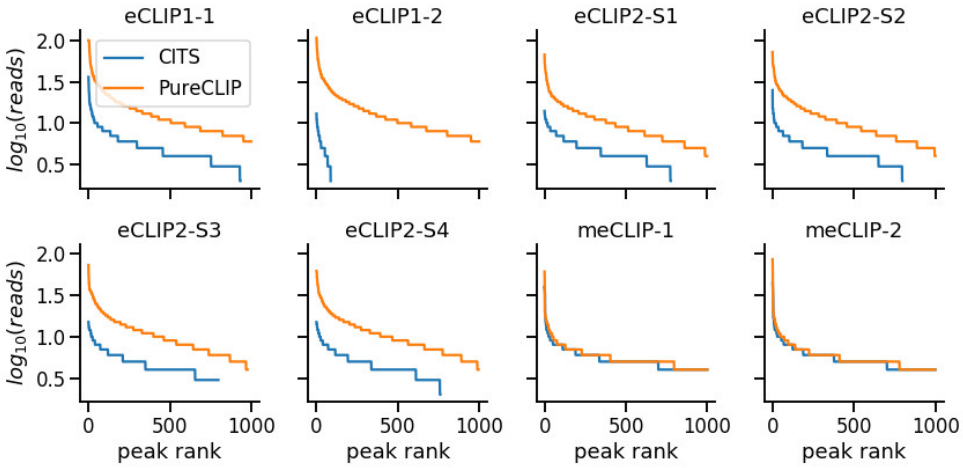

Figure 5.2: Log-transformed read count of peaks detected with PureCLIP and CITS inside the EJC canonical region. After sorting by decreasing number of reads, we selected the top 1000 peaks detected by each caller in separate data sets. CITS data sets with less than 1000 peaks are wholly shown.

5.2 Reproducibility of peak detection

Next, we proceeded to evaluate the reproducibility of the results. First, we computed the IDR value of PureCLIP peaks, following the ENCODE recommendations (see section 3.4.1). We observed that for most comparisons, IDR values rapidly increased with lower peak ranks (Fig. 5.3a). We computed the fraction of reproducible peaks with IDR <0.05 for each comparison, and found highly variable results ranging from 0.3% to around 85.9% (Fig. 5.3b). Yet, the number of reproducible peaks was rarely above 50 (with the exception of the eCLIP1-2/eCLIP2-S3 comparison). Altogether, the IDR values indicated a poor peak reproducibility of PureCLIP peaks, despite the specificity of their signal.

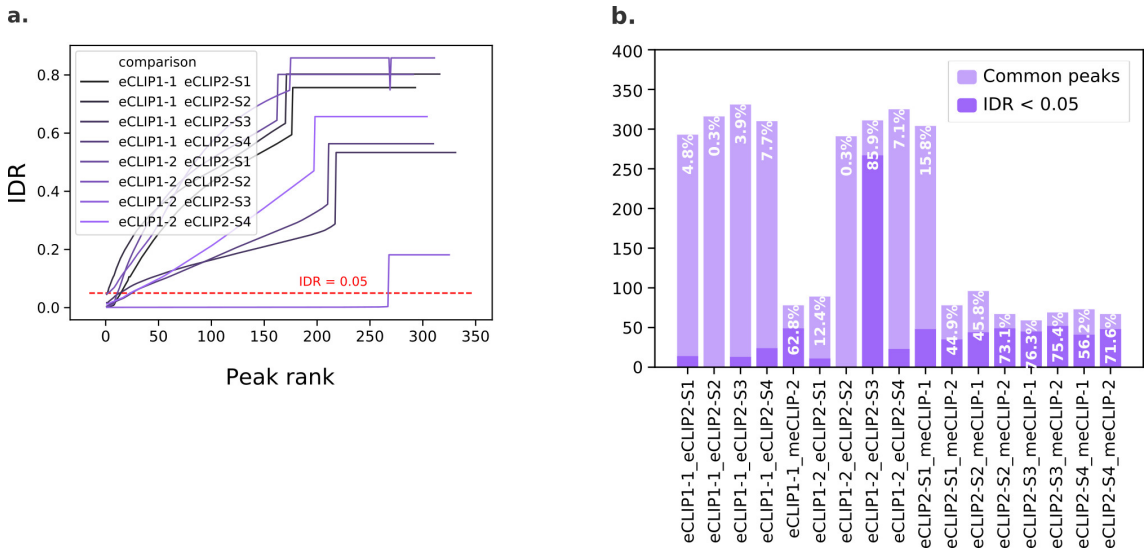

Figure 5.3: **a.** IDR curves of PureCLIP peaks detected on eCLIP data sets. The dotted red line indicates the IDR threshold of 0.05. **b.** Number of common peaks between CLIP replicates (in light purple), compared to the number of reproducible peaks (in darker purple). The percentage of reproducible peaks is indicated in white. IDR: irreproducibility discovery rate.

As the IDR method computes the reproducibility of commonly detected peaks,

we computed the Jaccard index to quantify the level of overlap between replicates. While we found a small number of peaks with IDR < 0.05 values, the number of commonly detected peaks was relatively low to begin with (never exceeding 350). We thus computed the Jaccard index between replicates, to quantitatively assess the pair-wise overlap of PureCLIP peaks (Fig. 5.4). Comparisons involving meCLIP replicates appear as the least reproducible with a Jaccard indexes of around 3%, despite having similar peak count to the rest of replicates. Regarding eCLIP replicates, we observe values between 15% and 17% across comparisons. These values indicate that more than 80% of peaks are not detected in both replicates. Thus, we found a low reproducibility rate of EJC peaks.

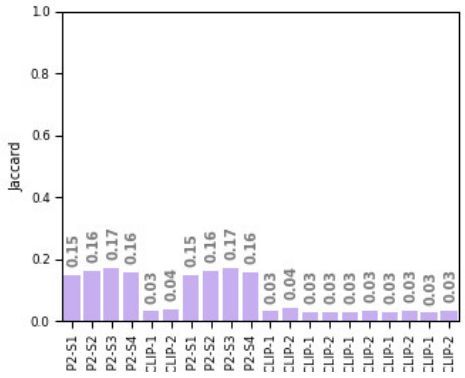

Figure 5.4: Jaccard indexes were computed by dividing the number of common peaks reported by the IDR software by the sum of the total number of peaks detected on each replicate separately, minus the number of common peaks.

5.3 The reproducibility of eCLIP is generally limited

To examine whether the low rate of reproducibility was specific to our eIF4A3 data, we studied CLIP data of other proteins. We downloaded eCLIP data of 72 different RBPs—two replicates each—from the ENCODE data portal (Davis et al. 2018). We detected peaks with PureCLIP, including SM-input as a co-variate, and with CITS using two different SM-input enrichment statistical tests. Next, we computed the Jaccard indexes for each RBP, and found that median values were approximately 0.20, 0.17, and 0.13, for PureCLIP, CITS-multitest, and CITS-permutation, respectively (Fig. 5.5). These values are similar to the ones we obtained with our EJC replicates, indicating that the low reproducibility rates are not specific to either our data or the eIF4A3 protein.

Next, we attempted to find an explanation to the observed Jaccard index values. First, we assessed the effect of the number of detected peaks, and found a low correlation to the Jaccard value (Fig. 5.6a). We then explored another peak detection quality score: the Fraction of Reads in Peaks (FRiP). Originally proposed for the assessment of ChIP-Seq data (Landt et al. 2012), the FRiP value indicates the fraction of reads inside peaks relative to the total number of reads in a library. The ENCODE consortium proposes that FRiP values under 1% indicate a good distinction between biologically relevant enrichment and background noise. This may appear counterintuitive, as one would expect that higher FRiP values indicate a better peak

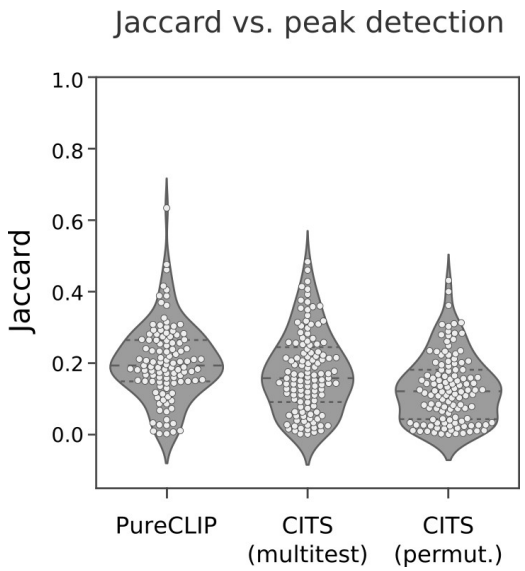

Figure 5.5: Jaccard index distribution of peaks detected with different peak callers in 72 different RBPs; CITS—multitest corresponds to SM-input enrichment statistical assessment proposed by Lovci et al. 2013, followed by multitest FDR correction; CITS-permutation corresponds to statistical assessment with p-values obtained by comparing the observed SM-input enrichment value to a distribution of values obtained with randomized peak positions within the gene.

detection. However, in ChIP-seq experiments, where background noise reads cover the majority of the genome, it is expected that only a small fraction of these reads are within specific binding sites.

We computed the FRiP values of the eCLIP replicates from the ENCODE project and found a moderate correlation to the Jaccard index of commonly detected peaks (Fig. 5.6b). This suggests that low FRiP values are associated to low reproducibility rates, despite being below the ENCODE recommended threshold. It should be noted that this recommendation concerns ChIP-seq peaks, and that there is no reference for the sparse, transcriptome-related CLIP-seq peaks. Furthermore, that FRiP is highly dependent on the number of binding sites expected for a particular NA-binding protein. For instance, a protein that targets very few sites across the genome/transcriptome will have low FRiP values regardless of the IP quality or the peak detection specificity. Conversely, proteins with a wider arrange of targets may yield higher FRiP values despite acceptable IP and peak detection results.

These results indicate that the reproducibility of eCLIP high-resolution peaks is generally low, regardless of peak calling strategy. Although we did not find a direct explanation to the low Jaccard values, we found a moderate correlation to the FRiP value. The reproducibility rate may be explained by multiple experimental factors that determine the signal-to-noise ratio in a CLIP library. Experimental factors aside, we observed a general reproducibility problem when detecting single-nucleotide binding sites in CLIP data.

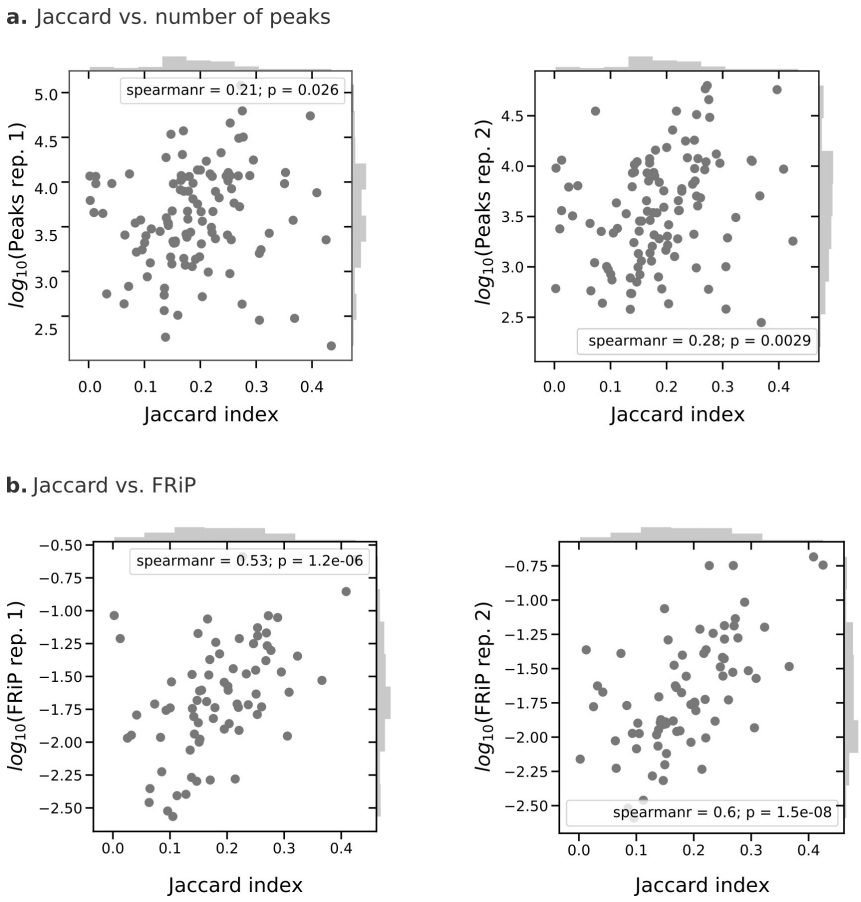

Figure 5.6: **a.** Scatter plot and correlation value comparing the log-transformed number of peaks in replicate 1 (left) and replicate 2 (right) against the Jaccard value of the 72 RBP peaks. **b.** Scatter plot and correlation value comparing the log-transformed Fraction of Reads in Peaks (FRiP) value in replicate 1 (left) and replicate 2 (right) against the Jaccard value of 72 RBP replicates.

5.4 The detection dilemma: precision vs. reproducibility

Thus far, we have applied the recommended analysis pipeline on our data. On one hand, the signal aggregation in the meta-exon shows a specific and reproducible enrichment in the EJC expected binding site. Yet, it provides only a global and qualitative profile of the data quality, with no information about binding sites or targets. On the other hand, peak detection results surfaced as a double-edged sword. The advantage of being the most informative level of detection is overshadowed by its low reproducibility rate. Hence, the levels of detection offered by current data analysis strategies are limited in information gain and reproducibility (illustrated in Fig. 5.7).

Yet, the meta-exon representations of the data indicates that there is EJC specific signal, despite the limitations we encountered in conventional peak detection. This prompted us to explore alternative strategies to process the EJC CLIP data. More precisely, we sought intermediate levels of detection between data aggregation (meta-analysis or along large regions), and single-nucleotide peaks. Our objective: to first obtain reproducible results that can be mined for relevant information.

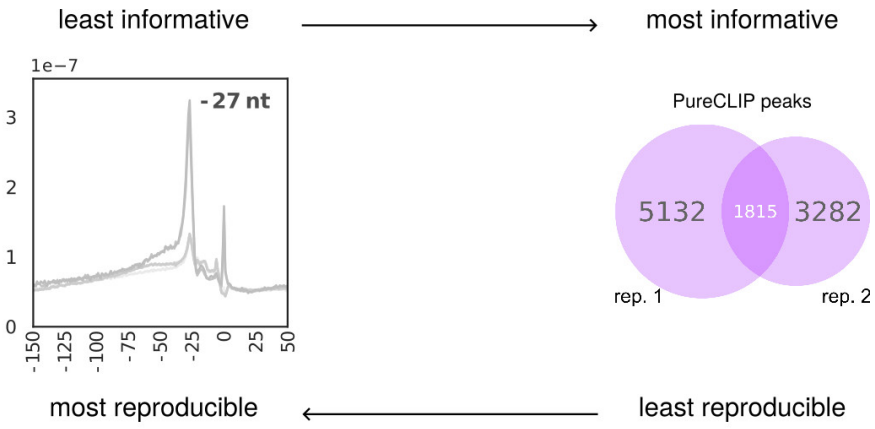

Figure 5.7: This illustration summarizes the main advantages and limitations of the recommended data analysis of CLIP-seq data. On one hand, Meta-exon plots (on the left) show reproducibly the specificity of the signal, but do not provide individual binding site information. On the other hand, peak detection (on the right) offers individual binding site information, but with a limited reproducibility rate.

Chapter 6

Reproducibility first: introducing an EJC-tailored pipeline

As shown in previous chapters, high-resolution binding site detection presents a low reproducibility rate. To overcome this limitation, we propose an EJC-specific and reproducibility-centered analysis pipeline. Our approach yielded better results than the more generalist conventional tools on EJC CLIP data.

6.1 Mining exon-level signal: the EJC Enrichment Score

We have presented the results and limitations of applying the recommended analysis pipeline to our data. In this section we will describe our first shift from the peak detection paradigm. This approach takes advantage of the data specificity revealed by the meta-exon plots (Fig. 5.7 and Fig. 6.1, left), and measures the signal enrichment at the exon-level. Because it focuses on the canonical EJC signal, we designated it EJC Enrichment Score, or EES. In this section we will detail how it is computed and present its performance on our data.

To measure the canonical signal enrichment in individual exons, we defined two 11-nucleotide long regions: a *canonical* region centered around the -27th nucleotide—spanning the -32nd to the -22nd positions upstream the 3'-end—, and a *non-canonical* region from the -15th to the -5th (Fig. 6.1, right). We chose the non-canonical region based on: a) the low enrichment showed by the meta-exon plot, and b) its presence in all exons regardless of their size (contrary to any region upstream of the canonical region, which may be absent from short exons). To exclude poorly covered regions, we only consider exons from genes whose read count is above the 90th percentile of each data set. To compute the EJC Enrichment Score (EES), we count the read 5'-end overlapping each region and calculate the ratio between the canonical value over the non-canonical (Equation 6.1). By comparing specific over non-specific signal in a particular exon, the EES aims to distinguish the EJC signal from background noise and identify EJC-loaded exons.

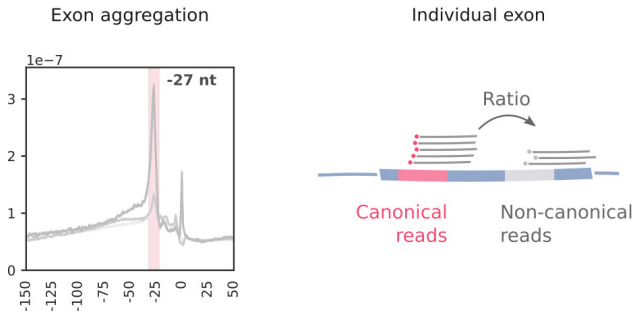

Figure 6.1: As shown on the left, aggregation of the signal detected in all exons reveals a sharp enrichment in a region around the -27th nucleotide. For individual exons, represented on the right, we compute the ratio between the number of reads inside this enriched region (canonical reads), and a downstream region with no apparent signal enrichment (non-canonical reads). The resulting value is the EJC Enrichment Score.

$$EES_i = \frac{n_{i,c}}{n_{i,nc} + 1} \tag{6.1}$$

Equation 6.1: The EJC Enrichment score of an exon i corresponds to the ratio between the number of read 5'-ends in the canonical region n_c, over the number of read 5'-ends in the non-canonical region n_{nc}. To avoid division by 0 when there is no signal in the non-canonical region, we systematically add 1 to the read count in the denominator.

We applied EES calculation on our CLIP data sets, as well as on the input and RNA-seq data sets to control the non-specific detection. Initially, we tested several EES thresholds to obtain sets of EJC-enriched exons. The idea was to find the threshold value that would optimize the Jaccard index of enriched exons between two replicates. However, as shown in Fig. 6.2, increasing the EES threshold consistently decreases the pair-wise Jaccard index. Thus, we arbitrarily defined exons with EES > 2 as enriched.

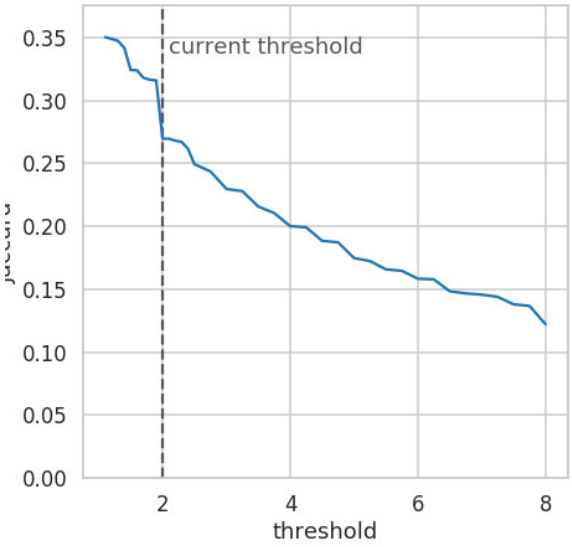

Figure 6.2: After computing the EES score on CLIP data sets, enriched exons were selected with different EES thresholds. Next, the Jaccard index between two sets of enriched exons was computed for each threshold. Instead of reaching an optimum Jaccard value, the Jaccard curve consistently decreases with higher EES values. The dotted line shows the chosen threshold to select enriched exons.

First, we assessed the number of enriched exons detected in CLIP and control data sets. To compare this strategy to peak detection with PureCLIP, we selected exons with at least one peak overlapping the canonical region as PureCLIP enriched exons. As observed in Fig. 6.3, EES yields between 4612 and 4807 enriched exons in CLIP data sets, whereas input controls yield between 3.5- to 7-fold less enriched exons. On the other hand, the number of PureCLIP exons is systematically lower than its EES counterparts. These results suggests EES offers a higher EJC-specific sensitivity than PureCLIP peaks.

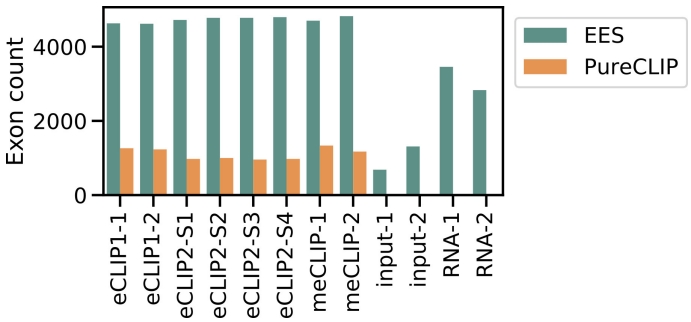

Figure 6.3: For CLIP data sets, we show the count of EJC-enriched exons (EES > 2), and of exons with at least 1 PureCLIP peak in the canonical region (from the -32th to the -22nd nucleotide). PureCLIP counts on control data sets are not shown because it is not designed to run on these data.

Surprisingly, RNA-seq data sets yielded only between 1.3 to 1.7 times less *enriched* exons than CLIP data sets. To assess whether this was due to a consistent accumulation of transcriptome reads in the canonical region, we compared the global distribution of EES values in all data sets (Fig. 6.4). We observed that scores were significantly higher in CLIP data sets than in both input controls and RNA-seq ($P < 0.05$), suggesting that there is local variation in the RNA-seq signal between the canonical and non-canonical regions, but that does not necessarily reflect EJC-enrichment. These results confirm that the EES retrieves EJC-specific signal from CLIP data.

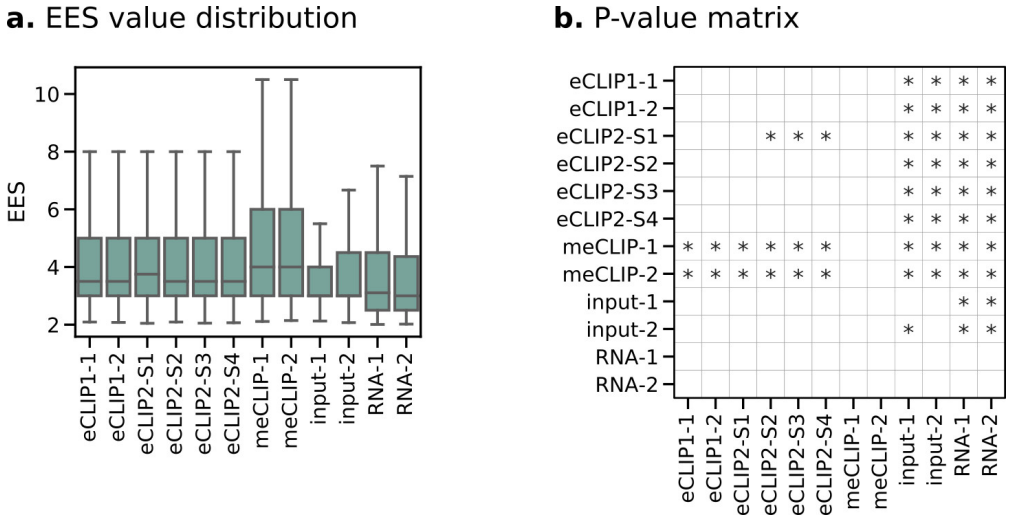

Figure 6.4: **a.** Distribution of EES values in CLIP and controls. Only exons with EES > 2 are presented. **b.** Mann-Whitney test p-value matrix. Pairwise statistical test results between data sets. Each cell represents the p-value returned by the Mann-Whitney test. * $P < 0.05$.

To assess the reproducibility rate of enriched exons, we computed pair-wise Jaccard indexes between all data sets (Fig. 6.5). For CLIP data sets, we observe slightly higher values within EES enriched exons than within PureCLIP exons (PC), with average Jaccard values of around 0.28 and 0.24 respectively. When comparing EES exons to PureCLIP exons, we obtain even lower values (0.12 on average), even for data sets stemming from the same library. Finally, comparisons between CLIP and control results show low values ranging from 0.02 to 0.06, regardless of the detection strategy. Interestingly, Jaccard values between RNA-seq replicates were higher than those between CLIP replicates. This suggests that the local coverage variations in RNA-seq that result in higher canonical signal are moderately reproducible. Taken together, these results show that canonical exon results are more reproducible than individual binding sites, which confirms the power of focusing on the specific EJC signal. Moreover, the low overlap between CLIP and control data sets (average Jaccard below 0.05) confirms the specificity of our results. Notably, the low overlap between EES and PureCLIP results (average Jaccard below 0.1) suggests that the two approaches find different sets of exons, and their results may be complementary.

In summary, in this section we have presented an alternative approach to detect EJC-loaded exons. With a simple fold-change calculation between two small regions of the exon, we quantify the signal and obtain EJC-specific results in our CLIP data sets. Although quantifying exon-level enrichment is a step forward compared to the qualitative nature of the meta-exon, we observe that reproducibility improvement

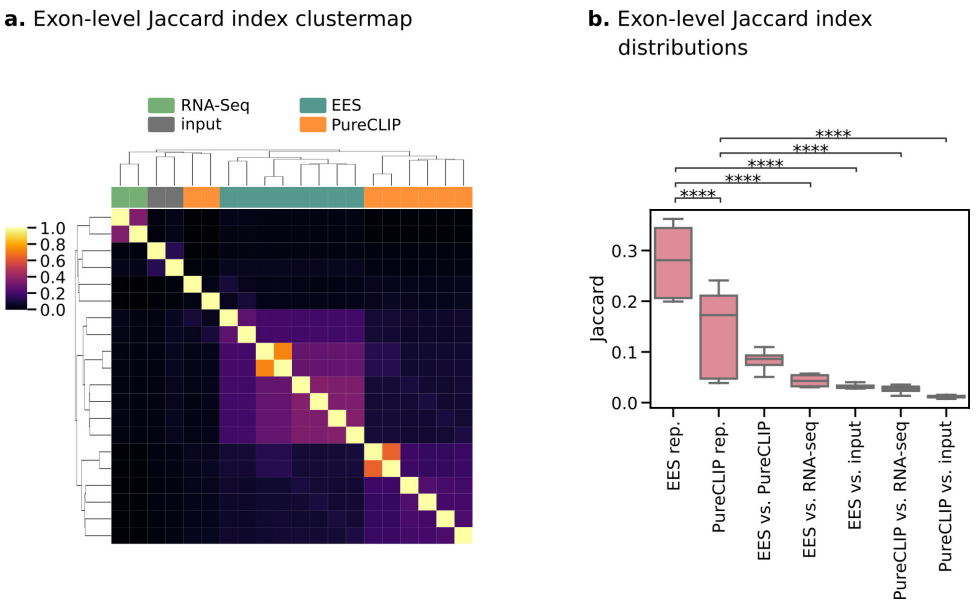

Figure 6.5: a) Jaccard matrix of enriched exons: each cell corresponds to the fraction of commonly enriched exons between a given pair of data sets. EES: enriched exons obtained with the EES strategy; PureCLIP: exons with at least one peak in the canonical region. For input and RNA-seq data sets, only EES enriched exons are shown. b) Jaccard index distribution among different comparisons, *EES rep.*: common exons between EES results in CLIP data sets; *PureCLIP rep.*: common exons between PureCLIP results in CLIP data sets; *EES vs. PureCLIP*: common exons between EES and PureCLIP results in CLIP data sets; *EES vs. RNA-seq*: common exons between EES results in CLIP data and EES results in RNA-seq data; *EES vs. input*: common exons between EES results in CLIP data and EES results in input data; *PureCLIP vs. RNA-seq*: common exons between PureCLIP results in CLIP data and EES results in RNA-seq data; *PureCLIP vs. input*: common exons between PureCLIP results in CLIP data and EES results in input data. **** P < 10-4 Mann-Whitney test.

is rather limited compared to the peak level. We found that over 70% of EES enriched exons are not reproducible, which corresponds to only a 10% reproducibility increment over peak detection. This prompted us to find a measure with even higher reproducibility rate.

6.2 Scoring EJC loading at the gene level: the Loaded Fraction

The EES approach has proven to mine EJC signal with higher sensitivity than PureCLIP. Although slightly improved relative to the peak level, exon-level reproducibility remains poor, as a large fraction of detected exons are not reproducible among CLIP data sets. In this section, I will present the Loaded Fraction (LF), a measure of EJC loading at the gene level. After assessing both its EJC specificity and reproducibility, we concluded that gene-level reproducibility is higher than both exon- and peak-level reproducibilities.

The LF value is the ratio between the number of EJC enriched exons over the total number of exons of a gene (see Equation 6.2). This results in a gene-level score of EJC loading, with values ranging from 0 (no loaded exons) to 1 (all exons loaded). For the sake of comparison, we computed LF values using EES-enriched exons (CLIP-EES), and exons with at least one PureCLIP peak in the canonical region (CLIP-PC). As controls, we used EES-enriched exons from input and RNA-seq data sets. When comparing the distributions, we observed significantly higher LF values in CLIP-EES data sets compared to input controls ($P < 0.05$), but not

when compared to RNA-seq data sets (Fig. 6.6). Similarly, CLIP-EES values were significantly higher than CLIP-PC values, even for the same libraries. This confirms the higher sensitivity of the EES strategy compared to PureCLIP. LF values higher than input controls confirm the specificity of the measure, although the similitude to RNA-seq values is puzzling.

$$LF_i = \frac{EnrichedExons_i}{TotalExons_i} \tag{6.2}$$

Equation 6.2: The Loaded Fraction of a gene i corresponds to the ratio between the number of Enriched Exons (EES > 2) in the gene over the total number of exons of the gene. For alternatively spliced genes, we select the isoform with the highest number of exons and the longest exonic length.

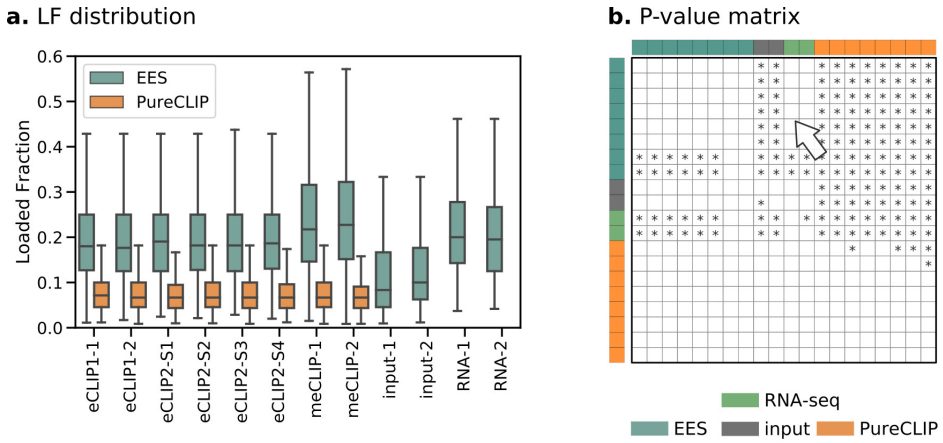

Figure 6.6: a) Distribution of LF values in CLIP and controls. Exons with EES > 2 or with PureCLIP peaks in the canonical regions were used to compute LF values per gene. b) Mann-Whitney test p-value matrix. Pairwise statistical test results between data sets. Each cell represents the p-value returned by the Mann-Whitney test. The arrow points at non-significant test results between EES and RNA-seq distributions. * $P < 0.05$.

Surprised by the similarity of LF values between CLIP-EES and RNA-seq, we computed pair-wise Jaccard indexes to quantify the level of overlap between data sets. As shown in Fig. 6.7, CLIP-ESS and CLIP-PC data sets present lower levels of overlap with RNA-seq data sets (average 0.26), than within each detection strategy (average 0.84 and 0.85, respectively). This reveals that the set of genes detected in CLIP is different from the ones detected in RNA-seq, despite their similar LF value distributions. Conversely, high Jaccard values between both detection strategies (average 0.88) show that they find similar sets of genes, despite the significantly higher LF values in EES genes. This results show that, although PureCLIP detection does not lack specificity, the EES strategy provides a higher sensitivity to EJC-specific signal.

Although the Jaccard index measures the level of overlap between two different sets, it does not quantify the reproducibility of the LF value itself. To refine the measure of reproducibility, we computed the pair-wise Pearson correlation coefficient of LF values of commonly detected genes using all data sets. As observed in Fig. 6.8b, correlation values among CLIP-EES replicates are significantly higher than CLIP-EES compared to RNA-seq (median 0.54 vs 0.06, respectively). This shows that for the small fraction of genes detected both in CLIP and RNA-seq, the exon loading detected in CLIP signal is not related to the detection in RNA-seq. Conversely, there is a higher correlation of LF values between CLIP and input controls

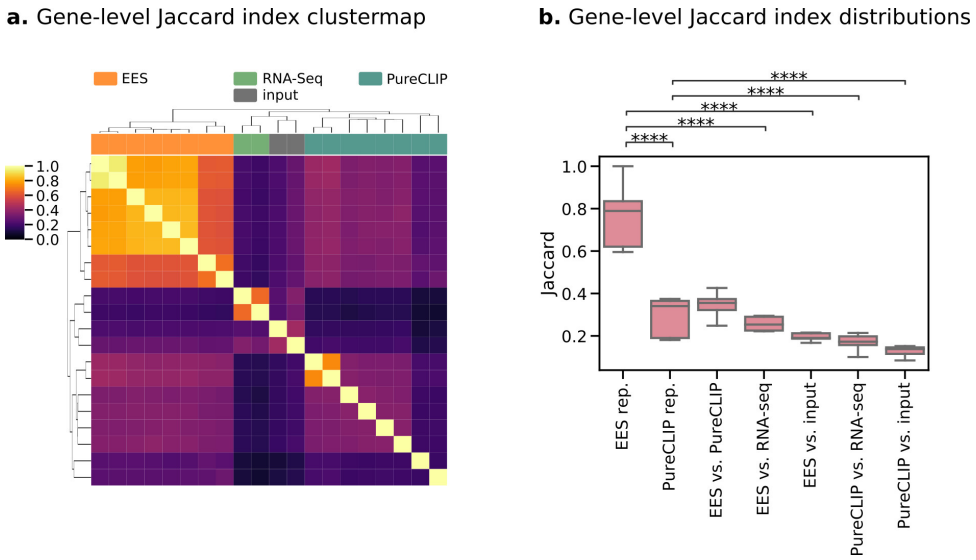

Figure 6.7: a) Jaccard matrix of loaded genes (LF > 0): each cell corresponds to the fraction of commonly detected genes between a given pair of data sets. EES: genes detected with the EES strategy; PureCLIP: genes detected with canonical region peaks. For input and RNA-seq data sets, only EES detected genes are shown. b) Jaccard index distribution among different comparisons, *EES rep.*: common genes between EES results in CLIP data sets; *PureCLIP rep.*: common genes between PureCLIP results in CLIP data sets; *EES vs. PureCLIP*: common genes between EES and PureCLIP results in CLIP data sets; *EES vs. RNA-seq*: common genes between EES results in CLIP data and EES results in RNA-seq data; *EES vs. input*: common genes between EES results in CLIP data and EES results in input data; *PureCLIP vs. RNA-seq*: common genes between PureCLIP results in CLIP data and EES results in RNA-seq data; *PureCLIP vs. input*: common genes between PureCLIP results in CLIP data and EES results in input data. **** P < 10-4 Mann-Whitney test.

compared to RNA-seq (average 0.26 vs. 0.06), despite the low exon-level overlap. This suggests that a fraction of input signal is detected in the same genes as CLIP signal, yet the relatively low correlation values show that LF values in CLIP data sets are EJC-specific.

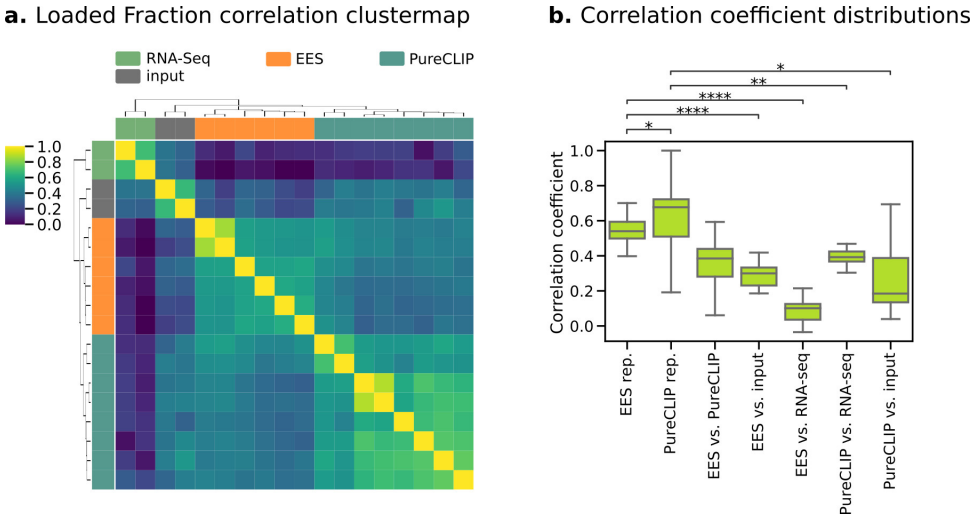

Figure 6.8: a) Pearson coefficient matrix of loaded genes (LF > 0): each cell corresponds to the Pearson correlation coefficient of LF between a given pair of data sets. EES: genes detected with the EES strategy; PureCLIP: genes detected with canonical region peaks. For input and RNA-seq data sets, only EES detected genes are shown. b) Pearson coefficient distribution among different comparisons, *EES rep.*: common genes between EES results in CLIP data sets; *PureCLIP rep.*: common genes between PureCLIP results in CLIP data sets; *EES vs. PureCLIP*: common genes between EES and PureCLIP results in CLIP data sets; *EES vs. RNA-seq*: common genes between EES results in CLIP data and EES results in RNA-seq data; *EES vs. input*: common genes between EES results in CLIP data and EES results in input data; *PureCLIP vs. RNA-seq*: common genes between PureCLIP results in CLIP data and EES results in RNA-seq data; *PureCLIP vs. input*: common genes between PureCLIP results in CLIP data and EES results in input data. * P < 0.05, ** P < 0.01, **** P < 10-4 Mann-Whitney test.

In conclusion, in this section we have proposed a gene-level measure of EJC detection. We found that gene-level Jaccard indexes between CLIP replicates were higher than exon-level Jaccard indexes (around 0.80 vs. around 0.28 on average, respectively). Furthermore, the correlation of LF values across replicates further confirms this gene-level reproducibility. Additionally, lower overlap and correlation values with both RNA-seq and input controls confirm the specificity of our strategy. This promising result motivated us to further explore the reproducibility of LF values, and thus extract the most reproducible information from our data.

6.3 Selecting reproducible LF values: Reproducibly Loaded Genes (RLG)

In the previous section we have shown a way to quantify the EJC detection at the gene-level with the Loaded Fraction (LF) score. We have shown that LF values of commonly detected genes are correlated across CLIP replicates. Yet, coefficient correlation values were rather moderate (0.54 on average), suggesting that among commonly detected genes, reproducibility of EJC loading remains limited. In this section, we will present how we selected *robust genes* using the reproducibility of LF values, then proceed to characterize them.

To measure the LF reproducibility, we computed the ratio between the LF value of each gene in a given pair of replicates (see Equation 6.3). Next, we needed to define a range of LF ratio values where a gene would be considered reproducibly detected. First, ratios were log_2-transformed to obtain symmetrical distributions centered around 0. Initially, we attempted to define a distribution range based on the standard deviation (SD) of the log-transformed ratios. However, the SD range boundaries varied from one pair of replicates to the other. Despite their variation, the boundaries were often close to a ratio value of 1.5. Thus, if the LF value in one replicate was less than 1.5 times greater than the LF value in the other replicate, a gene was considered to have a reproducible LF value.

$$ratio_{k,i,j} = \frac{LF_{k,i}}{LF_{k,j}} \tag{6.3}$$

Equation 6.3: The Loaded Fraction Ratio is the ratio between the LF value in replicate i over the LF value in replicate j for a gene k.

To define *reproducibly loaded genes* (RLG), we aggregated the information across replicates by counting the times a gene was reproducible between a given comparison. Thus far, our results suggest that meCLIP pseudo-replicates are the most different from other CLIP data sets. Initially, we aimed to keep a list of RLG whose LF values were reproducible in all CLIP data set comparisons. However, when including comparisons with meCLIP data sets, we obtained a list of no more than 30 genes. To have a broader view of EJC-loaded genes, we considered reproducibility only in eCLIP experiments, thus obtaining a list of 151 *RLG*.

Next, we aimed to determine whether RLG had any feature that would favor their detection across replicates. Thus, we compared the main characteristics of RLG to all expressed genes, and to the pool of genes detected in CLIP data (LF > 0).

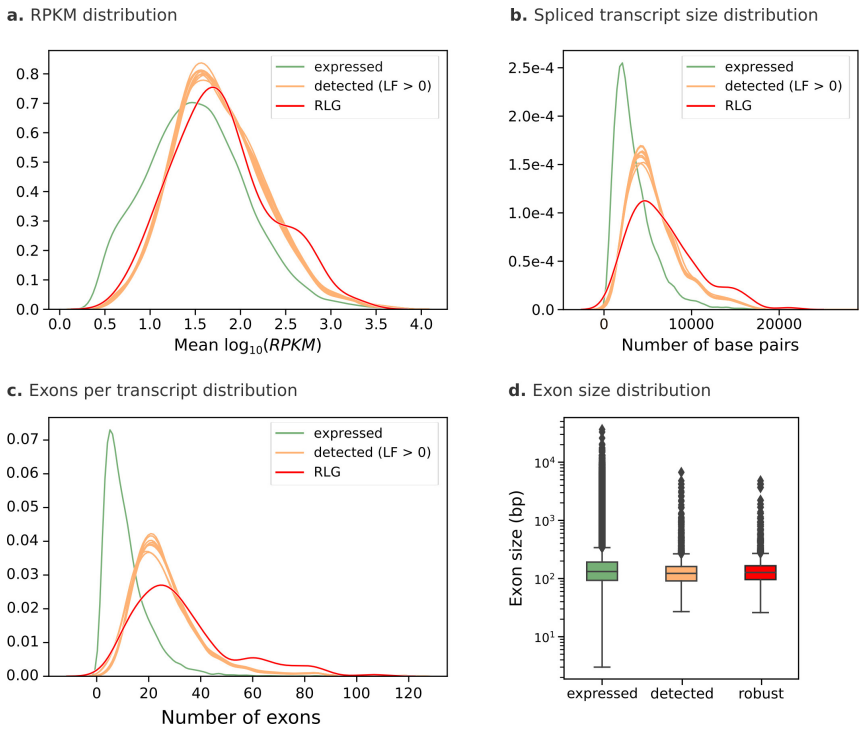

Figure 6.9: RLG were compared to all genes detected in CLIP (LF > 0, regardless of their reproducibility), and to all the genes expressed in HeLa cells. Comparison of distributions of a) transcript abundance, b) spliced transcript length, c) total number of exons, and d) exon size. As mentioned previously, for genes with multiple isoforms, only the one with the highest number of exons and the longest exonic size is used.

First, we used the RNA-seq data to quantify transcript abundance (RPKM), and plotted the distribution in each group of genes (Fig. 6.9a). We observe that detected and robust gene abundance is slightly skewed towards higher values compared to expressed genes. Then, we compared the spliced transcript size distribution, and found that detected and robust genes were consistently longer than all expressed genes (Fig. 6.9b). We analyzed the number of exons per gene and the length of individual exons, and found that detected and robust genes had more exons than expressed genes, while their median exon length is comparable (Fig. 6.9c-d). This shows that the difference in transcript length is due to a higher exon number rather than a higher exon size. Interestingly, these features were more similar between robust genes and all CLIP detected genes, than between CLIP genes and expressed genes (regardless of their reproducibility). This suggests that these characteristics do not favor the reproducible detection of robust genes. Yet, our detection strategy appears to favor exons in genes with a higher number of exons than the genes expressed in the cell.

We have established a list of reproducibly detected genes that offered the possibility to dive back into the study of individual exons. Furthermore, we have shown that gene-level characteristics were not related to the robustness of our detection. In the following sections, we will focus on the reproducibility and the factors influencing EJC detection on the exons of robust genes.

6.4 Exon-level reproducibility reveals EJC detection is not stochastic

In previous sections, we have employed the Jaccard index as a measure of reproducibility. In this section, we will assess the reproducibility of enriched exon detection within RLG by comparing the pairwise Jaccard index values of all data sets. First, we corroborated the specificity of the signal by observing higher values among CLIP data set comparisons than comparisons with control data sets (Fig. 6.10a-b). Importantly, we observe significantly higher exon-Jaccard values in RLG than in all genes with at least one loaded exon (LF $>$ 0) (Fig. 6.10b). This shows that highly reproducible genes present higher exon-level reproducibility as well, probably due to the removal of irreproducible noise stemming from other genes.

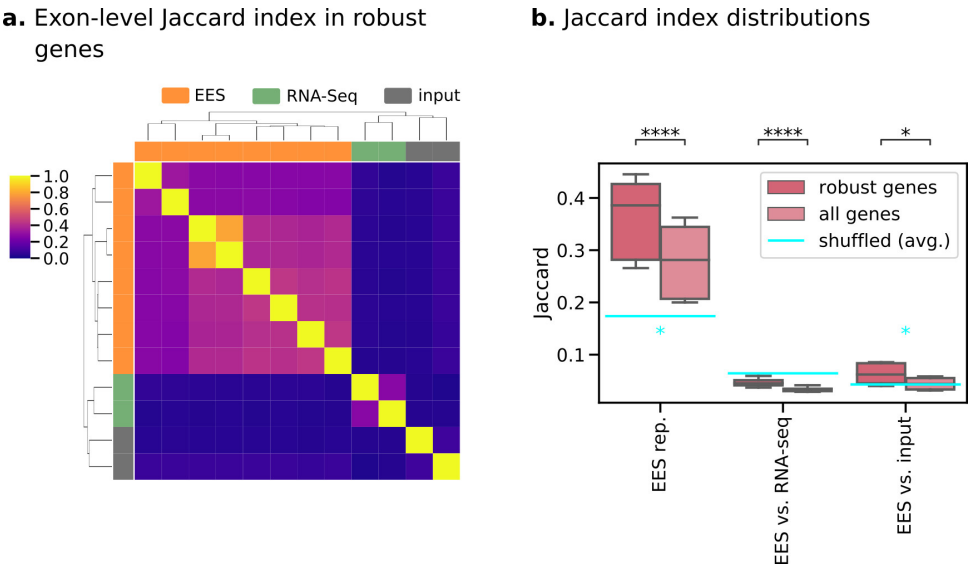

Figure 6.10: a) Exon-level Jaccard matrix of robust genes: each cell corresponds to the fraction of commonly detected exons between a given pair of data sets. b) Jaccard index distribution in robust and all detected genes; *EES rep.*: common exons between EES results in CLIP data sets; *EES vs. RNA-seq*: common exons between EES results in CLIP data and EES results in RNA-seq data; *EES vs. input*: common exons between EES results in CLIP data and EES results in input data. * P $<$ 0.05, **** P $<$ 10-4 Mann-Whitney test. In cyan, average Jaccard index in shuffled enriched exons. * P $<$ 0.05 permutation test.

Next, we aimed to prove the statistical significance of the observed Jaccard index values. We established an empirical null distribution of Jaccard values by 1) shuffling the position of enriched exons inside each robust gene, 2) computing the pairwise Jaccard index of commonly detected exons between all data sets, and 3) repeating this operation 1000 times. To test their significance, we compared observed values to the empirical null distribution and computed a p-value for each pairwise comparison. We found that all observed values were significantly greater than the null distribution (results summarized in Fig. 6.10b). This confirms that exon-level detection is not due to stochasticity.

The implications of this result are important. On one hand, it suggests that a reproducibly detected exon is likely to be loaded with EJC. Conversely, reproducibly *non-detected* exons are likely to be unloaded. Indeed, if non-detected exons were as likely to be detected in CLIP data, we would not observe any differences between the observed Jaccard values and their null distribution. Given that exon detection and non-detection is not stochastic, our next step was to investigate the reasons

behind this differential pattern.

6.5 Aggregating replicate information: The exon reproducibility score

Thus far, we have studied reproducibility in a pairwise manner with the Jaccard index. In order to analyze the exon level, we aggregated the information from all technical and sequencing replicates by computing the exon *Reproducibility score* (R score). It consists on counting the number of times a particular exon was loaded across the 8 replicates (Equation 6.4). Hence, an exon with R=8 is loaded in all replicates, whereas an exon with R=0 is loaded in none.

$$R_e = \sum_{i=1}^{N} 1_i(e)$$
$$1_i(e) = \begin{cases} 1, & \text{if } EES_{e,i} > 2 \\ 0, & \text{otherwise} \end{cases} \tag{6.4}$$

Equation 6.4: The Reproducibility score of an exon e is the number of times its EES value is greater than 2 for each replicate $i \in \{1...N\}$, where N is the number of CLIP replicates considered; in this case N=8.

We computed the R score for all the expressed exons of RLG, and counted the occurrences of each R score value (Fig. 6.11). We found that the majority of exons (n=2009) are reproducibly unloaded (R=0), whereas exons with higher R score values are more rare (205, 151, and 156 for R=6, R=7, and R=8, resp.). While the Jaccard index indicates that there is a global reproducibility rate, the R score is an assessment of individual exon reproducibility across all CLIP replicates. It reveals, on one hand, which exons are reproducibly loaded, and on the other, which exons are reproducibly unloaded.

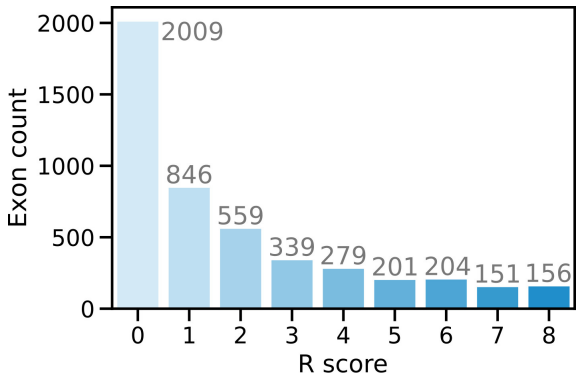

Figure 6.11: We computed the occurrence of exons with all possible values of the exon Reproducibility score (R). The maximum value is 8 because we considered exon loading across all CLIP data sets (n=8). R score was computed exclusively on exons from RLG.

After assigning a detection value to individual exons, we proceeded to further investigate the reason behind these detection patterns by correlating exon-level features to the R score values.

6.6 Canonical region T-content is directly related to robustness score

As discussed in section 6.4, reproducible detection of EJC-enriched exons is not stochastic. It is known that crosslink is more likely to occur between aromatic protein residues and uracil RNA bases (Ule, K. Jensen, et al. 2005; Krakau, Richard, and Marsico 2017). Hence, in this section we analyze and correct the influence of uracil content on the robustness score. Since we work with sequences from cDNA reads and the reference genome, we will refer to uracil content as T (thymidine) content.

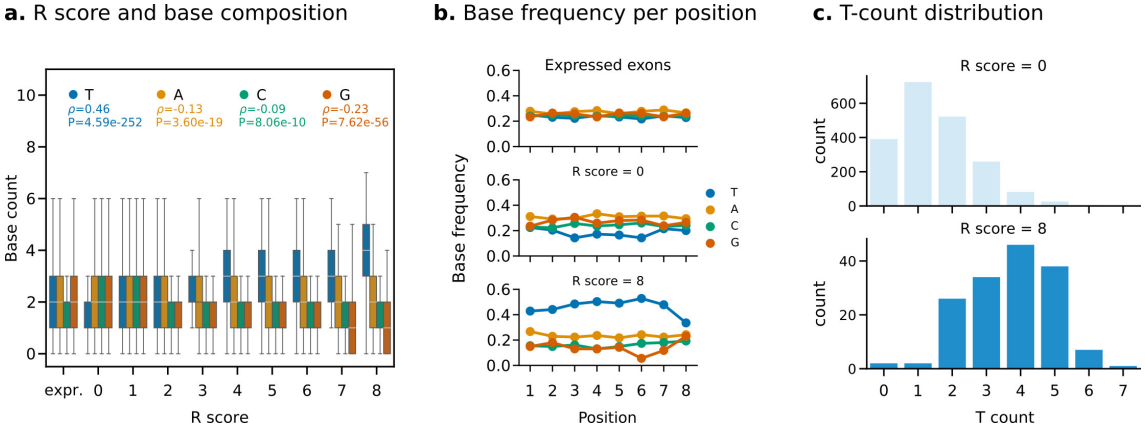

Figure 6.12: a) Base composition distribution in exons according to robustness score (*R score*) and in all expressed exons. We computed the occurrence of each nucleotide within a 8-bp window around the -27th position from the 3'-end of the exon. ρ: Spearman correlation coefficient between each base count and R score. P: p-value of the correlation coefficient compared to shuffled values. b) Frequency of each base per position computed with RSAT convert-matrix tool. c) Distribution of T counts in robustly non-detected exons (R = 0, light blue) and robustly detected exons (R = 8, dark blue).

First, we analyzed the base composition of the canonical region of RLG exons according to their R score value (Fig. 6.12a). We found that T content was significantly correlated to R score, with higher T counts in the canonical region with increasing R score values. Moreover, T counts were higher in robustly detected exons than in all exons expressed in the cell. We obtained the occurrence of each base along the canonical region to study whether the T content bias was position dependent. As shown in Fig. 6.12b, the occurrence of T is homogeneous along the canonical window of robustly detected exons, compared to robustly non-detected and expressed exons. This proves that the position within the binding region does not favor the occurrence of T. However, we do not observe a complete depletion of T in robustly non-detected exons, despite the relatively lower occurrences compared to other exons. When focusing on the distribution of T counts, we see an overlap between exons with opposite R scores (Fig. 6.12c). This suggests that, despite its strong effect, T content alone does not explain reproducible exon loading completely.

Knowing the relationship between T content and robustness score, we sought to reveal exons whose detection rate was independent of T content. One way to correct for this bias is to model the quantitative effect of T content on the R score. Thus, we fitted a simple linear model with the T count as the independent variable, and the R score as the response variable (Fig. 6.13a). The T-corrected R score, R_T, corresponds to the subtraction of the R score predicted by the linear model from the observed R score. With this strategy we detect a new distribution of robustness values independent of T-content. As shown in Fig. 6.13b, most exons have R_T scores

below 0, suggesting that they are more reproducibly undetected than expected by their T content. Conversely, several exons have R_T scores above 0, showing that they are detected at a higher rate than expected by their T content.

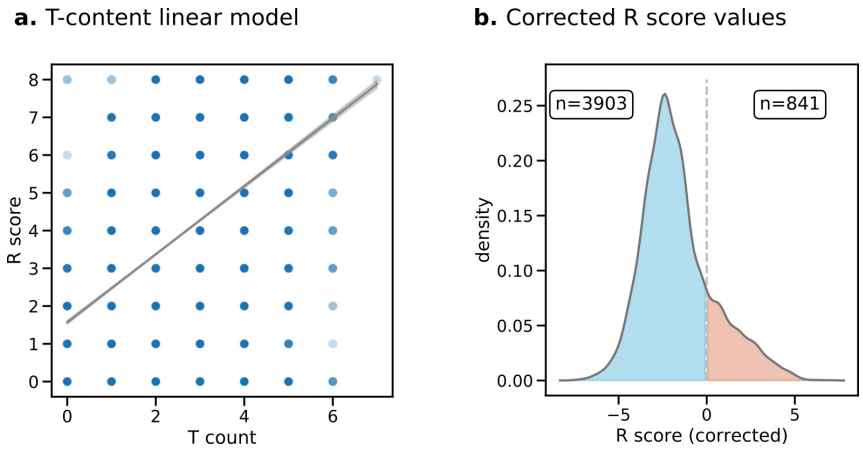

Figure 6.13: a) Linear model of T content as the explanatory variable and R score as the response variable. We corrected the R-score class imbalance by sub-sampling N exons from each class, where N was the minimum number of occurrences of a class (N=151, in our case). We fitted a simple regression model on the sub-sampled data and iterated 100 times to counter sub-sampling variance. The gray line depicts the average prediction of all models, with the standard deviation of prediction. b) Distribution of corrected R scores, R_T. In blue the fraction of exons with $R_T < 0$, and in red the fraction of exons with $R_T > 0$.

Altogether, our measures of detection reproducibility are indeed influenced by the crosslink bias towards T-rich sequences. Yet, modeling the effect of T content on the rate of exon detection allows quantifying and correcting our measures. With this approach, we prove that both robust exon detection, and non-detection, also occur independently of crosslink sequence bias.

6.7 Corrected robustness score rarely correlates with exon abundance

Discovering the T content bias on the robustness score, prompted us to further investigate another source of bias: exon abundance. We hypothesized that more abundant exons are more likely to be crosslinked than their counterparts, and thus be detected at higher rates. To test this hypothesis, we used RNA-seq read counts as a proxy of exon abundance, and removed exons with zero counts from our analysis. Next, we computed Spearman coefficient of correlation between exon abundance and exon R_T score for each gene individually (Fig. 6.14a). Each coefficient computation has an associated p-value to assess the significance of the correlation. We separated genes with non-significant correlation (n.s.) from genes with p-values under a confidence threshold of 0.05. We found that out of the 149 genes under study, only 9 had significant correlation values (Fig. 6.14b). Among these genes, only one showed a positive—and rather weak—correlation, while the others showed moderate negative correlations. It should be noted that these p-values were not adjusted for multi-testing. Therefore, we cannot conclude that for these genes there is a correlation between exon robustness and exon abundance. Yet, because this correlation was not significant for most genes, we did not take exon abundance as a factor that biases exon detection.

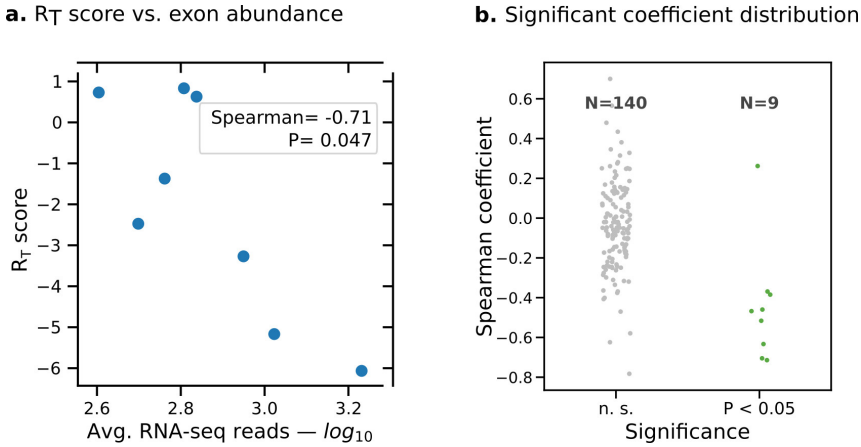

Figure 6.14: a) Comparison of RT score values and the number of RNA-seq reads of the exons in the *SRSF5* gene. The number of RNA-seq reads corresponds to the average between the two RNA-seq replicates described in Table 4.1. b) Distribution of Spearman coefficients of all robust genes according to the significance of correlation; *n.s.* not significant.

6.8 Conclusion

The CLIP protocol is challenging both at the experimental and the data processing levels. As discussed in section 3.1, the limited efficiency of both crosslink and IP hinder the generation of complex cDNA libraries. The main challenge is to retrieve a representative library of all mRNA targets of the EJC from a large pool of more abundant RNAs. Moreover, only a small percentage of the EJC-bound fragments are crosslinked, which hinders library complexity even further. Fixing the low complexity of CLIP libraries is beyond the scope of this work. However, we developed a data quality check strategy specific to EJC CLIP data. We combine the regular iCLIP pre-processing pipeline with additional NGS quality tools to better assess the complexity of our libraries. Performing small sequencing pre-runs and applying this data quality check, allowed us to accelerate the process of EJC CLIP optimization.

Regarding data processing, we found that current CLIP data analysis approaches show poor reproducibility and sensitivity on our EJC libraries. Thus, we developed an EJC-tailored data analysis strategy. We found that mining the signal within the specific binding region of the EJC yields more reproducible and sensitive results than conventional CLIP peak detection tools. Moreover, we developed several gene-level and exon-level measures to assess the reproducibility of EJC detection. This led to the definition of a robust gene list, where we distinguish the most reliable signal from noise. Next, we introduced the R score as a measure of exon-level reproducibility within robust genes across several replicate. We confirmed and corrected the T-content bias on the R score generated by crosslink, and discarded the influence of exon abundance on EJC detection. Altogether, our reproducibility-based approach has provided us with an unprecedented map of EJC-loaded and -unloaded exons.

Chapter 7

Studying the behavior of the EJC with CLIP data

In the previous section, we described the challenges that we faced analyzing EJC CLIP data and the strategy that we employed to obtain reproducibly loaded and unloaded exons. In this section, we attempt to use our measurements to gain some insight on the mechanism of EJC loading.

7.1 The EJC does not occupy all exons of a gene

One of the main questions regarding the EJC mechanisms in mammals is whether the EJC is present in all exons of a gene. Due to its splicing-dependent assembly, the targeting of specific junctions would imply the regulation of spliceosome composition at individual splicing events. Previous studies have indicated that not all exons of a transcript are loaded with EJC (G. Singh, Kucukural, et al. 2012; Saulière, Murigneux, et al. 2012, although they did not distinguish between canonical from non-canonical signal in their estimations. A differential EJC loading would imply that certain exons are targeted for EJC assembly, whereas others are not.

As discussed in section 6.5, our data analysis strategy revealed reproducibly detected exons within robust genes. Remarkably, we also discovered reproducibly *non-detected* exons (see Fig. 6.11). This is significantly different from a random configuration where every exons is equally likely to be detected or not detected (see Fig. 6.10b). Thus, the loading configurations that we observe within a gene are not due to stochastic EJC assembly or random detection across replicates.

Although robust detection is highly influenced by canonical region T-content, we observed differential exon loading after T-content correction (see Fig. 6.13b). This suggests that robust EJC *non-detection* is only partially explained by low T-content, probably due to the chemical nature of crosslink. Similarly, the abundance of an exon within a transcript cannot explain the EJC *non-detection*. Thus, the differential loading may not completely be due to technical biases.

Interestingly, both the R_T score distribution and the EJC loading map show that the majority of exons are not loaded within robust genes. Yet, further experimental validation is necessary to determine whether all these low-score exons are truly *unloaded*. We cannot dismiss a possible effect of library complexity on the detection of loaded exons, as will be discussed in the next section. Altogether, these results show a novel EJC loading map that indicates a differential exon loading along the gene.

7.2 Loading rate varies with sequencing depth, not transcript abundance

Previous EJC binding site analyses estimated a loading rate of maximum 80% of exons of detected genes. More importantly, these data suggested that loading rate was positively correlated to transcript abundance (G. Singh, Kucukural, et al. 2012; Saulière, Murigneux, et al. 2012). These results, however, were obtained with the lower resolution peak detection strategies and bypassing reproducibility. The higher resolution of our

CLIP data, and the higher reproducibility of our EJC detection strategy prompted us to reevaluate these previous estimations.

The eCLIP2-1 and eCLIP2-2 libraries were down-sampled to the exact same number of reads from eCLIP1-1 and eCLIP1-2 data sets respectively, due to a higher number of enriched exons (EES > 2), and overall higher LF values. The complexity of this library gave us the opportunity to assess the effect of sequencing depth on EJC signal detection. We thus established several fractions of both sequencing runs of eCLIP2 to compute EES values and select enriched exons. As shown in Fig. 7.1, the curve of enriched exons increases rapidly for small fractions, then approaches a plateau with higher number of reads. This means that if eCLIP2 were to be re-sequenced, increasing the sequencing depth would not reveal a dramatically higher number of enriched exons. More importantly, this curve shows the importance of sequencing depth in EJC signal detection.

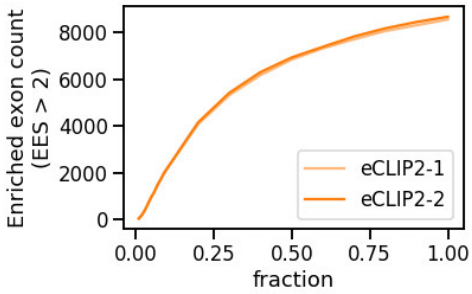

Figure 7.1: Number of EJC enriched exons detected in several randomly selected fractions of the most complex CLIP data set, eCLIP2

Next, we studied the relationship between Loaded Fraction and transcript abundance. Given the influence of sequencing depth on detection, we computed LF values for the genes detected in the sub-sampled fractions of eCLIP2 data sets. Then, we used the average RPKM values of detected genes and created equally sized bins of transcript abundance. We observed that LF values increased with higher transcript abundance only in fractions with 50% of total reads, but were independent of abundance in fractions with a higher number of reads (Fig. 7.2). Moreover, we found that the median LF values increased in higher fractions, independently of the relationship with transcript abundance. Thus, we first conclude that the EJC loading of a transcript is independent of its abundance. Secondly, the depth of sequencing, and certainly the library complexity, influence the gene-level detection of EJC loading.

These results prove the importance of assessing library complexity and the effect of sequencing depth on binding site detection. We have shown that sub-sampled fractions that are not close to the detection plateau yield a biased relationship between EJC loading and transcript abundance. This suggests that EJC detection is favored in highly abundant transcripts in CLIP libraries with low coverage. Furthermore, the overall distribution of LF values is influenced by sequencing depth as well, and thus should always be considered in the context of particular libraries before drawing conclusions.

Figure 7.2: Loaded Fraction distribution according to transcript abundance in sub-samplings of the most complex CLIP data set (eCLIP2). Transcripts were divided in 10 equally sized bins according to their average RPKM values from RNA-seq data sets. Only three representative fractions are shown alongside the results in the whole library.

7.3 Loaded genes are not functionally related

The functional relevance of the EJC has been proven in several model organisms, and confirmed in humans by genetic syndromes that affect the expression of its components. Therefore, analyzing the function of EJC loaded genes in a genome-wide study is not only interesting, but crucial to further characterize the functions of the complex within the cell.

To test whether there was a particular enrichment of functionally related genes among robustly detected genes, we ran the Panther software using default parameters (Mi et al. 2019; "The Gene Ontology Resource" 2019; Ashburner et al. 2000). We used genes detected in CLIP prior to selection by reproducibility (LF > 0) as background for the enrichment tests. We tested the enrichment of Gene Ontology terms (biological process, molecular function, cellular component), from both the GO consortium and the Panther database, as well as Reactome and Panther pathways. All enrichment tests were negative, regardless of database, showing no significant enrichment of known gene functional annotations among EJC loaded genes.

It should be noted that a) libraries were obtained from a cell culture of HeLa cells, and b) that we are not comparing different biological conditions. In other words, we are analyzing data that comes from a steady biological state. Our results suggests that in this context, EJC loading does not occur on a particular subset of genes. Yet, this does not prove that EJC loading does not occur on functionally related genes in cells under differentiation, for instance. A more informative study would consist in selecting the genes with a significant EJC loading variation between distinct biological contexts, and reassessing the enrichment of functional annotations.

7.4 Gene structure features do not correlate with EJC loading

In this section, we explore the correlation between EJC detection and exon definition, since the EJC is assembled and deposited on mRNA during splicing. First, we focused on the simple structure of robust genes by computing the correlation between the R_T score and exon and flanking intron sizes (Fig. 7.3a). We thus hypothesized that EJC loading was dependant on the strength of splicing site, rather than the basic structure of exons and introns. To test this hypothesis, we ran the MaxEntScan tool to score the splice junctions of robust genes exons. This tool proposes different algorithms to compute splice site strength using exon-intron junction sequences, from both the 5'- and the 3'-splice sites (5'-SS and 3'-SS, respectively). We found no significant correlation between EJC loading and splice site score, both 5'-SS and 3'-SS, regardless of the algorithm employed by the scoring tool (Fig. 7.3b-c). These results show that these *cis* features do not explain the EJC loading we find in our data.

7.5 The EJC does not have a preferred position inside the mRNP

It has been shown that EJC particles are removed by the ribosome at the first round of translation (Lejeune et al. 2002; Dostie and Dreyfuss 2002). As CLIP represents a snapshot of the processes being carried out in the cell, one might suspect that actively translated genes show a biased EJC distribution towards the 3'-end of the transcript. We represented the EJC binding landscape as a heat map of the R_T score across the exons of all robust genes (Fig. 7.4a), and did not observe any positional bias at first glimpse. We thus aggregated the information of all robust genes by comparing the R_T score to the relative rank of the corresponding exon inside the gene (Fig. 7.4b). Under the hypothesis of a positional bias, we would observe higher R_T values for normalized exon ranks closer to 1. Yet, we observe no clear relationship between exon rank and R_T score, suggesting that the EJC is not detected in preferentially towards the 3'-end of transcripts.

As there is evidence that suggests that loading on a given exon has an effect on neighboring exons (Ashton-Beaucage et al. 2010; Roignant and Treisman 2010; Malone et al. 2014), we refined the EJC localization study within a transcript by analyzing consecutive pairs of exons. We thus used a R_T threshold t to define two distinct exon-loading states: ***loaded*** if $R_T > t$, and ***unloaded*** otherwise. To determine the influence of EJC loading in consecutive exons, we defined a *state pair* as the couple made up of the state of a given exon with rank i, and the state of the exon with rank $i+1$ (Fig. 7.5a). We then counted the occurrence of all possible state pairs within a gene, and constituted an observed distribution of state pairs for all robust genes (Fig. 7.5b).

To test whether the observed counts were significant, we shuffled the ranks of exons within a gene and computed state pair counts, repeating the process 1000 times, thus establishing a null empirical distribution. Under the alternative hypothesis that observed counts were higher than the null distribution, we computed adjusted p-values and tested their significance with a confidence of $\alpha = 0.05$ (Fig.

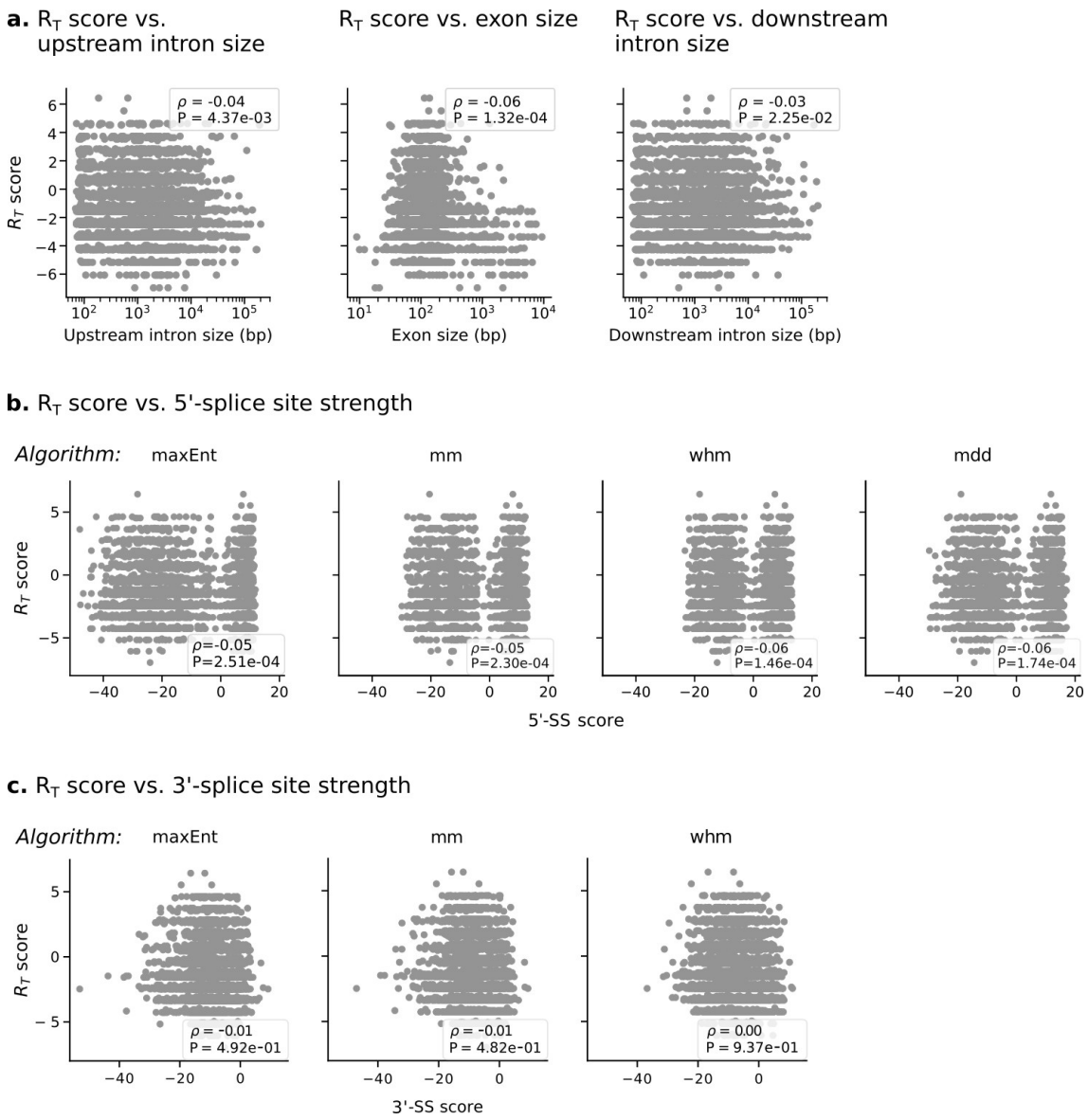

Figure 7.3: **a.** Correlation between corrected robustness score (RT) of an exon and the size of: the upstream intron (left panel), the exon itself (center panel), and the downstream intron (right panel). All sizes are in base pairs (bp). Correlation between corrected robustness score (RT) of an exon and **b.** the strength of the 5'-splice site, and **c.** the strength of the 3'-splice site. Splice site strength scores were obtained using the MaxEntScan tool. We applied all available algorithms on the sequences overlapping the splice junctions of all robust gene exons. ρ: Spearman coefficient of correlation. P: p-value associated with the coefficient. 5'SS: 5'-splice site. 3'-SS: 3'-splice site. maxEnt: Maximum Entropy Model; mm: first-order Markov-Model; whm: Weight Matrix Model; mdd: Maximum Dependence Decomposition Model.

7.5c); we performed this test for each state pair combination separately. To assess the impact of the threshold used to define exon states, we performed the test using threshold values ranging from -2.0 to +2.0, and counted the number of genes with significant state pair counts (Fig. 7.5d). We observed that the number of genes with significant state pairs that included loaded exons was high for relaxed thresholds, and decreased for more stringent thresholds. Conversely, significant unloaded-unloaded state pairs were not detected for relaxed thresholds, but increased dramatically for stringent thresholds. This shows that the significance of state pair counts is highly dependent on the threshold chosen to define the loading state. Thus, we cannot conclude that neighboring exons tend to have the same loading state.

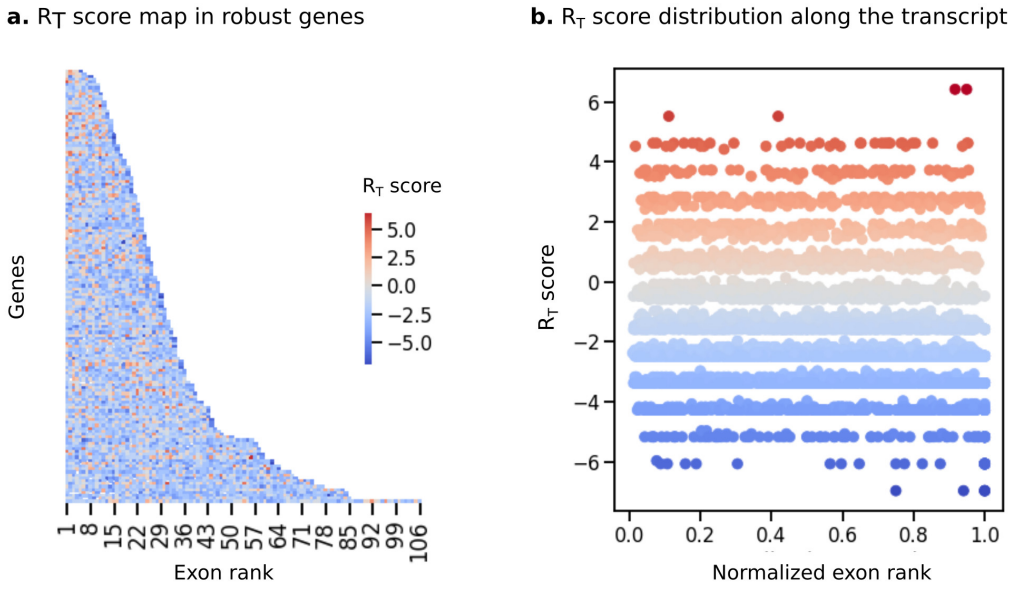

Figure 7.4: a) Map of the corrected robustness score (R_T) in robust genes ranked by number of exons. b) Comparison of RT score and the relative position of exons within a transcript; normalized exon rank corresponds to the rank of an exon divided by the total number of exons of a gene. Lower RT values are shown in blue; higher R_T values are shown in red.

7.6 Conclusion

In this chapter we have explored the EJC loading map obtained with the approaches described in chapter 6. Our main objective was to find features that could explain the EJC loading we detect with the R_T score. These explanatory features hold the potential to reveal the code that dictates the EJC deposition on exons. However, our analyses have only scratched the surface of what is seemingly a complex orchestration of factors.

We studied the relationship between transcript abundance in a steady state, but did not address transcription or degradation rate. We assessed localization bias within the transcript, but did not quantify their translation efficiency, nor did we study the relationship between the latter and EJC detection reproducibility. We assumed that EJC loading was localized in clusters of continuous exons, but the data does not support this assumption. The code behind EJC deposition thus remains to be determined.

Nevertheless, we now possess the tools to find the most reliable signal provided by CLIP data. We have paved the way to find the binding site map of the EJC. The advancements in library construction and the data analysis approach we present in this work, allowed to determine that the fraction of loaded exons depends on the quality of the data rather than reflecting a biological reality. We have proven the existence of both robustly detected and undetected exons, suggesting that the deposition of EJC along the exon junctions of a transcript is not homogeneous. Finally, although we ignore the reasons behind this observation, our results show that it is not a stochastic process.

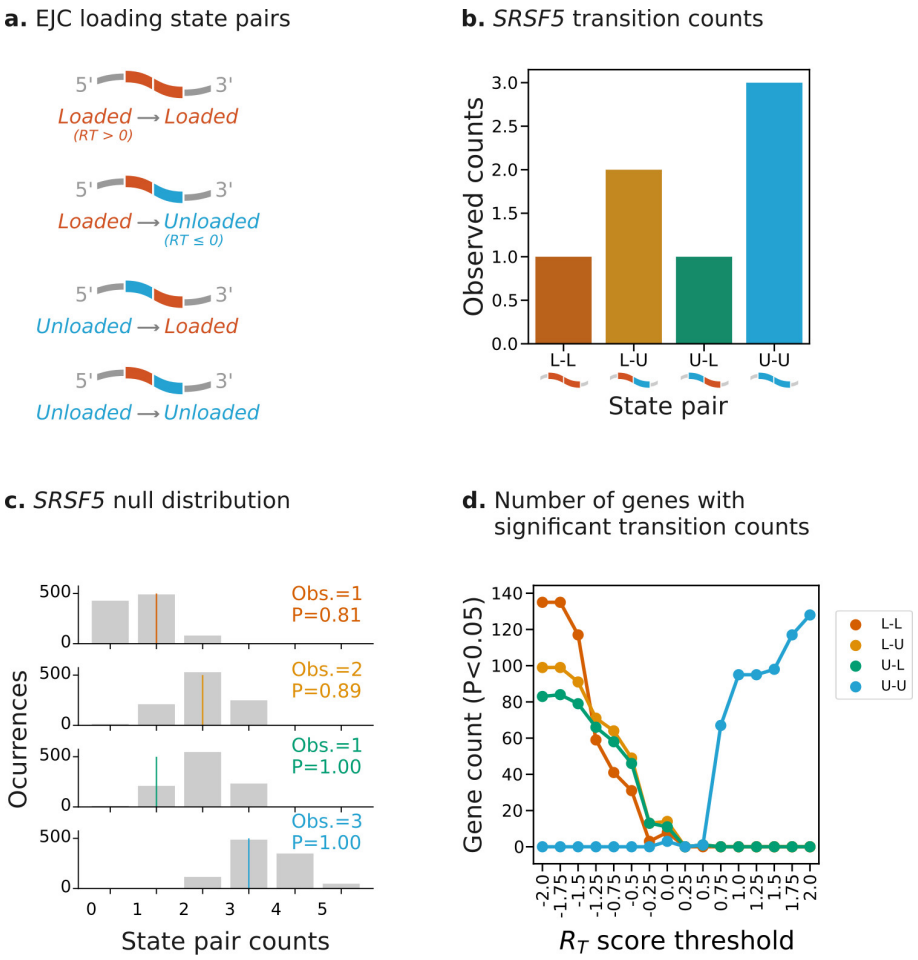

Figure 7.5: **a.** Definition of an exon loading state according to the RT score. In this example, an exon is considered as loaded if RT > 0, otherwise it is considered unloaded. A state transition compares the state of one exon to the state of the downstream exon. **b.** Distribution of observed state pairs in the gene SRSF5. **c.** Comparison of observed state pair counts to an empirical null distribution, with the corresponding adjusted p-value (FDR multi-test correction). **d.** Variation of number of genes with significant ($P < 0.05$) state pair counts according to different RT thresholds.

PART III

CONCLUSIONS AND PERSPECTIVES

Chapter 8

Discussion

PTGR is a complex network that requires RBP binding site determination to further understand it. This is especially true for the EJC, as it impacts various aspects of mRNA life. Establishing the EJC binding map is crucial to determine whether it regulates specific events of particular genes.

In this work, we present a dedicated strategy to mine EJC-specific CLIP-seq data while selecting the most reproducible signal. As a result, we have established an novel binding site map of the EJC within a set of reproducibly loaded genes (RLG). This allowed us to reach an important conclusion: EJC appears to be loaded specifically in *some* exons of human transcripts. Moreover, this occurs at a lower rate than previously estimated. Further analyses will unravel the functional implications of this discovery.

8.1 Lessons learned from CLIP-seq data analysis

Generating CLIP-seq libraries is a challenging task, especially when aiming to detect reproducible binding sites. From HeLa RNA-seq experiments, we estimated around 10 000 expressed genes (>2.3 RPKM), which translates to over 118,000 expressed exons. The eIF4A3 eCLIP data sets that we successfully obtained detect between 5,000 to 8,000 EJC enriched exons. This corresponds to 4% to 7% rate of detection, which strikes as relatively low despite mining signal from the canonical region.

This may be explained by the binding mechanism of the EJC. As discussed in section 2.2.2, eIF4A3 adopts a closed conformation on RNA and interacts with the ribose-phosphate backbone rather than with the nucleotide bases. Thus, it is probably harder to capture a considerable amount of EJC-RNA interactions, as UV-light irradiation is more likely to create links between RNA residues. Correspondingly, our results suggest that uracil content promotes robust EJC loaded exons. Again, UV-light irradiation is more likely to create covalent bonds between pyrimidines and (mostly) aromatic residues. Thus, combined with the interaction between eIF4A3 and RNA, the chemistry of UV crosslink may explain the overall low sensitivity of binding site detection.

Yet, eIF4A3 is not the only EJC protein that contacts RNA. As CASC3 directly contacts the base of one nucleotide (Andersen et al. 2006), it may be contemplated as a more efficient CLIP candidate than eIF4A3. However, recent reports suggest that CASC3 may not be a constitutive core component on all genes, with notably a functional implication in NMD-sensitivity Mabin et al. 2018. Thus, *CLIPping* CASC3 might reveal a sub-population of EJCs with specific roles, rather than wider panorama of EJC-mediated regulation.

In this work we have set up a pre-processing strategy that accelerates the EJC CLIP library production. We now estimate the complexity and signal specificity of libraries in a pre-sequencing runs. This allows us to focus on the data sets with the potential to represent the majority of EJC-loaded exons, while discarding poor-quality or failed CLIP libraries early in the process. It is thus a matter of time before we obtain better EJC CLIP libraries that reveal a higher number of robustly

detected genes—either in human, drosophila, or other species.

8.2 Beyond the EES

With the EJC enrichment score, we propose a strategy that specifically detects exon-level enrichment of EJC signal. Furthermore, we propose an approach that prioritizes reproducible detection by: a) selecting robustly detected genes with similar Loaded Fraction values, and b) establishing the *Robustness* score as a way to quantify exon detection across several replicates. This strategy allowed us to reach an unprecedented conclusion: the EJC is not detected in all exons of human genes. This suggests that EJC assembly on specific junctions is a regulated process, and entails specific regulatory consequences for transcripts.

Our strategy, however, presents some limitations. First, only the top 10% of genes with the highest number of reads are selected prior to EES computation. With this first filter, we aimed at excluding poorly covered genes and background noise. In practice, however, this skews the distribution of detected gene length towards longer genes, as they are more likely to have a higher amount of reads than their shorter counterparts. In truncation-based peak calling approaches, the number of read 5'-ends is compared to an empirical background and statistically assessed with a Chi-square or Fisher exact test (Lovci et al. 2013; Shah et al. 2017; Chakrabarti et al. 2018). Similarly, EJC enrichment may be assessed by establishing a contingency table with the number of reads inside the canonical and non-canonical regions of a particular exon, and the total number of reads in all canonical and non-canonical regions of all the exons of the gene (see Table 8.1 below). Then, the fraction of canonical reads can be assessed with the appropriate test to determine whether the enrichment is significant.

Table 8.1: Contingency table of the number of canonical & non-canonical reads at the exon and gene levels. n: number of read. c: canonical region. nc: non-canonical region.

	Canonical	**Non-canonical**	**Total**
Exon (i)	$n_{i,c}$	$n_{i,nc}$	$n_{i,c} + n_{i,nc}$
Gene (with k exons)	$\sum_{j=1}^{k} n_{j,c} - n_{i,c}$	$\sum_{j=1}^{k} n_{j,nc} - n_{i,nc}$	$\sum_{j=1}^{k} n_{j,c} + \sum_{j=1}^{k} n_{j,nc} - (n_{i,c} + n_{i,nc})$
Total	$\sum_{j=1}^{k} n_{j,c}$	$\sum_{j=1}^{k} n_{j,nc}$	$\sum_{j=1}^{k} n_{j,c} + \sum_{j=1}^{k} n_{j,nc}$

In this setting, the signal within each exon is compared to the signal of the gene that contains it. Assessing each exon in its particular context instead of filtering by gene coverage may prevent the gene size bias that we encountered. Moreover, it may reduce false positives caused by background noise in long genes, and reveal enriched exons in shorter genes. Implementing this statistical framework may be a perspective for future work.

Another limitation of our approach is the exclusion of input controls from the computation of enrichment. We computed EES values on input data sets separate from CLIP data sets, then compared the number of detected exons and computed LF correlations to conclude specific EJC detection in CLIP compared to input. However, this does not control the significant enrichment of individual exons relative to input. Future work should consider how to incorporate input signal into the computation of EES values. Possible ways include: a) establishing a model with input signal as a confounding variable (as done by PureCLIP, Krakau, Richard, and Marsico 2017), or b) computing an *a posteriori* EES fold-enrichment (as done by CLIPper, Lovci et al. 2013). However, high EES values in input may not necessarily correspond to non-specific enrichment. Since we consider the reads inside the canonical region, part of the EJC signal present in the SM-input may be detected. Further thought must be put into the role of input control in the EES strategy.

Finally, the EES value is EJC-specific. This strategy takes advantage of the specific deposition site relative to the exon junction. We define a canonical region where the EJC signal is expected to be greater, and a non-canonical region to compare with. Thus, this approach cannot not be directly extrapolated to all RBPs, especially those with less specific spatial binding. An equivalent of EES may be computed for RBPs with known binding motifs or specific locations within a transcript. It may not be as simple for RBPs with more dynamic binding modalities. For instance UPF1, which is a helicase that translocates along RNA, yields a CLIP signal that is broadly distributed along the 3'-UTR (Hurt, Robertson, and Burge 2013; Zünd et al. 2013). Thus, although our approach mines specific signal, it *sacrifices* its direct generalization to all RBPs.

8.3 The amount of EJC per gene

In this work, we estimate the amount of EJC per gene with the Loaded Fraction (LF) value, computing the ratio of enriched exons over the total number of exons of a gene. We mainly use LF to select similarly loaded genes and constitute a list of robustly detected genes. Yet, this value is useful to compare previous estimations of EJC-occupied exons per gene. Saulière and colleagues performed HITS-CLIP of eIF4A3 and ran FindPeaks, a ChIP-seq peak caller, to detect binding sites (Fejes et al. 2008). They found 50% of peaks were outside the canonical region (Saulière, Murigneux, et al. 2012), and estimated that 80% of exons contained at least one peak (canonical or non-canonical). Our own estimation is not as straight-forward. We found that LF values were dependent on the complexity and sequencing depth of the CLIP library, as higher sequencing depth results in higher median LF values. In comparison, we found a maximum 30% of loaded exons per gene on average. It should be noted that our strategy only considers enriched signal in the canonical region. This means that a) we cannot estimate the percentage of non-canonical signal enrichment, and b) we do not take into account non-canonical loaded exons to compute LF values. However, the sharp canonical enrichment in our data suggests that the proportion of non-canonical signal may be lower than previous estimations. This is in agreement with the results that suggested non-canonical EJC corresponds to co-purification of its partners (G. Singh, Kucukural, et al. 2012). Altogether, our estimation indicates a con-

siderably lower EJC deposition rate on exons, which can have important functional implications for individual transcripts.

8.4 A sequence bias with biological implications?

A recent publication found a sequence-bias in splicing regulation (Lemaire et al. 2019. Knock out of multiple splicing factors followed by RNA-seq identified two distinct populations of regulated exons: one of GC-rich and flanked by short introns, the other AT-rich and flanked by long introns. Splicing factors involved in GC-rich exon splicing were found to be involved in unwinding or preventing the formation of RNA secondary structures, which would otherwise hinder spliceosome recognition. On the other hand, splicing factors involved in AT-rich exon splicing were found to prevent spliceosome recognition of spurious polypyrimidine tracts and branching points. The genes containing these exons shared the same sequence bias, i.e. GC-rich exons were part of GC-rich genes, and AT-rich exons were part of AT-rich genes. These exons were also found within genome *isochores*—regions with uniform GC or AT content. This indicates that RNA processing undergoes specific pathways according to sequence composition.

In this work, we primarily associate sequence content bias to the chemistry of crosslink. We claim that higher T-content favors robust detection of enriched exons across replicates. However, we cannot completely discard a biological implication of sequence bias in our results. Testing the correlation between robustness score and gene-level GC-content, and chromatin domains may reveal a potential biological cause for the observed sequence bias in EJC CLIP data. It should be noted, however, that our data only suggests a positive T-content correlation, rather than an AT-content correlation. This may indicate a strong crosslink effect that may overshadow other meaningful sequence bias.

8.5 Displaced to the 3'-end?

EJC disassembly is translation dependent. In mammals, it is a process mediated by the PYM protein and associated to the first round of translation. At the moment of crosslink, the ribosome may have removed the EJC deposited in the mRNP 5'-end, without yet reaching the 3'-end. In this context, one would expect to observe an EJC distribution bias towards the 3'-end of the transcript. Yet, our data on RLG does not support this notion.

This observation may have two explanations. The first is that the *in vivo* snapshot that we obtain with CLIP mainly captures transcripts that are not actively translated, showing heterogeneous binding sites that do not follow a global pattern along the transcript. The second explanation is that we observe the profile of a population of transcripts that undergo asynchronous translation. These explanations may not be mutually exclusive, as some transcripts may be under translation arrest during their transport to specific sub-cellular locations (Johnstone and Lasko 2001),

although this may not be the case for all mRNPs. Translation efficiency can be computed with ribosome profiling data (Ribo-seq), which quantifies the RNA fragments protected by active ribosomes (Ingolia 2014). EJC CLIP data needs to be crossed with translation efficiency data to determine which explanation suits our observations.

Consecutive EJC loading

In *D. melanogaster* and human, the presence of an EJC has an impact on the splicing of neighboring junctions (Ashton-Beaucage et al. 2010; Roignant and Treisman 2010; Malone et al. 2014; Blazquez et al. 2018; Boehm et al. 2018). In human cells, neighboring exons communicate through multimeric EJC higher-order interactions that include other RBPs, resulting in highly packaged mRNPs. Taking this into account, we explored whether consecutive exons were more frequently loaded with EJC in human.

Our data does not suggest clusters of consecutive loaded exons within a transcript. One possible explanation is a variable stability of the EJC in certain junctions, even in a scenario where it is assembled uniformly across the transcript. Another possibility is that EJC deposition on one exon does not influence the deposition on neighboring exons in human genes. In this scenario, the EJC-mediated regulation of splicing effects may follow more complex patterns than the ones observed in *D. melanogaster*, where EJC deposition facilitates exon definition of weak splicing sites.

Nevertheless, further correlation tests may be performed using our binding map. In this work, we only assessed the correlation between the R_T score and the splicing site strength of the detected exon. Yet, we did not assess the correlation between EJC detection and the splicing strength of flanking exons. As suggested by a recent study, splicing of neighboring introns is a coordinated process, both in human and drosophila (Drexler, Choquet, and Churchman 2020). Therefore, studying the splicing features of neighboring junctions may reveal whether there is any impact of EJC deposition on the splicing of neighboring junctions in human.

8.6 How do EJC-loaded exons regulate splicing?

Applying our dedicated CLIP data analysis approach resulted in an EJC binding map for a set of robustly detected genes. Nevertheless, computational approaches are only as worthy as the experimental validation that follows. Our EJC binding map identifies EJC-loaded genes that may be used as reporters to elucidate the role of the EJC in their regulation. Inspired by previous reporter assays, one may constitute intronless variants of robust genes to test the impact of the absence of EJC in specific exons.

Another interesting approach would be to intersect the list of robust genes with the EJC-dependent splicing events observed by Wang and colleagues upon knock-down of core components (Z. Wang, Murigneux, and Le Hir 2014). Additionally, it would be interesting to study potential cryptic splicing sites within the robustly

detected genes, that may be subject to recursive splicing (Blazquez et al. 2018; Boehm et al. 2018). This would allow to correlate the detection of the EJC to the regulation of constitutive splicing, as well as the repression of recursive splicing in human cells.

Finally, the current model for EJC-mediated splicing regulation proposes a transcription slow-down caused by the EJC assembly. This would facilitate the recognition of weaker splicing sites and ensure proper splicing (Le Hir, Saulière, and Z. Wang 2016). To further investigate this model, our binding map may be compared to nascent transcript sequencing data (NET-seq, or GRO-seq), to estimate the correlation between EJC deposition and the transcription rate of robust genes. Along with the analyses mentioned above, this would allow to elucidate the EJC role in splicing regulation in humans.

8.7 Elucidating a *simpler* code?

The study of the EJC-mediated localization of the *oskar* mRNP in *D. melanogaster* revealed that only splicing of the first intron was necessary for correct oocyte polarization (Ghosh, Marchand, et al. 2012). Furthermore, screening assays revealed EJC-mediated splicing of weak intron splicing sites of the *mapk* and *piwi* transcripts Roignant and Treisman 2010; Malone et al. 2014). These highly specific roles in *D. melanogaster* suggest straight-forward roles of the EJC in gene expression regulation, which have not been observed in human cells.

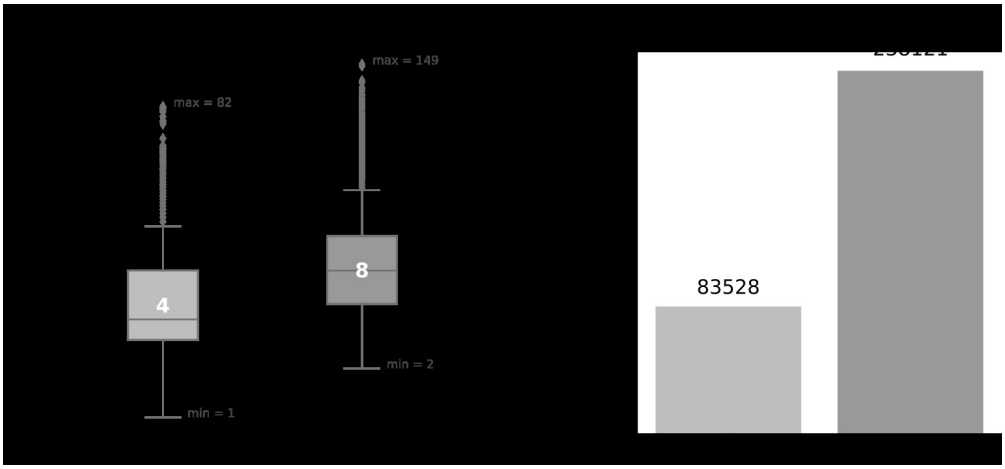

Figure 8.1: a. Distribution of the number of exons per transcript in *D. melanogaster* and *H. sapiens*. b. Total number of annotated exons in *D. melanogaster* and *H. sapiens*. Drosophila annotations were obtained from the *FlyBase* consortium, version BDGP6. Human annotations were obtained from Ensembl, version GRCh38; only the longest isoform and with experimental evidence were used for quantification.

The regulation of the *D. melanogaster* transcriptome may be less complex than human. Using genome annotations, we found that overall, drosophila transcripts contain less exons than human transcripts. Indeed, the median coding gene contains 4 exons in drosophila, half of the median gene in humans (Fig. 8.1a). Moreover, the total drosophila transcriptome contains approximately 3 times less exons than the human transcriptome (Fig. 8.1b). Finally, 90% of drosophila genes have under 10 isoforms (Brown et al. 2014), whereas estimates in human suggest that the average gene generates at least 7 (Pan et al. 2008). This suggests that drosophila transcripts

are less prone to complex alternative splicing patterns than human genes. Thus, in addition to a less complex regulation, the drosophila transcriptome may be easier to cover with CLIP.

Both the EJC known mechanisms and the characteristics of the drosophila transcriptome make it a promising system to study the binding site landscape of the EJC. Applying the strategy presented in this work to drosophila cells would result in a reproducible binding site map. Under simpler rules of deposition in this organism, it may be more straight-forward to model the mechanisms of regulation and functional implications of EJC binding on specific exons.

8.8 Conclusion

CLIP-seq techniques are an approach to discover the transcriptome-wide binding modalities of RBPs. However, they are a challenging endeavor at the experimental and data analysis front.

CLIP protocols involve multiple steps that span several days. Obtaining a high-quality library once all steps have been performed depends on the nature of the protein of interest and numerous factors. This requires investing time and energy into optimizing many experimental conditions, specifically for the targeted RBP. Here, we implemented a data pre-processing strategy that accelerated the optimization process for eIF4A3. Yet, much work remains to optimize CLIP for other proteins and in other organisms.

The experimental results impact directly the quality of the data for analysis. Regardless of the quality, however, data analysis is not straightforward. The main limitation in the detection of single-nucleotide binding sites is reproducibility. We have shown that binding sites detected with CLIP peak callers have limited reproducibility. We thus developed a dedicated strategy to mine EJC CLIP data in a reproducible way, which may be extrapolated to some other RBPs with similar binding modalities.

Our method allowed us to re-evaluate with high confidence the loading rate of EJC onto human transcripts. We found lower EJC presence than previously estimated. The absence of EJC from certain junctions raises questions about the differential regulation of specific splicing events. Further analyses may reveal the factors underlying the loading and detection of EJC-loaded exons, both in human and drosophila, as well as its functional implications in gene expression regulation.

Chapter 9

Annexes

Published online 25 September 2018 *Nucleic Acids Research, 2018, Vol. 46, No. 21* ***11553–11565***
doi: 10.1093/nar/gky858

Monitored eCLIP: high accuracy mapping of RNA-protein interactions

Rémi Hocq†, Janio Paternina †, Quentin Alasseur, Auguste Genovesio* and Hervé Le Hir *

Institut de biologie de l'Ecole normale supérieure (IBENS), Ecole normale supérieure, CNRS UMR8197, INSERM U1024, PSL Research University, 75005 Paris, France

Received May 03, 2018; Revised August 31, 2018; Editorial Decision September 11, 2018; Accepted September 13, 2018

ABSTRACT

CLIP-seq methods provide transcriptome-wide snapshots of RNA-protein interactions in live cells. Reverse transcriptases stopping at cross-linked nucleotides sign for RNA-protein binding sites. Reading through cross-linked positions results in false binding site assignments. In the 'monitored enhanced CLIP' (meCLIP) method, a barcoded biotinylated linker is ligated at the 5′ end of cross-linked RNA fragments to purify RNA prior to the reverse transcription. cDNAs keeping the barcode sequence correspond to reverse transcription read-throughs. Read through occurs in unpredictable proportions, representing up to one fourth of total reads. Filtering out those reads strongly improves reliability and precision in protein binding site assignment.

INTRODUCTION

Post-transcriptional gene regulation is governed by hundreds of RNA binding proteins (RBPs). RBPs form ribonucleoprotein complexes with all kind of RNAs to function as genetic information support, structural scaffold, interaction guide, or enzyme. The repertoire of eukaryotic RBPs comprises over 1500 different RBPs in human (1). In the case of human messenger RNAs (mRNAs), literally covered by proteins, RNA is in direct contact with >800 different RBPs (2,3), which modulate transcript processing and destiny (4). Despite the physiological importance of RBPs evidenced by their implication in diverse pathologies (5), the precise function of most RBPs remains obscure. The development of the cross-linking and immunoprecipitation (CLIP) method represented a pioneering step in the quest of RBP mapping (6). The basic principle of this strategy is the covalent binding of RBPs with their direct RNA targets by ultraviolet (UV) light irradiation. Once cross-linked, RNA digestion separates RNA-protein complexes before immunoprecipitation under stringent washing conditions. Coupled to high throughput sequencing, CLIP offers a transcriptome-wide snapshot of RNA-protein interactions in live cells as covalent links are formed before any disturbing purification step (7). The importance of CLIP methods prompted the community to further improve their efficiency, specificity, and accuracy, as reviewed recently by Lee and Ule (8).

A major caveat of CLIP methods is the poor efficiency of UV-C crosslinking, which is estimated not to exceed a few percent (9). The crosslinking efficiency *per se* can be strongly improved by using photoactivatable ribonucleosides combined with UV-A irradiation (PAR-CLIP) (10). However, incorporation in living cells of nucleoside analogs into RNA is likely to introduce a bias in the RNA sequences that interact with RBPs.

In addition to cross link, cDNA library preparation further decreases CLIP efficiency. After purification and protein digestion, cross-linked peptides remain attached to RNA fragments. This cross-linking mark partially blocks reverse transcriptase (RTase) progression during cDNA synthesis (11). This issue is circumvented by CLIP strategies in different ways. In the HITS-CLIP protocol, cDNA library preparations are based on adaptors ligated at both RNA extremities. Hence, cDNA fragments terminated at the cross-linking site do not harbor the 5′ adaptor and cannot be amplified by PCR. Thus, only cDNA fragments resulting from RTase bypassing the cross-linking site (read-through) are sequenced (Figure 1A). It was then suggested that the center of these read-through reads corresponds on average to the binding site (12), and thus shorter RNA fragments provide higher binding site accuracy. This is limited, however, by the minimal read length (of around 20 nt) required for an unambiguous mapping (13).

The individual-nucleotide resolution CLIP (iCLIP) protocol was conceived to recover truncated cDNA, which may constitute a large fraction of the total cDNA fragments (14). With this approach, a single adaptor is ligated to the 3′-end of RNA fragments before reverse transcription. After circularization and relinearization, cDNAs are amplified by PCR independently of cDNA termination (Figure 1B). The

*To whom correspondence should be addressed. Tel: +33 1 44 32 39 45; Fax: +33 1 44 32 39 45; Email: lehir@ens.fr
Correspondence may also be addressed to Auguste Genovesio. Email: auguste.genovesio@ens.fr
†The authors wish the first two authors to be regarded as joint First Authors.
Present address: Rémi Hocq, IFP Energies nouvelles, Département Biotechnologie, Rueil-Malmaison 92852, France.

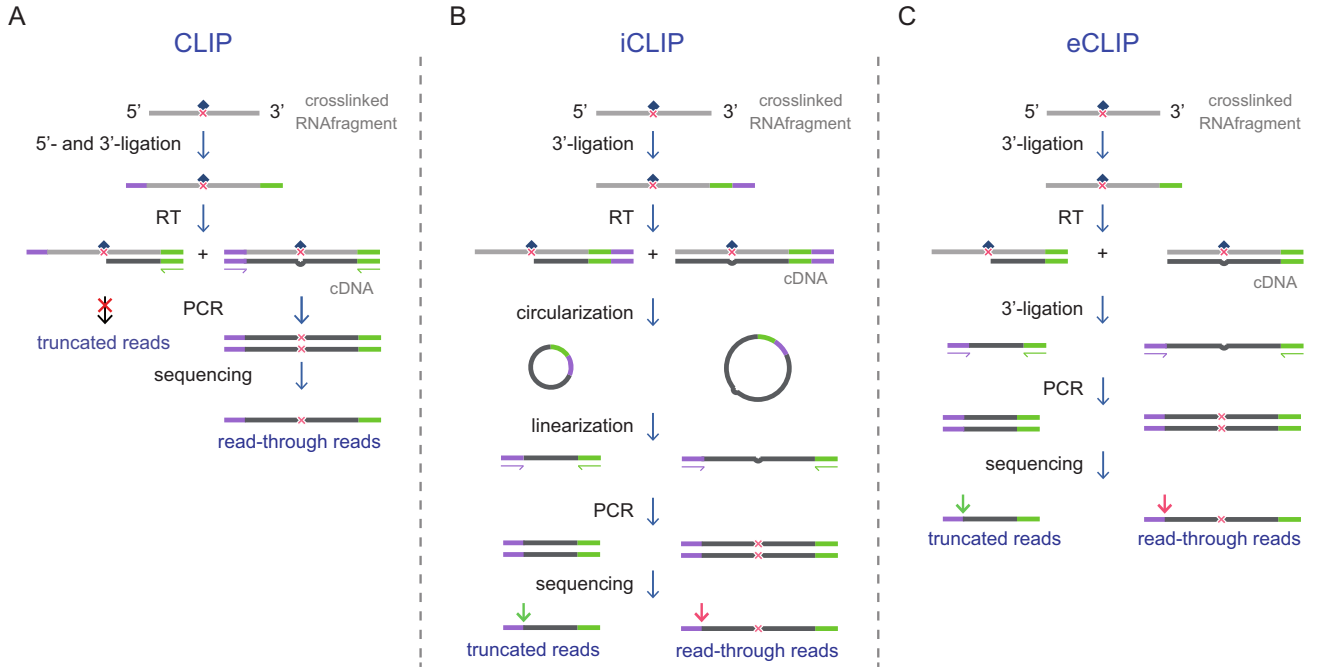

Figure 1. Comparison of CLIP, iCLIP, and eCLIP procedures. Scheme of the CLIP protocol. Immunoprecipitated RNA fragments are coupled to a peptide (blue square) at the crosslinking site (red cross). Reverse-transcription (RT) either stops or reads through the crosslinking site. (**A**) For CLIP, adaptors (purple and green) are ligated at both extremities of the crosslinked RNA fragments. Only read-through cDNAs can be amplified by PCR using primers complementary to adaptors generating read-through reads. (**B**) For iCLIP, a single bipartite adaptor is ligated at the 3′ extremity of the crosslinked RNA fragments. Full-length or truncated cDNAs are circularized and then linearized leading to the presence of adaptors at both extremities. PCR amplifies both truncated and read-through reads. (**C**) For eCLIP, a single adaptor (green) is ligated at the 3′ extremity of the crosslinked RNA fragments. After RT, a second adaptor (purple) is ligated to the 3′ extremity of the cDNAs. PCR amplifies both truncated and read-through reads. The green arrows point towards the position of the crosslinking site. The red arrows point towards the 3′ extremity of the cDNA, upstream the crosslinking site.

first sequenced nucleotide of truncated reads, after 5′ adaptor removal, corresponds to the nucleotide where the reverse transcriptase stopped, one nucleotide downstream of the cross-linking site (14). More recently, the enhanced CLIP (eCLIP) (15), infrared-CLIP (irCLIP (16)) and bromodeoxyuridine CLIP (BrdU-CLIP) (17) methods also suggested improvements of the library construction in order to capture all cDNAs (18). In the case of eCLIP notably, adaptors are ligated first at the 3′-end of RNA and next at the 3′-end of the cDNA, hence bypassing a relatively low-yield circularization step (Figure 1C). In addition, eCLIP includes a parallel analysis of the size-matched input (SM-input) control to identify the most abundant non-specific RNA fragments contributing to background signal (15).

iCLIP- and eCLIP-related methods provide single nucleotide resolution of the cross-linked site as reverse transcriptase tends to stop one nucleotide downstream of the cross-linking site (14) (Figure 1B, C). While in theory the truncation site is independent of the read length, considering various cDNA lengths helps RBP binding site assignment (19). This may result partly from a non-negligible population of read-through reads, whose mapping precision is affected by read length. This population can nonetheless be computationally estimated. Indeed, when passing through the cross-linked nucleotide, reverse transcriptases generate mutations (20). These cross-linking induced mutation sites (CIMS) are valuable both for CLIP-related methods to localize the binding sites and for iCLIP/eCLIP derivatives to estimate the proportion of read-through reads (21). However, CIMS occurrence is variable between RBPs and often remains low among read-through reads, thus preventing a precise binding site mapping (12,21). Furthermore, when long RNA fragments are purified, many sequenced reads are too short to reach the CIMS. In the case of RBPs recognizing specific motifs, adding a motif search can help to map binding sites (12,19,21). However, using motif search is limited by RBPs generally targeting low complexity sequences (22).

For both iCLIP and eCLIP, mapping accuracy depends upon the proportion of truncated reads versus read-through reads, as the latter correspond to spurious cross-linking sites. Consequently, a major hurdle hit by CLIP-related methods is the unpredictable behavior of reverse transcriptases: they either stop or read through the crosslinking site on RNA fragments in a stochastic way. Read-through reads may represent a large percentage of total reads, hence may distort binding site assignment. To discriminate truncated reads from read-through reads, we modified the eCLIP pipeline to establish a 'monitored eCLIP' (meCLIP) protocol. The major modification consisted in adding a biotinylated oligonucleotide ligation to the 5′-end of the RNA fragments to discriminate read-through from truncated cDNA fragments. The ability of the reverse transcriptase to read through the RBPs cross-linking site is monitored to systematically discriminate and filter out read-through reads that generate imprecise peaks. Here, we applied meCLIP to the

human RNA helicases, eIF4A3 (eukaryotic initiation factor 4A3), a core component of the exon junction complex (EJC) (23), and UPF1 (up-frameshift factor 1), an essential factor for the nonsense-mediated RNA decay (NMD) (24).

MATERIALS AND METHODS

Plasmids and molecular cloning

Genome edition of endogenous *eIF4A3* was achieved through expression of Cas9 (*Streptococcus pyogenes*) nickase from pX335 (Addgene) and sgRNAs from sgRNA expression vectors (kind gift from Edouard Bertrand). The sgRNA expression vector displays an optimized sequence for the improved expression of the sgRNA as previously described (25). Briefly, the sgRNA scaffold has been engineered to remove an RNA polymerase III stop motif and to stabilize the hairpin structure recognized by Cas9. sgRNA protospacers sequences were designed using eCRISP (http://www.e-crisp.org/E-CRISP/). Insertion of those sequences in the sgRNA expression vectors was done by Golden Gate assembly into Bbs1 restriction sites. For UPF1 genome edition, sgRNA sequence from pX335 was replaced with the one from the sgRNA expression vector and a second one was added with the Bsa1 restriction site by Gibson assembly (26). Corresponding sgRNAs targeting the C-terminal region of UPF1 were then cloned by Golden Gate assembly in Bbs1 and Bsa1 restriction sites.

eIF4A3 homology regions (800 bp upstream and downstream the stop codon, with modification of PAM sequences to prevent re-cutting) were chemically synthesized (Genewiz) in pUC19. UPF1 homology regions (1000 bp upstream and downstream the stop codon) were amplified by PCR on HeLa genomic DNA. pUC57 vectors comprising sequences coding for the 3× HA affinity tag, an IRES2, the puromycin resistance gene and the SV40 polyadenylation signal were a gift from E. Bertrand's laboratory. For UPF1 edition, the tagging cassette was modified by replacement of the TEV cleavage site by a 3C proteolytic site repositioned after the HA affinity tag and addition of a 3xFLAG tag. These modifications were ordered as a gBlock DNA fragment (IDT). Final repair plasmids were obtained by assembly of the homology regions and the tagging cassettes by Gibson assembly (26).

Cell culture

Human HeLa cells were grown in DMEM supplemented with GlutaMAX, 4.5 g/l glucose, 110 mg/l sodium pyruvate, 10% fetal bovine serum, 100 U/mL penicillin and 100 μg/ml streptomycin (Life Technologies). Cells were passaged every 3–4 days following standard procedure and cultivated in a humidified incubator at 37°C with 5% CO_2. Five million cells at 50% confluency were co-transfected using homemade JetPEI reagent with 0.5 μg pX335, 1.5 μg of each of the sgRNAs expression vectors and 1.5 μg of repair plasmid (eIF4A3) or 0.5 μg of modified pX335 and 4.5 μg of repair plasmid (UPF1). Cells were split in five 15 cm dishes 24h post-transfection and puromycin (InvivoGen) at various concentrations (0, 250, 500, 1000 and 2000 ng/ml) was added 24 h later. Medium was replaced every 3–4 days with fresh antibiotic for 10–15 days. Clones obtained at the highest puromycin concentration were picked in 96-well plates and expanded for an additional 15 days.

Transgene integration and expression

For genomic DNA extraction, cells were lysed in PXL lysis buffer containing proteinase K and RNase A at 37°C. Lysates were centrifuged and the supernatants were then precipitated with 2 volumes of ethanol following phenol–chloroform–alcohol isoamyl (25:24:1) extraction. Transgene integration at the correct locus was verified by PCR with primers annealing upstream the targeted region and in the insertion. Homozygosis was then investigated by PCR with primers annealing upstream and downstream the homology regions. Expression was tested by western blot on the soluble fraction of a cell lysate with antibodies directed against the endogenous proteins and/or the affinity tag itself. Correct integration of eIF4A3 into the native EJC was tested on cell line eIF4A3-HA (clone B) by co-immunoprecipitation and western blot.

Antibodies for immunoprecipitation

Anti-eIF4A3 were previously described (27). HA & Flag tagged proteins were respectively immunoprecipitated with Pierce anti-HA magnetic beads (Life Technologies) and M2 anti-FLAG magnetic beads (Sigma).

Oligonucleotides design and sequences

RNA and DNA linker sequences from the published eCLIP procedure (15) were modified in order to allow sequencing of the library in single-end mode and to be compatible with the P3/P5 PCR primers from Solexa used in standard iCLIP. Random and multiplex barcodes were placed on the second ligation primer. All the sequences are available in Supplementary Text S1. All the oligonucleotides were purchased from IDT and Eurofins Genomics and were ordered desalted, except for the P3/P5 primers that were ordered PAGE purified.

eCLIP and meCLIP library preparation

eCLIP procedure is similar to the one developed by Van Nostrand and colleagues with a few modifications notably towards oligonucleotides sequences (cf. Results section). A few kit-based manipulations were also replaced by conventional biology methods, such as ethanol precipitation. eCLIP and meCLIP step-by-step protocols are available in Supplementary Text S1. Briefly, 20 million cells per sample were crosslinked at 150 mJ/cm^2. Sample underwent partial RNase 1 digestion. The soluble fraction was precleared (wild-type eIF4A3) with unconjugated protein A beads before IP or directly immunoprecipitated (eIF4A3-HA and UPF1-FLAG) on pre-coupled corresponding magnetic beads. Two percent of RNase-treated lysate was kept at 4°C to be used as SM-input negative control. RNP complexes were washed stringently with a buffer containing 1M NaCl and 2M Urea. Cross-linked RNAs were subsequently 5′ and 3′ dephosphorylated, followed by 3′ RNA linker ligation as previously reported (15). In meCLIP experiments,

a 5′ phosphorylation event was added with T4 PNK to allow the subsequent 5′ ligation of the biotinylated linker. Resulting RNPs and SM-input control were purified by SDS-PAGE and transferred onto a nitrocellulose membrane. Size selection was performed by comparison to a radiolabeled control (5% of the beads can be radiolabeled with γ-^{32}P ATP and T4 PNK) and elution of RNAs was achieved by proteinase K treatment, acid phenol–chloroform extraction and ethanol precipitation. SM-input samples were 5′ and 3′ dephosphorylated. 3′ RNA linker was then ligated and the resulting RNAs, as well as the eCLIP samples, were reverse transcribed. cDNAs were purified by Exo1 treatment to remove unused RT primers and alkaline treatment to remove RNAs. A second 3′ ligation step was then performed in conditions optimized by Van Nostrand *et al.* Ligation products were then purified with Agencourt AMPure XP beads modified with a cutoff set at 50-mer (28). Final quantities of the libraries were estimated using qPCR and samples with close Cp were multiplexed prior to final PCR amplification. PCR product were size-selected (175–300 bp) by PAGE and eluted by diffusion. Samples were then precipitated and submitted to single-end sequencing on a NextSeq 500 sequencer (Illumina) in two separate runs (t1 and t2).

Reverse transcription assays

As indicated, SuperScript-IV (Invitrogen) was used on extracted RNAs according to the manufacturer's protocol. Following denaturation, reverse transcription was performed at a high temperature (55°C) to decrease RNA secondary structures. SuperScript-III (Invitrogen) was used in a similar manner, but at a lower temperature (42°C, then 50°C) since this enzyme is less thermostable than SSIV. AffinityScript (Agilent Technologies) was used at 55°C, as previously described (15). TGIRT-III (InGex) was used at 60°C, as previously described (16). Detailed protocols are available in the Supplementary Text S1.

Read pre-processing and mapping

We performed de-multiplexing of raw reads using a custom script that identifies sample 5′ end barcodes (four nucleotides within a 9 bp randomer). We applied PCR duplicate removal on the de-multiplexed data, and once again after merging the reads from the same sample originating from different lanes; reads with the exact same sequence, including the 9 bp randomer were considered as PCR duplicates. After barcode trimming, we used cutadapt (version 1.10) to trim the 13 bp 5′ end linker (CAGTCCGACGATC) of read-through reads and simultaneously separate them from truncated (untrimmed) reads. Finally, we trimmed the 3′-Illumina adaptor and poor-quality bases with trimmomatic (v.0.36), discarding reads that were <20 bp long after trimming (options ILLUMINACLIP:/path/to/Trimmomatic-0.36/adapters/TruSeq2-PE.fa:2:30:1 SLIDINGWINDOW:5:25 LEADING:25 TRAILING:25 MINLEN:20). To sort the reads into the categories of short and long cDNA fragments, we used cutadapt after PCR duplicate removal with the following options:

```
cutadapt -a AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG –m 20 –untrimmed-output = my_sample_long_fragments.fastq my_sample.fastq > my_sample_short_fragments.fastq
```

Untrimmed reads correspond to long fragments that were not fully sequenced and thus lack the Illumina 3′ end adapter. All sorted reads were mapped separately against the reference genome.

For genome visualization, datasets were mapped to the human genome (hg38, Ensembl 85, with processed transcripts and pseudo genes masked), using STAR (version 2.5.1b) with the following parameters:

```
STAR –readFilesIn raw_reads.fastq.bz2\
–outFileNamePrefix/path/to/output/mapped_reads.\
–readFilesCommand bunzip2-c\
–outReadsUnmapped Fastx\
–genomeDir/path/to/genome/index/–sjdbOverhang 100–sjdbGTFfile/path/to/annotation/hg38.gtf\
–outFilterType BySJout –alignSJoverhangMin 8–alignSJDBoverhangMin 1\
–outFilterMatchNminOverLread 0.4–outFilterScore MinOverLread 0.4–outFilterMultimap Nmax 20\
–outFilterMismatchNmax 999–outFilterMismatchNoverLmax 0.06–alignIntronMin 20\
–alignIntronMax 1000000–outSAMattributes All\
–outSAMtype BAM SortedByCoordinate\
–outWigStrand Stranded –quantMode GeneCounts
```

For the meta-exon profiles, we downloaded spliced transcript sequences from Ensembl (hg38, Ensemble 85), and mapped the reads using bowtie2 (version 2.3.2) with its default parameters. We used spliced transcriptome sequences as reference rather than the genome sequence because we systematically obtained a depletion of reads in the 6nt region upstream of the exon junction. In both cases, we mapped reads to one representative transcript per gene, selecting the isoform with the maximum number of exons, using the longest exonic size as a tiebreaker. To compute the number of uniquely mapped reads, we used the number of uniquely mapped reads reported in the STAR final log file. We computed the read-through percentage of each meCLIP library as the number of uniquely mapped read-through reads over the sum of uniquely mapped truncated and read-through reads.

Peak detection and intersection

We used the CTK suite (Shah *et al.*) to detect cross-linking induced truncation sites (CITS) (https://zhanglab.c2b2.columbia.edu/index.php/CTK_Documentation) using the following commands:

```
perl /path/to/ctk/parseAlignment.pl -v –map-qual 255 \
–min-len 18 –mutation-file mutations.txt - \
parsedReads.bed
perl /path/to/ctk/getMutationType.pl
-t del mutations.txt \
parsedReads.deletions.bed
perl /path/to/ctk/CITS.pl -big -gap 10 -p 0.001 \
parsedReads.bed \
```

parsedReads.deletions.bed \
cits.output.bed

The CITS detection tool finds significant truncation sites when the number of truncations per position are compared to a shuffled read-start distribution. It is therefore suitable for single-nucleotide resolution binding site assignment in iCLIP and eCLIP experiments. Next, we computed 2-fold SM-input enrichment of the CITS reported by CTK and of the same set of CITS with shuffled genomic coordinates. For downstream analyses, we selected CITS with a 2-fold enrichment higher than the 95th percentile of the shuffled distribution (which corresponds to a P-value <0.05). Multitest correction was not used due to a major loss of power (high number of false negatives).

Prior to CITS intersection, we increased the CITS region with bedtools slop (version 2.27.1), using the option -b 5. We carried out all read and truncated read CITS intersection with BedTools' intersect (version 2.27.1) with options –c –s –a truncated.cits.bed –b all.reads.cits.bed to obtain the number of common peaks; to retrieve the peaks only found in either dataset, options –v and –s were used. We used matplotlib-venn (version 0.11.5) to plot the Venn diagrams. The percentage of common CITS corresponds to the Jaccard index multiplied by 100; the percentage of CITS found only on either all reads or truncated reads was computed by dividing the respective number of peaks by the total number of peaks detected on both datasets.

regplot function from the Seaborn python library (version 0.8.0) was used to plot both the percentage of read-through reads and the number of uniquely mapped reads against the fraction of peaks detected in the all read datasets.

Peak correlation scatterplots

To assess the correlation between replicate peaks, we followed the ENCODE procedure to find significantly SM-input enriched peaks (Ref. Yeo). We calculated SM-input fold-enrichment of all CITS detected by CTK and applied False Discovery Rate as multiple test correction. Peaks with fold-enrichment higher or equal to 8 and P-value under 10^{-5} were considered significantly enriched.

Next, we obtained the CITS in common between replicates by intersecting all detected CITS (independently of their SM-input enrichment), using the same parameters as the intersection described above. We plotted the fold-enrichment value of each replicate for the common peaks, coloring the fraction of peaks that were significantly enriched in the first replicate. We computed the squared Pearson's correlation coefficient (R^2) using the SciPy python library (version 0.19.1), both on fold-enrichment values of all CITS and of significantly enriched CITS of the first replicate.

Distribution of 5′ ends relative to the exon junction (meta-exon plot)

We used BedTools intersect (version 2.27.1) to intersect uniquely mapped reads to Ensembl85 exon annotations of the hg38 assembly of the human genome, which was generated using the header of the transcript sequences downloaded from Ensembl; the genomic coordinates of exons were converted into transcript coordinates to be consistent with the mapping output from bowtie2. We only considered reads mapped to protein coding genes, mapped to exons longer than 30 bp, and whose 5′ end mapped inside the boundaries of the exon; the distance of the 5′ end of each read was plotted to either the start or the end of the exon, correcting the exon distribution and library size by dividing the counts at each relative position by the number of exons covered at that position and the total number of mapped reads. For genome visualization, BAM files (STAR aligner output) were converted to bedGraph files using BedTools genomecov function. To find 'canonical exons', we automated the retrieval of individual exon coordinates by selecting exons with a high proportion of 5′ ends inside the approximate canonical binding region (between 29 and 19 nucleotides upstream of the exon junction); similarly, we identified exons with no read-through signal by selecting exons which intersect with truncated reads but do not intersect with read-through reads.

RESULTS

Immunoprecipitation strategy for CLIP

CLIP efficiency suffers from the caveats inherent to IP such as epitope accessibility, affinity and antibody specificity. This is particularly critical when using large-scale proteomic or transcriptomic approaches to characterize protein complexes. Indeed, the depth of such strategies determines signal discrimination form noise. Unfortunately, suitable commercial antibodies for IP are not always available, especially for newly discovered RBPs. Furthermore, dedicated antibody production is long and uncertain. As an alternative option, exogenous proteins fused to well-characterized affinity tags can be used. However, expression conditions of recombinant proteins (strong synthetic promoters, optimized codons, high gene copy number/cell ratio) may generate artefactual interactions that differ from the endogenous cellular context. Moreover, the competition between endogenous and recombinant RBPs may provoke biases in the outcome.

Recently, Van Nostrand and colleagues addressed some of the aforementioned immunoprecipitation issues by using a version of eCLIP (TAG-eCLIP) (30), in which CRISPR-Cas9 mediated gene editing was used to generate endogenous RBPs fused with affinity tags. However, when gene modifications are heterozygous, only a portion of the protein of interest is concerned by the immunoprecipitation of the affinity tag, thus impacting IP yield and RNA-protein complexes recovery. Here, we used CRISPR/Cas9 (31) (Supplementary Figure S1) to knock-in an affinity tag and a selection cassette to select positive insertions (Figure 2A, Supplementary Figures S2 and S3). To reduce the number of clones resulting from random plasmid integration and increase the yield of homozygous insertions, a selection marker was inserted at the C-terminal region of the RBP loci as an independent open reading frame driven by an internal ribosomal entry sequence (IRES) to avoid its fusion to the tagged protein. We successfully obtained homozygous HeLa cell lines making eIF4A3 and UPF1 proteins

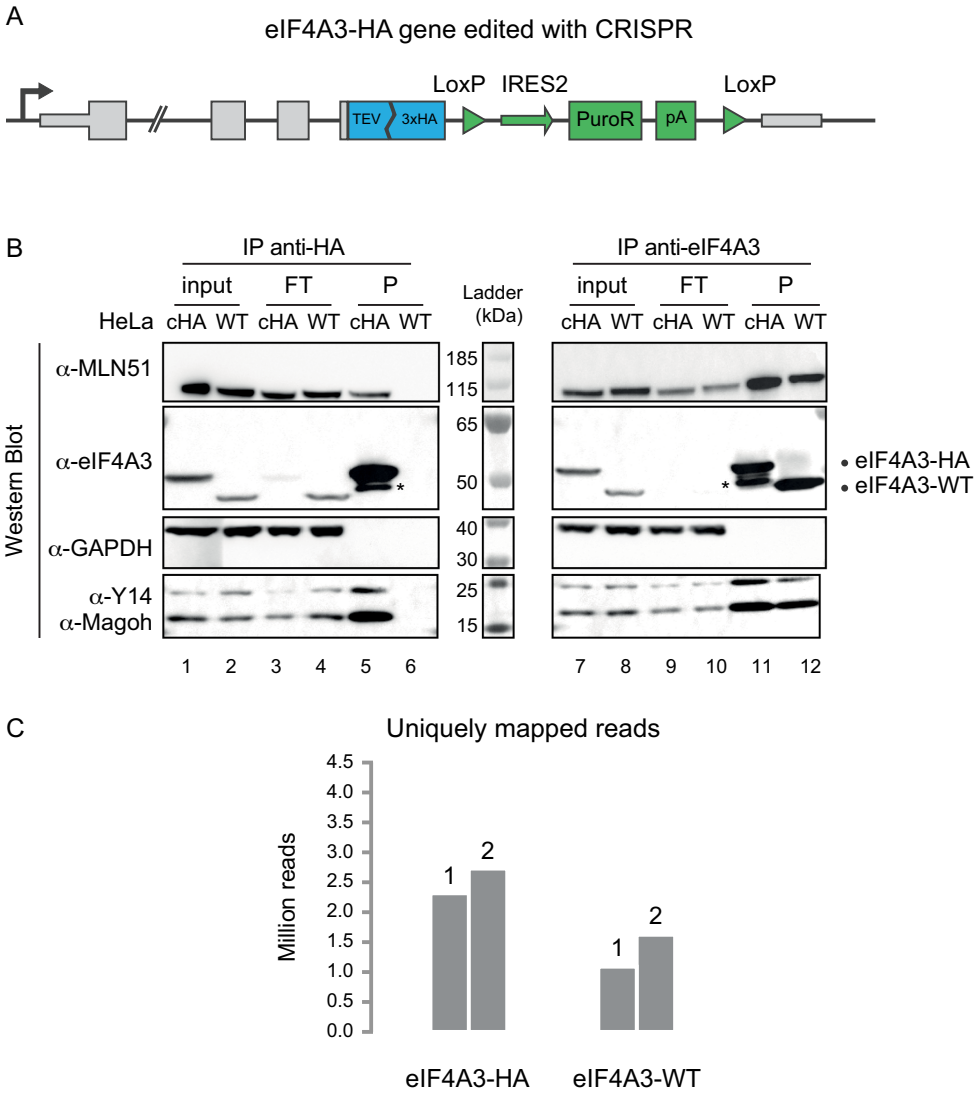

Figure 2. CRISPR/Cas9 editing of eIF4A3. (**A**) Schematic representation of the edited eIF4A3 gene. C-terminal insertion harbors a TEV proteolytic cleavage site and a 3xHA affinity tag, fused to a Internal Ribosomal Entry Site (IRES2)-controlled puromycin (PuroR) selection cassette encompassed by LoxP recombination sites. pA: poly A signal. (**B**) Lysates from wild type (WT) or eIF4A3-edited (cHA) Hela cells were immunoprecipitated with anti-HA or anti-eIF4A3 antibodies and probed for EJC core subunits by Western Blot. FT: flow-through; P: precipitate. The star indicates a C-terminal truncated form of eIF4A3-HA. **C.** Number of uniquely mapped reads from eIF4A3 eCLIP libraries obtained either with endogenous (eIF4A3-WT) or with anti-HA (eIF4A3-HA).

fused to either 3xHA or 3xFLAG affinity tags. Sequencing of the corresponding genomic regions showed that both gene alleles were correctly edited (Supplementary Figures S2 and S3). eIF4A3 expression levels were compared in both WT and edited HeLa cells using anti-eIF4A3 or anti-HA antibodies. Single bands showed that both *eIF4A3* alleles had been successfully modified and the HA tag was confirmed not to affect protein expression levels (Figure 2B, lanes 1, 2, 7, 8). Moreover, immunoprecipitation with anti-HA or anti-eIF4A3 antibodies confirmed that both forms of the protein co-precipitated as efficiently their core EJC partners MAGOH, Y14 and MLN51, demonstrating that editing neither altered eIF4A3 expression, nor its incorporation into EJCs.

We next employed the eCLIP pipeline (15) to compare eCLIP efficiency performed with anti-eIF4A3 and anti-HA antibodies. After sequencing, read pre-processing and mapping against the human genome, we found that a higher number of uniquely mapped reads was obtained using the anti-HA antibody for the immunoprecipitation step (Figure 2C). Thus, the high affinity anti-HA antibody against CRISPR-tagged eIF4A3 improves eCLIP library preparation efficiency.

Sorting out truncated cDNA reads using 'monitored eCLIP' or meCLIP.

A bottleneck of current CLIP procedures resides in their inability to determine the proportion and variability of

reverse transcriptase that read through cross-linked nucleotides, and the consequences on the accuracy of RBP binding site assignment. To alleviate this hurdle, we modified the standard eCLIP pipeline, to establish 'monitored eCLIP' or meCLIP. The major modification of meCLIP compared to eCLIP consists in ligating an oligonucleotide containing a barcode at the 5′ end of RNA fragments before RT (Figure 3). The 5′ linker is reverse transcribed if, and only if, the RT manages to pass the RNA-peptide cross-linking site. It is then possible to quantify the ratio between the numbers of reads that harbor the 5′ linker and those that do not. However, the ligation yield significantly varies across experiments (10,32) and unligated RNAs may significantly bias such an approach. To circumvent this obstacle, the 5′ linker was biotinylated. Purification of biotinylated RNA fragments eliminates unligated RNAs before reverse transcription. After sequencing, reverse transcription termination events are then easily monitored by detecting the biotinylated linker sequence at the beginning of the read (Figure 3). The steps of eCLIP and meCLIP methods shown in Figure 3 are described in detail in the Methods section. In addition to the 5′ linker ligation, we replaced all adaptors for single-end sequencing compatibility. Since both the cross-linking site and the biotinylated linker are at the 5′-end of the read, pair-end sequencing is not necessary. Additionally, random and multiplexing barcodes were placed on the 3′ DNA linker, avoiding the use of costly RNA multiplexing linkers. In summary, with the biotinylated 5′ RNA linker enables it is possible to distinguish truncated reads from read-through reads.

meCLIP reveals a highly variable proportion of read-through reads

In order to estimate the impact of the extra steps added to the eCLIP protocol onto library preparation efficiency, we first performed four eCLIP and meCLIP experiments in parallel using anti-HA to target eIF4A3-HA. Quantitation of cDNA libraries by RT-qPCR revealed that meCLIP preparation is on average only 3.5 times less efficient than eCLIP. Moreover, we verified that fragments that are not ligated to the biotinylated primer are not retained on streptavidin beads, which indicated the high specificity of biotinylated fragment purification. Then, we performed meCLIP using anti-eIF4A3 antibodies, anti-HA (to target eIF4A3-HA) or anti-FLAG (to target UPF1-FLAG) using *SuperScript IV* RTase. We observed that the quantity of read-through reads reached up to one-fourth of the total number of reads (Figure 4A). In addition, there were great variations in read-through proportions between replicates or between targeted RBPs. It is important to note that read-through read percentage reflects the ability of a reverse transcriptase to bypass a cross-linked nucleotide. As this ability may differ from one RTase to another, we repeated eIF4A3-HA meCLIP with three additional RTases: *AffinityScript*, *SuperScript III* and *TGIRT III*. These enzymes have different biological origins and have also been employed for various CLIP experiments (14–16). Each reaction was carried out at the optimal conditions of each enzyme. In addition, to test the variability of this feature and to simulate laboratory-to-laboratory variations, meCLIP experiments

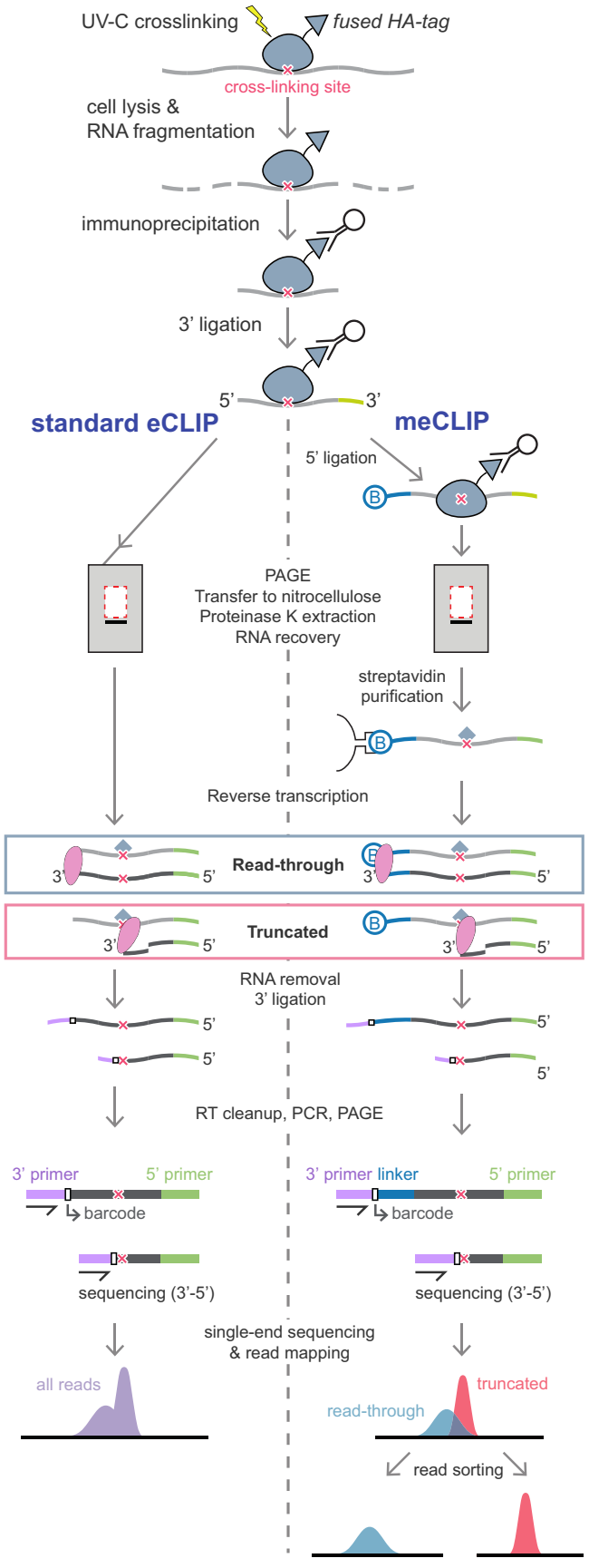

Figure 3. Detailed comparison of the eCLIP and meCLIP protocols. Presentation of the different steps involved in eCLIP and meCLIP procedures.

were done in duplicate by two different experimenters and in independent sequencing runs. For the various eIF4A3-HA meCLIP libraries, the percentage of read-through reads was highly variable (from 2 to 24%) and depended on the RTase used (Figure 4B). The lowest proportion of read-through reads was observed with *AffinityScript* while the three other enzymes generated more variable and greater proportion of read-through reads (at least 8–25%). qPCR quantification of cDNA obtained after RT showed variable efficiency measures among different enzymes (Supplementary Figure S4A). However, different cDNA yield did not correlate directly with PCR duplicate rate in sequenced libraries (Supplementary Figure 4B, C). These results illustrate how the meCLIP protocol can sort out the highly variable and significant amount of read-through reads from the final dataset independently of experimental conditions.

meCLIP removes noise from binding site detection

Detection of significant peaks among mapped reads is a critical step in CLIP-seq downstream analysis to identify potential RBP binding sites. To assess the impact of read-through reads on peak detection, we used the cross-link induced truncation site (CITS) detection software from the CTK suite (29) on meCLIP datasets corresponding either to all reads (equivalent to an eCLIP dataset), or to truncated reads only. As detailed in the Methods section, we selected CITS with a significant SM-input enrichment ($P < 0.05$). The comparison is illustrated first by a few examples that show how peaks are distributed on some annotated exons (Figure 5). On exon 47 of the *LAMA1* gene, two of the four CITS (underlined by a black line) are detected on the all read signal. When applying CITS detection on truncated reads only, these truncation sites are no longer detected, and the right-most CITS is shifted upstream toward a stronger truncation signal. (Figure 5A). The binding site detected in all reads on exon 7 of *FBLX6* corresponds mainly to read-through read signal and thus is not likely to be a truncation site (Figure 5B). In contrast, the binding site on exon 11 of *INPPL1* corresponds exclusively to truncated read signal, and most likely corresponds to a cross-linked RNA region (Figure 5C). These examples illustrate that, within a given eCLIP experiment, the proportion of read-through versus truncated reads is highly variable and unpredictable from one transcriptome region to another (Figure 5A–C). CITS detected in the whole eCLIP dataset but absent in the truncated reads dataset clearly constitute incorrect cross-linking sites.

To determine the impact of read-through reads on CITS detection at the genome scale, we intersected the sets of CITS detected on all reads and on truncated reads separately. In the case of *SuperScript III*, replicate 1, most peaks are common (~82%), a small proportion (0.34%) are detected only in the truncated read dataset, and 17% of the peaks are detected only in all reads (Figure 5D). Since the only difference between all read and truncated read data sets is the presence of read-through reads, these CITS were considered as originating from the read-through signal. By comparing eIF4A3 datasets obtained with different RTases, as well as the UPF1 replicates, we found a direct relationship between the percentage of 'read-through CITS' and the proportion of read-through reads in each library (Figure 5E), but not with the total number of mapped reads (Figure 5F); we did not observe this relationship when comparing with the CITS detected exclusively on truncated reads (Supplementary Figure S5). Taken together, these results show that read-through reads generate imprecise binding sites that can be sorted out with the meCLIP procedure.

Next, we followed the ENCODE eCLIP pipeline (15) to assess the reproducibility of meCLIP peaks detected on all reads and on truncated reads (Supplementary Figure S6). Using this approach, comparison of meCLIP datasets corresponding to the unrelated proteins eIF4A3 and UPF1 showed very poor correlation coefficients. In contrast, comparison of meCLIP replicates showed higher correlation coefficients, varying between 0.19 and 0.58. Overall, removing read-through reads had little to no effect on replicate reproducibility, as shown when comparing truncated reads to all reads.

meCLIP improves precision of cross-linking site positioning

We next investigated the influence of read-through reads on the localization of the binding sites of a given protein. We used the data of eIF4A3 which has been shown to bind to a precise position upstream of the exon junction (23) and not the data of UPF1 that has been shown to be widely dispersed over mRNA 3′UTR regions (33,34). We first compared the transcriptome-wide distribution of 5′ ends using the addition of truncated and read-through datasets and positioned them relative to the exon junction. A sharp peak was observed centered on the 27th nucleotide upstream the exon junction (Figure 6A). Most of the reads around this position belong to the truncated read category. In contrast, the 5′ ends of read-through reads are distributed upstream of the 27th nucleotide peak, as expected for reads bypassing the cross-linking site. We next examined the 5′ end position of truncated and read-through reads on individual exons. In agreement with the transcriptome-wide distribution, read-through signals often appear upstream of a cluster of truncated reads 5′ ends (Supplementary Figure S7A and B). Upstream read-through signal varies in positions and intensities from one exon junction to another. They are sometimes absent despite a read-through percentage of over 10% in the *SuperScript III*, replicate 1 library (Supplementary Figure S7C). Additionally, we found examples of CITS far away from the canonical EJC deposition site that may correspond to non-canonical EJC binding sites (27,33) (Supplementary Figure S6D). These examples illustrate how the meCLIP

Following UV crosslinking, RNase treatment, and RBP purification, an RNA adaptor (green) is ligated at the 3′ end. For meCLIP, a biotinylated RNA linker (blue) is incorporated at the 5′ end. RNAs are fractionated by electrophoresis and eluted from gels. For meCLIP, biotinylated RNAs are purified on Streptavidin beads using stringent conditions. Reverse transcription (RT) is then performed, which leads to two distinct cDNA populations. One of them bears the 5′ linker if the reverse transcriptase reads through the crosslinked peptide (read-through cDNAs). The other one lacks the 5′ linker due to a stop of RT at the crosslinked peptide. A second adaptor (purple) is ligated at the 3′ end of the cDNAs which are next amplified by PCR and submitted to high-throughput sequencing. For meCLIP, two populations of reads are easily sorted out based on the presence or absence of the biotinylated linker sequence.

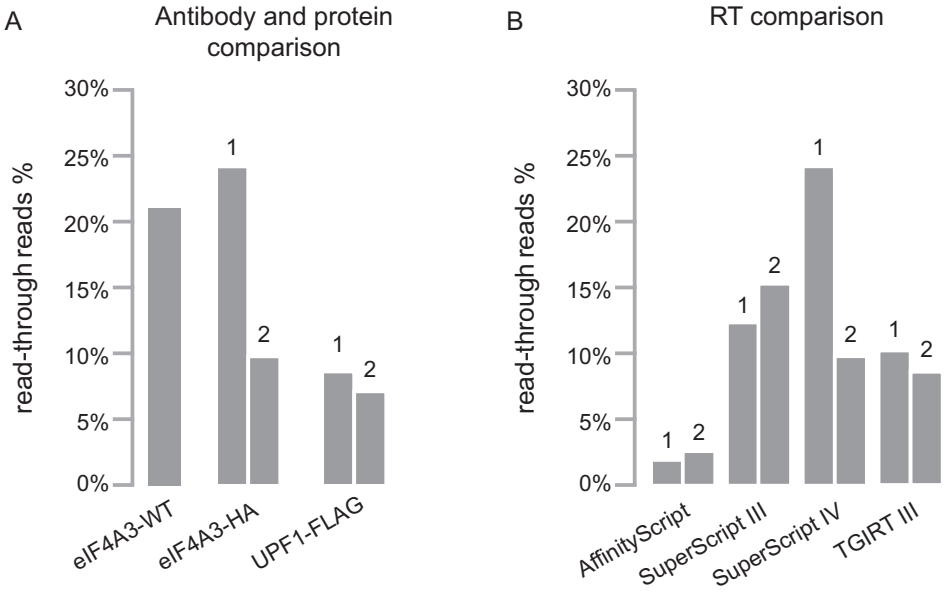

Figure 4. Read-through reads percentage for meCLIP depends on experimental conditions. Libraries generated using (**A**) different cross-linked proteins and (**B**) different reverse transcriptases (RTase) and HA-tagged eIF4A3.

protocol reduces the noise of RBP signals and helps improving the precision of binding site localization.

Read length was recently debated (13,19) as having an impact on the precision of 5′ end of reads obtained from iCLIP, and iCLIP-derived experiments. Hauer *et al.* pointed out that using the center of short fragments instead of the 5′ end of iCLIP reads increased the precision of RBP mapping. To verify this feature, we sorted truncated and read-through reads based on the presence (short fragments) or absence (long fragments) of the Illumina 3′-adapter sequence. The same sharp peak centered on the 27th nucleotide upstream the exon junction was observed for both short and long truncated reads (Figure 6B). However, the signal upstream of the main peak was weaker for short truncated reads compared to long truncated reads. Notably, the signal shifts upstream for both short and long read-through reads. Thus, the use of short truncated reads further increases the precision of binding site assignment.

Another strategy for binding site assignment consists in detecting cross-linking induced mutation sites (CIMS), which stem from errors during reverse transcription. This strategy is used in the case of HITS-CLIP and PAR-CLIP data analysis, where only read-through reads are exploited. We used the short read-through fraction of reads, which is more likely to harbor most CIMS, to quantify the number of mutations in our libraries. We found that deletions occur in $<0.5\%$ of short fragments (with the exception of one library that reaches over 1.2%); for libraries obtained with AffinityScript, they are not detectable at all; insertions are consistently negligible (Supplementary Figure S8). Considering that short fragments are about 50% of all uniquely mapped reads, a sensitive CIMS detection would require high sequencing depth, which can be especially costly in the case of multiplexed experiments. In contrast, meCLIP, as well as other iCLIP-derived methods, identifies the crosslinking-induced truncation sites by using directly the 5′-ends of the truncated reads, which on average make up approximately 90% of uniquely mapped reads.

The sharp enrichment of truncated reads observed in the meta-exon plot for the eIF4A3 datasets (Figure 6A) prompted us to compare the distribution of reads relative to the exon junction of four successive CLIP protocols: the original CLIP or HITS-CLIP (27), iCLIP (19), eCLIP and meCLIP. By applying exactly the same data analysis pipeline to all datasets, we observed both a sharpening and an increase of the meCLIP signal at the 27th nucleotide upstream of the exon junction relative to the other protocols (Figure 6C), indicating that a higher proportion of meCLIP truncated reads map to this precise position. Altogether, we demonstrated that the coupling of a nucleotide-resolution CLIP method to the identification of read-through reads and to CRISPR/Cas9-mediated genome editing for affinity tagging, improves the accuracy of the cross-linking site localization.

DISCUSSION

Although extremely informative on transcriptome-wide RBP binding sites, performing a CLIP experiment is a difficult task. Following UV irradiation, a small amount of cross-linked RNA fragments must be isolated by immunoprecipitation from a tremendous excess (possibly $>10^6$ fold) of undesired RNA fragments. Despite this challenge, the analysis of CLIP reads generated by deep sequencing are expected to narrow down the assignment of the RBP binding sites to one nucleotide. In this study, we show that ligation of a biotinylated barcode linker to the 5′ end of RNA fragments markedly improves the CLIP cDNA library preparation to identify and discard misassigned binding sites. Furthermore, genome edition brings a convenient strategy to

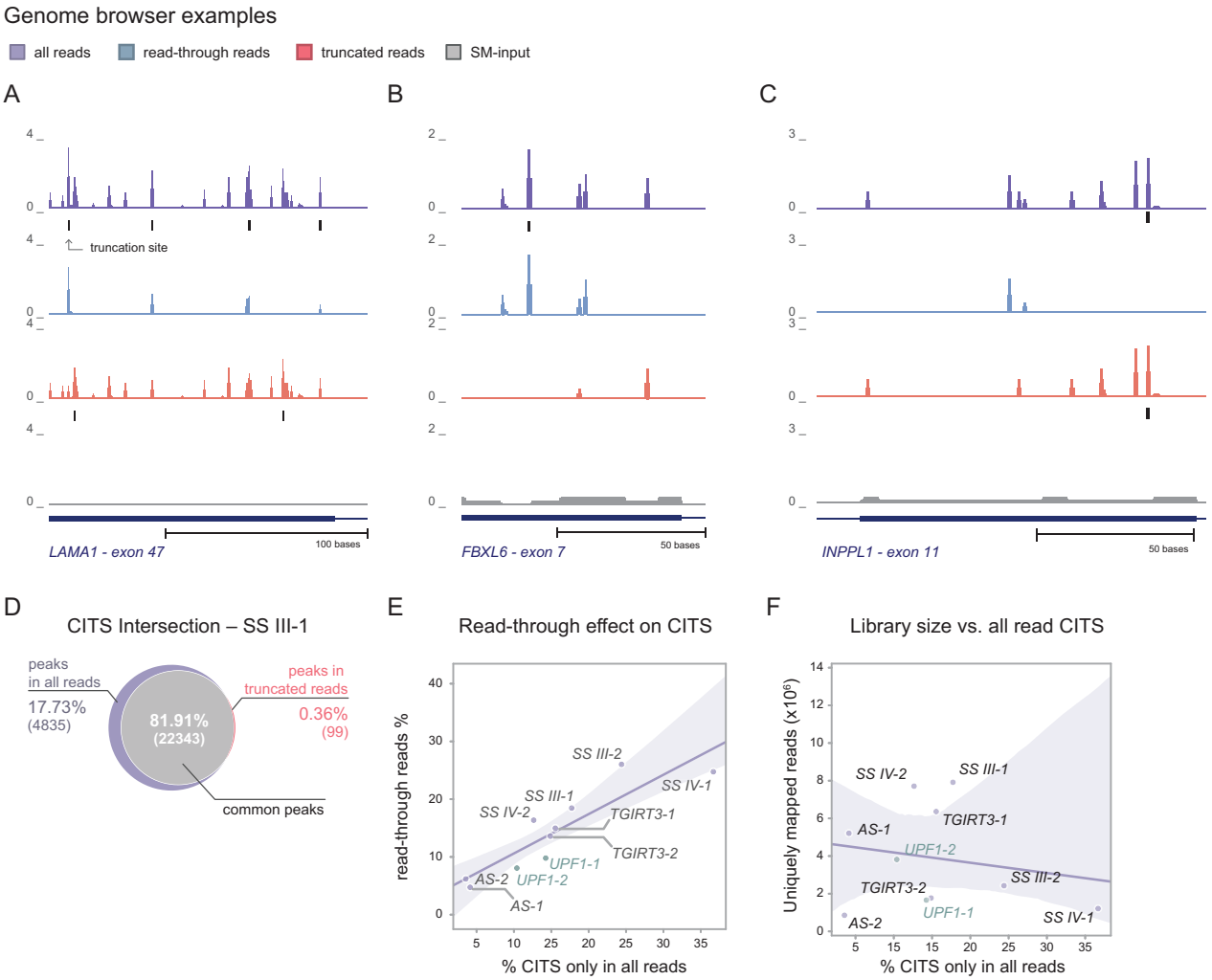

Figure 5. CITS detection is biased by read-through reads. (**A–C**) meCLIP reads mapped on three examples of exons. Each black underline corresponds to a CITS detected with CTK. Read coverage is in Reads Per Million (RPM). (**D**) Venn diagram representing the intersection of peaks detected in the unfiltered (all reads) and filtered (truncated reads) data sets from *SuperScript-IV* replicate 1. (**E**) Read-through reads percentages are plotted against the percentage of truncation sites (CITS) detected exclusively in all reads (purple in D). (**F**) Number of uniquely mapped reads versus the percentage of CITS detected exclusively in all reads; the shade around the line indicates the confidence interval (95%) of the linear regression. *AS*: *AffinityScript*, *SSIII*: *SuperScript III*, *SSIV*: *SuperScript IV*.

bypass the need of specific antibodies for immunoprecipitation.

The tagging of endogenous proteins by genome editing has recently proved to be an alternative to antibodies for CLIP analysis (15). Consistently with the work of van Nostrand *et al.*, we successfully obtained cell lines expressing tagged RBPs. In the case of eIF4A3, this incorporated modification did not alter its expression nor its capacity to correctly assemble the EJC. We optimized the genomic modification strategy to obtain a high yield of homozygous clones (Supplementary Figures S1–S3). Homozygosis prevents competition between tagged and untagged versions of the protein within the same cell, while also ensuring that a higher quantity of edited protein is available. The comparison of the expression of native or HA-tagged eIF4A3 in HeLa cells, and the comparison of the CLIP data in the two cell lines (Figure 2) confirmed that the CRISPR-Cas9 mediated fusion of affinity tags to human RBPs constitutes a compelling alternative to specific antibodies (15). Furthermore, this strategy offers the opportunity to target proteins for which no antibodies suitable for CLIP are available, broadening perspectives for the vast number of poorly characterized RBPs. Simultaneous CRISPR–Cas9 protein tagging of several proteins with different tags offers the possibility of co-purifying proteins of interest—a strategy certainly appropriate to shed light on the dynamics of RBPs, which often function in several different RNP complexes.

After obtaining cell lines expressing tagged RBPs, the eCLIP protocol was engineered to distinguish CLIP reads resulting from either reverse transcriptase termination or read-through at the cross-linking site. Until now, the frequency of RT stalling at the peptide–RNA cross-linking site has never been strictly assessed and indirect estimations suggested that it could be variable (12,21). Our strategy al-

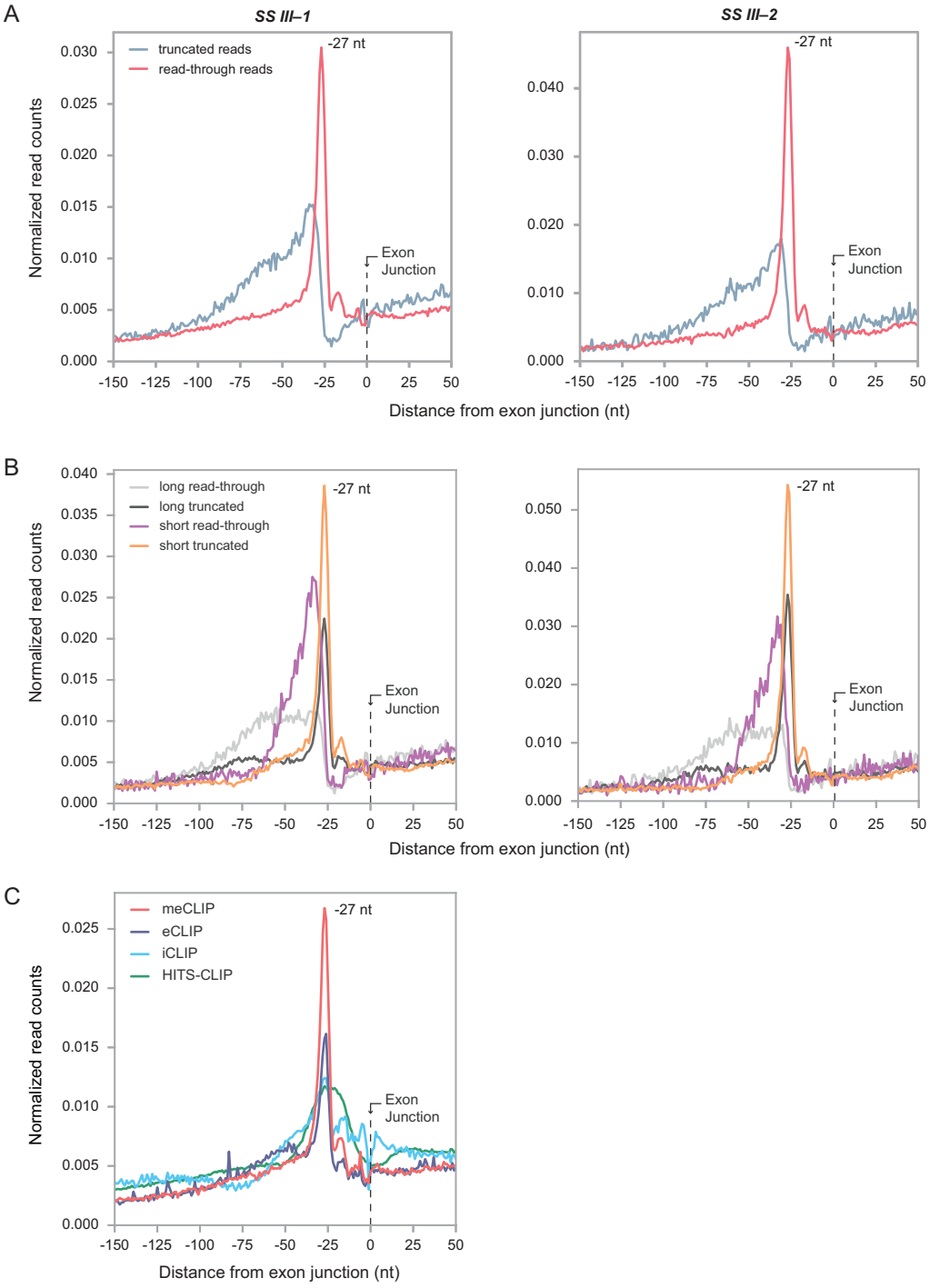

Figure 6. Increased accuracy of cross-linking site positioning. Positioning of 5′ ends of meCLIP reads relative to the exon junction. (**A**) Distribution of eIF4A3-HA meCLIP replicates: truncated reads (red) and read-through reads (blue). (**B**) Distribution of eIF4A3-HA meCLIP replicates: short truncated reads (orange), short read-through reads (purple), long truncated reads (dark grey), and long read-through reads (light gray). (**C**) Distributions of eIF4A3 reads obtained with the meCLIP, eCLIP, iCLIP and HITS-CLIP procedures. meCLIP signal corresponds to truncated reads, and is normalized using the number of uniquely mapped truncated reads.

lowed to precisely measure for the first time the proportion of RT from the truncated ones. The meCLIP data set analysis shows that the percentage of read-through reads reaches up to 24% of all reads and is also highly variable and unpredictable (Figure 5 and Supplementary Figure S6). Among the 10 meCLIP experiments performed for this study, the percentage of read-through reads varied from 2 to 24%. Variations from 7 to 24% exist for identical meCLIP libraries prepared by two different experimenters, or by the same experimenter at different times, which clearly demonstrates that read-through frequency cannot be predicted even for a single RBP. Importantly, we show that meCLIP distinguishes read-through reads from truncated reads despite variable proportions of read-through, different RT experimental conditions, and different efficiency levels among reverse transcriptases. For two different RBPs, meCLIP datasets contained a relatively high percentage of read-through reads. This underlines the importance of using meCLIP to eliminate the unpredictable proportion of spurious reads.

Moreover, we found that mapped read-through reads generate a signal that leads to incorrect RBP binding site assignment, as seen by the direct relationship between read-through CITS and read-through read percentage (Figure 5E). Thus, discarding the more imprecise read-through signal is necessary for precise single-nucleotide binding site assignment. The benefit of read-through reads elimination for RBP mapping was clearly visible in the case of eIF4A3, which is known to be deposited upstream of mRNA splice junctions. Visualization of individual exons shows that meCLIP signal corresponding to read-through reads is in most cases located upstream of signal of truncated reads. This trend is even more visible on the transcriptome-wide distribution of the relative distance of the 5′-end of meCLIP reads to the exon junction. Indeed, read-through read signal is enriched upstream of the canonical position of eIF4A3, which is centered 27 nucleotides upstream of spliced junctions. Although this meta-exon analysis is highly reproducible between replicates (Figure 6), we observed similar correlation coefficients between eCLIP and meCLIP when comparing replicates at the peak level (Supplementary Figure S6). However, the moderate correlation of significantly enriched peaks shows that more efforts are necessary to improve single nucleotide CLIP reproducibility.

Importantly, the proportion of eIF4A3 read-through signal greatly varies from one exon to another revealing that even within the same experiment the frequency of RT stalling is highly variable and unpredictable. As previously noticed in the case of eIF4A3 mapping (13,19), the consideration of short reads further increases the precision of binding site assignment. This is probably due to the fact that the longer the RNA fragments are prior to RT, constraints such as non-specific cross-links or secondary RNA structures are more likely to impair RT and cDNA truncation at the *bona fide* crosslinking sites. However, despite a slight enrichment upstream eIF4A3 canonical position, long truncated reads accumulate in a similar pattern as their short counterparts. As long reads represent a large fraction of total reads, discarding them risks decreasing sequencing depth and negatively affecting downstream analyses. A better option would be to select experimental conditions that maximize the proportion of short fragments (e.g. optimized ribonuclease treatment, size-selection). Finally, we demonstrated the improvement brought by the meCLIP strategy through the comparison of four successive CLIP protocols targeting eIF4A3: the original CLIP or HITS-CLIP (27), iCLIP (19), eCLIP and meCLIP (Figure 6c), under similar experimental conditions (such as cell lines and antibodies) and using the same analysis pipeline. Future meCLIP analyses of eIF4A3 will certainly help to understand the mechanisms that regulate the EJC deposition. More generally, future analyses of meCLIP shall tell whether this method adds a quantitative and localized dimension to the study of RBP dynamics and function.

In summary, our meCLIP method significantly improves the accuracy of RBP binding site mapping by unambiguously filtering out CLIP reads that do pinpoint RBP crosslinking sites and consequently translate into biased peaks. Furthermore, the combination of genome editing to fuse efficient affinity tags to RBP of interest with meCLIP paves the way to elucidate poorly characterized RBP functions and their role in post-transcriptional gene regulation.

DATA AVAILABILITY

meCLIP data sets have been deposited to the Sequence Read Archive (accession number SRP154888).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank lab members and J. Ule for fruitful discussions. We are grateful to E. Bertrand for plasmids and to L. Bastianelli for bioinformatics contribution.
Author contributions: R.H. and H.L.H. designed the genome editing and CLIP experiments. R.H. performed genome editing. R.H. and Q.A. performed CLIP experiments. J.P. did the computational analyses. J.P., A.G., R.H. and H.L.H. designed bioinformatics strategies. H.L.H., R.H., A.G. and J.P. wrote the paper.

FUNDING

Centre National de la Recherche Scientifique; Ecole Normale Supérieure; Agence Nationale de la Recherche (ANR) [2011-BLAN-01801 and ANR-13-BSV8-0023 to H.L.H.]; program ≪ Investissements d'Avenir ≫ launched by the French Government and implemented by ANR [ANR-10-LABX-54 MEMOLIFE and ANR-10-IDEX-0001-02 PSL* Research University]. Funding for open access charge: Agence Nationale de la Recherche.
Conflict of interest statement. None declared.

REFERENCES

1. Gerstberger,S., Hafner,M. and Tuschl,T. (2014) A census of human RNA-binding proteins. *Nat. Rev. Genet.*, **15**, 829–845.
2. Baltz,A.G., Munschauer,M., Schwanhausser,B., Vasile,A., Murakawa,Y., Schueler,M., Youngs,N., Penfold-Brown,D., Drew,K., Milek,M. *et al.* (2012) The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell*, **46**, 674–690.

3. Castello,A., Fischer,B., Eichelbaum,K., Horos,R., Beckmann,B.M., Strein,C., Davey,N.E., Humphreys,D.T., Preiss,T., Steinmetz,L.M. *et al.* (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, **149**, 1393–1406.
4. Singh,G., Pratt,G., Yeo,G.W. and Moore,M.J. (2015) The clothes make the mRNA: Past and present trends in mRNP fashions. *Annu. Rev. Biochem.*, **84**, 325–354.
5. Castello,A., Fischer,B., Hentze,M.W. and Preiss,T. (2013) RNA-binding proteins in Mendelian disease. *Trends Genet.*, **29**, 318–327.
6. Ule,J., Jensen,K.B., Ruggiu,M., Mele,A., Ule,A. and Darnell,R.B. (2003) CLIP identifies Nova-regulated RNA networks in the brain. *Science*, **302**, 1212–1215.
7. Licatalosi,D.D., Mele,A., Fak,J.J., Ule,J., Kayikci,M., Chi,S.W., Clark,T.A., Schweitzer,A.C., Blume,J.E., Wang,X. *et al.* (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464–469.
8. Lee,F.C.Y. and Ule,J. (2018) Advances in CLIP technologies for studies of Protein-RNA interactions. *Mol. Cell*, **69**, 354–369.
9. Darnell,R.B. (2010) HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip. Rev. RNA*, **1**, 266–286.
10. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M. Jr, Jungkamp,A.C., Munschauer,M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
11. Urlaub,H., Hartmuth,K. and Luhrmann,R. (2002) A two-tracked approach to analyze RNA-protein crosslinking sites in native, nonlabeled small nuclear ribonucleoprotein particles. *Methods*, **26**, 170–181.
12. Zhang,C. and Darnell,R.B. (2011) Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat. Biotechnol.*, **29**, 607–614.
13. Hauer,C., Curk,T., Anders,S., Schwarzl,T., Alleaume,A.M., Sieber,J., Hollerer,I., Bhuvanagiri,M., Huber,W., Hentze,M.W. *et al.* (2015) Improved binding site assignment by high-resolution mapping of RNA-protein interactions using iCLIP. *Nat. Commun.*, **6**, 7921.
14. Konig,J., Zarnack,K., Rot,G., Curk,T., Kayikci,M., Zupan,B., Turner,D.J., Luscombe,N.M. and Ule,J. (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.*, **17**, 909–915.
15. Van Nostrand,E.L., Pratt,G.A., Shishkin,A.A., Gelboin-Burkhart,C., Fang,M.Y., Sundararaman,B., Blue,S.M., Nguyen,T.B., Surka,C., Elkins,K. *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods*, **13**, 508–514.
16. Zarnegar,B.J., Flynn,R.A., Shen,Y., Do,B.T., Chang,H.Y. and Khavari,P.A. (2016) irCLIP platform for efficient characterization of protein-RNA interactions. *Nat. Methods*, **13**, 489–492.
17. Weyn-Vanhentenryck,S.M., Mele,A., Yan,Q., Sun,S., Farny,N., Zhang,Z., Xue,C., Herre,M., Silver,P.A., Zhang,M.Q. *et al.* (2014) HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep.*, **6**, 1139–1152.
18. Martin,G. and Zavolan,M. (2016) Redesigning CLIP for efficiency, accuracy and speed. *Nat. Methods*, **13**, 482–483.
19. Haberman,N., Huppertz,I., Attig,J., Konig,J., Wang,Z., Hauer,C., Hentze,M.W., Kulozik,A.E., Le Hir,H., Curk,T. *et al.* (2017) Insights into the design and interpretation of iCLIP experiments. *Genome Biol.*, **18**, 7.
20. Granneman,S., Kudla,G., Petfalski,E. and Tollervey,D. (2009) Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 9613–9618.
21. Sugimoto,Y., Konig,J., Hussain,S., Zupan,B., Curk,T., Frye,M. and Ule,J. (2012) Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol.*, **13**, R67.
22. Helder,S., Blythe,A.J., Bond,C.S. and Mackay,J.P. (2016) Determinants of affinity and specificity in RNA-binding proteins. *Curr. Opin. Struct. Biol.*, **38**, 83–91.
23. Le Hir,H., Sauliere,J. and Wang,Z. (2016) The exon junction complex as a node of post-transcriptional networks. *Nat. Rev. Mol. Cell Biol.*, **17**, 41–54.
24. Isken,O. and Maquat,L.E. (2008) The multiple lives of NMD factors: balancing roles in gene and genome regulation. *Nat. Rev. Genet.*, **9**, 699–712.
25. Chen,B., Gilbert,L.A., Cimini,B.A., Schnitzbauer,J., Zhang,W., Li,G.W., Park,J., Blackburn,E.H., Weissman,J.S., Qi,L.S. *et al.* (2013) Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell*, **155**, 1479–1491.
26. Gibson,D.G., Young,L., Chuang,R.Y., Venter,J.C., Hutchison,C.A. 3rd and Smith,H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, **6**, 343–345.
27. Sauliere,J., Murigneux,V., Wang,Z., Marquenet,E., Barbosa,I., Le Tonqueze,O., Audic,Y., Paillard,L., Roest Crollius,H. and Le Hir,H. (2012) CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. *Nat. Struct. Mol. Biol.*, **19**, 1124–1131.
28. Clarke,A.C., Prost,S., Stanton,J.A., White,W.T., Kaplan,M.E., Matisoo-Smith,E.A. and Genographic,C. (2014) From cheek swabs to consensus sequences: an A to Z protocol for high-throughput DNA sequencing of complete human mitochondrial genomes. *BMC Genomics*, **15**, 68.
29. Shah,A., Qian,Y., Weyn-Vanhentenryck,S.M. and Zhang,C. (2017) CLIP Tool Kit (CTK): a flexible and robust pipeline to analyze CLIP sequencing data. *Bioinformatics*, **33**, 566–567.
30. Van Nostrand,E.L., Gelboin-Burkhart,C., Wang,R., Pratt,G.A., Blue,S.M. and Yeo,G.W. (2017) CRISPR/Cas9-mediated integration enables TAG-eCLIP of endogenously tagged RNA binding proteins. *Methods*, **118–119**, 50–59.
31. Sander,J.D. and Joung,J.K. (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.*, **32**, 347–355.
32. Zhang,Z., Lee,J.E., Riemondy,K., Anderson,E.M. and Yi,R. (2013) High-efficiency RNA cloning enables accurate quantification of miRNA expression by deep sequencing. *Genome Biol.*, **14**, R109.
33. Hurt,J.A., Robertson,A.D. and Burge,C.B. (2013) Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. *Genome Res.*, **23**, 1636–1650.
34. Zund,D., Gruber,A.R., Zavolan,M. and Muhlemann,O. (2013) Translation-dependent displacement of UPF1 from coding sequences causes its enrichment in 3' UTRs. *Nat. Struct. Mol. Biol.*, **20**, 936–943.

Characterizing the splicing factor CWC27

Elucidating the mechanism of the spliceosome-dependent assembly of the EJC is one of the central themes of our laboratory. eIF4A3 pull-down in native conditions followed by mass-spectrometry revealed two major protein partners: splicing factors CWC22 and CWC27.

The aim of the study presented in the following article was to characterize the CWC27 interactions with CWC22 and eIF4A3, as well as its role in eIF4A3 assembly during the splicing reaction. My contribution to this work was to select Differentially Expressed Genes (DEG) in both CWC22 and CWC27 knock-downs relative to control experiments. We showed that several DEG were common to both conditions and their level of over- or under-expression was highly correlated.

This result, as well as several phenotypic similarities and overlapping alternative splicing events, suggested a common role of these two splicing factors in gene expression regulation.

5670–5683 *Nucleic Acids Research, 2020, Vol. 48, No. 10* *Published online 24 April 2020*
doi: 10.1093/nar/gkaa267

Structural and functional insights into CWC27/CWC22 heterodimer linking the exon junction complex to spliceosomes

Virginia Busetto[1], Isabelle Barbosa[1], Jérôme Basquin[2], Émelie Marquenet[1], Rémi Hocq[1], Magali Hennion[1], Janio Antonio Paternina[1], Abdelkader Namane[3], Elena Conti[2], Olivier Bensaude[1] and Hervé Le Hir[1,*]

[1]Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure, CNRS, INSERM, PSL Research University, 46 rue d'Ulm, 75005 Paris, France, [2]Department of Structural Cell Biology, MPI of Biochemistry, Munich, Germany and [3]Génétique des Interactions Macromoléculaires, Genomes and Genetics Department, Institut Pasteur, 25-28 rue du docteur Roux 75015 Paris, France

Received February 13, 2020; Revised April 01, 2020; Editorial Decision April 05, 2020; Accepted April 22, 2020

ABSTRACT

Human CWC27 is an uncharacterized splicing factor and mutations in its gene are linked to retinal degeneration and other developmental defects. We identify the splicing factor CWC22 as the major CWC27 partner. Both CWC27 and CWC22 are present in published B^{act} spliceosome structures, but no interacting domains are visible. Here, the structure of a CWC27/CWC22 heterodimer bound to the exon junction complex (EJC) core component eIF4A3 is solved at 3Å-resolution. According to spliceosomal structures, the EJC is recruited in the C complex, once CWC27 has left. Our 3D structure of the eIF4A3/CWC22/CWC27 complex is compatible with the B^{act} spliceosome structure but not with that of the C complex, where a CWC27 loop would clash with the EJC core subunit Y14. A CWC27/CWC22 building block might thus form an intermediate landing platform for eIF4A3 onto the B^{act} complex prior to its conversion into C complex. Knock-down of either CWC27 or CWC22 in immortalized retinal pigment epithelial cells affects numerous common genes, indicating that these proteins cooperate, targeting the same pathways. As the most up-regulated genes encode factors involved in inflammation, our findings suggest a possible link to the retinal degeneration associated with CWC27 deficiencies.

INTRODUCTION

Splicing of pre-messenger RNA (pre-mRNA) is performed by a very large RNA protein complex: the spliceosome. The stepwise assembly of spliceosomes involves the recruitment of snRNP (small nuclear ribonucleoproteins) and numerous proteins (1). Extensive rearrangements in composition and conformation accompany the formation of successive complexes named: E (early), A (pre-spliceosome), B (pre-catalytic spliceosome), B^{act} (activated spliceosome), B* (catalytically activated spliceosome), C, C* (catalytic spliceosome), P (post-catalytic splicesome) and ILS (Intron Lariat Spliceosome). B* spliceosomes catalyse the first catalytic step generating cleaved 5′-exon and intron/3′-exon lariat intermediates while C* spliceosomes catalyse the second step yielding ligated exons and intron lariat (2).

The yeast CWC27 (Complexed with Cef1 27) interacts with Cef1 protein, an essential splicing factor. The human CWC27 homologue is also named NY-CO-10. In both human and yeast spliceosomes, CWC27 is part of the B^{act} complexes (3–5) and leaves before its conversion to B* (5,6). CWC27 comprises an inactive N-terminal peptidyl-prolyl isomerase (PPIase) domain that has been conserved throughout evolution from yeast to mammals, followed by an elongated, unstructured and solvent-exposed C-terminal domain (7). Mutations that are expected to generate truncations of CWC27 unstructured C-terminal domain have been identified in human patients with retinal degeneration with or without other developmental defects (8). In mouse models, CWC27 knock-out is lethal while a C-terminal protein-truncating mutation leads to retinal degeneration, suggesting that the N-terminal CWC27 PPIase domain is essential for viability (8). Despite being associated with the spliceosome at a specific step, the molecular function of CWC27 remains unknown.

To unravel its function, we investigated CWC27 co-immunoprecipitating proteins. We found CWC22 (Complexed with Cef1 22), another evolutionarily conserved splicing factor, to be the CWC27 major interaction partner.

*To whom correspondence should be addressed. Tel: +33 144323945; Fax: +33 144323941; Email: lehir@ens.fr

In both *Saccharomyces cerevisiae* and human spliceosomes, CWC22 borders the spliceosome 'exon binding channel' and stabilizes the 5′ exon before the first step of splicing (3,5). In humans, CWC22 has been proposed to escort eIF4A3, a core exon junction complex (EJC) subunit, to the spliceosome (9,10). The EJC is an RNA binding protein complex found in metazoans and deposited around 27 nt upstream exon–exon junctions (11,12). It is composed of four core subunits (eIF4A3, MAGOH, Y14 and MLN51) and interacts with various peripheral factors (13). The EJC is recruited by spliceosomes and accompanies spliced mRNAs from the nucleus to the cytoplasm where it is removed by the first translating ribosome. It participates to pre-mRNA splicing regulation and contributes to mature mRNA export, localization, translation and degradation (13,14). According to published cryo-EM spliceosome structures, the complete EJC is bound to the 5′ exon in the spliceosome C complex (15,16). However, how and when the four core EJC subunits are recruited and assembled onto mRNA remains largely unknown.

Using purified recombinant proteins, we reconstituted a CWC27/CWC22/eIF4A3 ternary complex and solved its 3D structure by X-ray crystallography. This structure possibly corresponds to eIF4A3 earliest contacts with the spliceosome. We propose that CWC22 and CWC27 in the B^{act} complex form a landing platform for eIF4A3 before the release of CWC27 and the assembly of a complete EJC core bound to CWC22. In addition, transcriptomic data of knock-downs of CWC27 and CWC22 in an immortalized retinal pigment epithelial cell line revealed that these proteins target the same pathways. Noteworthy, genes in the inflammation pathways are among the most strongly upregulated, suggesting a link between retinal degeneration and CWC27 deficiency.

MATERIALS AND METHODS

Cells maintenance and transfections

Human HeLa and Hek293T cells were propagated at 37°C in a humidified 5% CO_2 atmosphere in high glucose DMEM medium (31966-021, Life Technologies) supplemented with 10% fetal bovine serum and 100 U/ml Penicillin-Streptomycin (Life Technologies). For overexpression of CWC27 and eIF4A3 constructs, cells were transfected with JetPrime (Polyplus) according to manufacturer's instruction. Full-length human CWC27 cDNA was PCR amplified with Phusion DNA polymerase (New England Biolabs) from a HCT116 cDNA homemade library and cloned into p 3xFLAG-CMV-10 (Sigma). Truncated CWC27 versions were generated by inverted PCR from p 3xFLAG-CMV-10 CWC27. p3XFLAG-CMV eIF4A3 was obtained from (M.J. Moore). Point mutant D270G in eIF4A3 was generated by subjecting the p3XFLAG-CMV eIF4A3 plasmid to QuickChange Site-Directed Mutagenesis.

hTERT-RPE-1 cells were propagated at 37°C in a humidified 5% CO_2 atmosphere in DMEM/F-12 GlutaMAX medium (31331-028, Life Technologies) supplemented with 10% FBS and 100 U/ml penicillin–streptomycin 1× (Life Technologies). HeLa and hTERT-RPE-1 cells were genotyped by Eurofins Forensik and routinely tested for mycoplasma by PCR.

Immunoprecipitation and Western Blotting

Cells were lysed with RIPA buffer (20 mM Tris–HCl pH 7.5, 150 mM NaCl, 1 mM Na2EDTA, 1 mM EGTA, 1% NP40, 1% sodium deoxycholate, RQ1 DNase (Promega, 1:50) and Protease inhibitor (Sigma, 1:100)). RNase A+T1 (Thermo Scientific, 1:200) was added or not to the sample. IP was performed overnight with 1 mg of total protein and 40 μl of Anti-FLAG M2 Magnetic Beads (Sigma) or 40 μl of Dynabeads Protein A (Life Technologies) linked to the desired antibody. Washes were performed with IP150 buffer (10 mM Tris–HClL (pH 7.5), 150 mM NaCl, 2.5 mM $MgCl_2$, 1% NP-40). After elution with SDS loading dye, samples were separated by electrophoresis in 4–12% Tris-glycine SDS/PAGE (Life Technologies) and were transferred onto 0.2-μm nitrocellulose membranes (Protan-BA83; GE Healthcare) using Thermofisher Transblot systems. Membranes were blocked in PBS with 10% (w/v) milk and 0.05% Tween-20 (Euromedex) before incubation with primary antibodies diluted 1:1000 in PBS 0,05% Tween for 1 h at RT or overnight at 4°C. Anti-CWC22, anti-eIF4A3, anti-MAGOH, anti-Y14 (9) and anti-CWC27 (Atlas, #HPA020344), anti-GAPDH (Cell signaling Technology, 2118S), anti-FLAG (Sigma, F7425) were used. After washing with 1× PBS, membranes were incubated with stabilized goat anti-rabbit secondary antibodies (1:10 000; Promega) and visualized using SuperSignal West Femto (Thermo Scientific) with LAS 4000 mini (GE Healthcare).

Immunofluorescence

Cells were seeded on coverslips coated with poly-lysine (Sigma, P1524) and fixed in 4% paraformaldehyde before permeabilization in PBS-Triton X (0.1%) for 2 min. After blocking, coverslips were incubated for 1 h at RT with the primary antibody diluted in PBS–BSA 1%. Nuclei were stained with Hoechst (diluted 1:400 in PBS-BSA 1%). Coverslips were then incubated for 1 h at room temperature with secondary antibodies (conjugated with Alexa Fluor 488 or Alexa Fluor 546 or Alexa Fluor 647 fluorochrome) diluted in PBS-BSA 1%. Coverslips were mounted in 5 μl of Fluoromount-G (Southern Biotech®) medium. Pictures were taken on Nikon Ti LGM. Images were processed and analyzed with Fiji software.

Genome editing

HeLa cells were co-transfected with two plasmids. The first one derived from pX335-U6-Chimeric_BB-CBh-hSpCas9n(D10A) (from E. Bertrand, IGMM Montpellier), expresses the nickase version of *Streptococcus pyogenes* Cas9 (Cas9n) and the two gRNAs (gRNA1 (5′-GGCCGCTCTCATCCCCCGTA-3′) and gRNA2 (5′-GCTCATCTTGGTCAGTACAA-3′). The other plasmid contains the repair sequence comprising the puromycin gene flanked by two lox sites. Puromycin-resistant colonies were isolated and expanded. Homozygous edited cell clones were identified by PCR on genomic DNA and by western

blot with an anti-CWC27 antibody (Atlas, #HPA020344). To remove the puromycin gene, a FLAG-CWC27 expressing clone was transfected with a plasmid expressing both the Cre-recombinase and the Geneticin-resistance genes (X. Morin, IBENS, Paris) and maintained in Geneticin (G418, *Thermo Scientific*) containing medium for 40 hours to select transiently transfected cells. After clonal isolation and expansion, removal of the puromycin gene was checked by PCR on genomic DNA and CWC27 expression was analysed by Western Blot with an anti-CWC27 antibody.

Mass spectrometry

LC–MS/MS analysis of FLAG-CWC27. 10^7 HeLa WT and HeLa FLAG-CWC27 cells were lysed in HKM300 buffer (10 mM HEPES pH 7.5, 10 mM KCl, 1.5 mM $MgCl_2$, 300 mM NaCl, 0.2 mM EGTA, 0.5% NP-40). A nuclear fraction (P) was pelleted 10 min at 400 g. The P fraction was resuspended with 1 ml HKM300 digested by DNase RQ1 (1:50) 10 min on ice, sonicated and centrifuged at 10 000 g for 10 min at 4°C. The supernatant (1 mg of protein in a final volume of 1 ml) was incubated overnight at 4°C with 40 μl of Anti-FLAG M2 Magnetic Beads (Sigma). The beads were washed three times 5 min at 4°C in HKM300.

Spin-dried beads were digested overnight at 37°C by sequencing grade trypsin (12,5 μg/ml; Promega Madison, WI, USA) in 20 μl of 25 mM NH_4HCO_3. The digested peptide mixture was loaded on a Q-Exactive plus system coupled to a Nano-LC Proxeon 1000 column equipped with an EASY-Spray ion source (Thermo Scientific). Peptides were separated by chromatography on Acclaim PepMap100 C18 pre-column (2 cm, 75 μm i.d., 3 μm, 100 Å), Pepmap-RSLC Proxeon C18 column (50 cm, 75 μm i.d., 2 μm, 100 Å) with a gradient from 95% solvent A (water, 0.1% formic acid) to 35% solvent B (100% acetonitrile, 0.1% formic acid) over a period of 97 min at 300 nl/min flow rate. Peptides were analysed in the Orbitrap cell, in full ion scan mode, at a resolution of 120 000 (at m/z 200), with a mass range of m/z 350–1550 and an AGC target of 4 × 105. Fragments were obtained by high collision-induced dissociation (HCD) activation with a collisional energy of 30%, and a quadrupole isolation window of 1.6 Da. MS/MS data were acquired in the Orbitrap cell. Precursor priority was highest charge state, followed by most intense. Peptides with charge states from 2 to 8 were selected for MS/MS acquisition. The maximum ion accumulation times were set to 100 ms for MS acquisition and 60 ms for MS/MS acquisition.

LC–MS/MS analysis of CWC22 and eIF4A3 immunoprecipitates. Nuclei were prepared from HeLa cells essentially as previously described by A. Lamond (17). The clean pelleted nuclei were resuspended in 5 ml RIPA buffer (50 mM Tris pH 7.5, 150 mM NaCl, 1% NP-40, 0,5% deoxycholate) with antiprotease cocktail, RQ1 RNAse-Free DNAse (1:50 volume, Promega), RNAse A (1:100 volume, Thermo Scientific) and RNAse T1 (1:100 volume, Thermo Scientific), sonicated at 4°C. The lysate was centrifuged at 2800 g for 10 min at 4°C. Supernatants were incubated for 2 h at 4°C with 100 μl protein A-coupled Dynabeads (Life Technologies) either or not (control) crosslinked with dimethylpimelidate to affinity-purified anti-eIF4A3 or anti-CWC22. The beads were next washed three times with IP buffer 300 (10 mM Tris–HCl, pH 7.5, 300 mM NaCl, 2.5 mM $MgCl_2$, 1% NP-40, 1% protease-inhibitor mixture) and incubated 20 min at 25°C with 50 U RNase A and 25 U RNAse T1 in 200 μl of IP buffer 150. Following three washes with IP buffer 300. Proteins were eluted with 20 ng/μl of the appropriate immunogenic peptide.

A short SDS-PAGE (dye-front at 1 cm from the bottom of the well) was used as a cleanup step. Gel slices were washed in water and proteins were reduced with 10 mM DTT before alkylation with 55 mM iodoacetamide. After dehydration with 100% (v/v) acetonitrile, we performed in-gel digestion using trypsin/Lyc-C (Promega) overnight in 25 mM NH_4HCO_3 at 30°C. The peptide mixture was analyzed by LC-MS/MS using an RSLCnano system (Ultimate 3000, Thermo Scientific) coupled to an Orbitrap Fusion mass spectrometer (Thermo Scientific). Peptide separation was performed on a C18-reversed phase column (75 mm ID × 50 cm; C18 PepMapTM, Dionex) at a flow rate of 400 nl/min and an oven temperature of 40°C. The loading solution was 0.1% trifluoroacetic acid and 2% acetonitrile and for elution 100% water with 0.1% formic acid for channel A, and 0.085% formic acid and 100% acetonitrile for channel B. The peptides were eluted with a linear multistep gradient of 1–6% solution B in 1 min, of 6–9% solution B in 11 min of 9–32% solution B in 82 min, and of 32–40% solution B in 6 min. We acquired Survey MS scans in the Orbitrap on the 400–1500 m/z range with the resolution set to a value of 120 000 and a 4 × 105 ion count target. Each scan was recalibrated in real time by co-injecting an internal standard from ambient air into the C-trap. Tandem MS was performed by isolation at 1.6 Th with the quadrupole, HCD fragmentation with normalized collision energy of 35, and rapid scan MS analysis in the ion trap. The MS2 ion count target was set to 104 and the max injection time was 100 ms. Only those precursors with charge state 2–7 were sampled for MS/MS. The dynamic exclusion duration was set to 60 s with a 10 ppm tolerance around the selected precursor and its isotopes. The instrument was run in top speed mode with 3 s cycles.

Mass spectrometry data analysis

The raw mass spectrometry data were analyzed by MaxQuant software (version 1.6.3.4) (18) using the embedded Andromeda search engine using the human protein database downloaded from Uniprot (20181204, 95146 entries) and completed with the contaminant list from MaxQuant. Up to two missed cleavages were allowed. The precursor mass tolerance was set to 4.5 ppm and the fragment mass tolerance to 0.5 Da. Carbamidomethylation of Cysteine residues was set as fixed modification and acetylation of protein N-terminus, oxidation of Methionine and deamidation of Asparagine and Glutamine were set as variable modifications. Minimal peptide length was set to seven amino acids. Second peptide option search was allowed. A false discovery rate (FDR) of 1% was independently applied for both peptide and protein identification. The 'match between runs' (MBR) option was allowed with a match time window of 1 min and an alignment time window of 20 min.

In the case of identified peptides that are shared between two proteins, these were combined and reported as a single protein group. Label-free quantification (LFQ) option was enabled, with at least two peptides required for LFQ measurements. LFQ was done using both unique and razor peptides for each protein.

Bioinformatic analysis of the MaxQuant/Andromeda workflow output and the analysis of the abundances of the identified proteins was performed from the 'proteinGroups.txt' of MaxQuant output file with the Perseus software (version 1.6.2.3) (19). The lists of identified proteins were filtered to eliminate reverse hits and known contaminants. LFQ values were further transformed to a $\log_2$ scale. The missing values were imputed from normal distribution with a width of 0.3 and a down-shift of 1.8 to simulate signals from low abundant proteins. To distinguish specifically interacting proteins from the background, protein abundances were compared between sample and control groups, using the Student's t-test statistic (FDR ≤ 0.01, S0 $= 2$, $n =$ 3 independent measurements), and results were visualized as volcano plots.

Recombinant proteins and *in vitro* interaction assays

Purification of recombinant proteins CBP-CWC22-S (pHL599), CBP-eIF4A3 (pHL241), eIF4A3 (pHL48) was previously described (9,20). For CBP-CWC22-MIF4G (pHL988), residues 117–406 were inserted in a variant of pET28a to fuse an N-terminal CBP tag and a C-terminal His6 Tag (Chamieh *et al*. 2008). For PTS-CWC27 (pHL1584), coding sequences of human CWC27 (1-472, Uniprot Q6UX04) were cloned between SalI and NotI in pET28a (Novagen) allowing for the fusion of N-terminal tags Protein A and TwinStrep between NheI and SalI and a C-terminal His6 Tag. For CWC27-C (pHL1553), residues 354–472 were PCR amplified and inserted between NheI and XhoI sites in pET28a allowing fusion of a C-terminal His6 Tag. The PTS-CWC27 protein fragments were successively purified on Nickel column (Ni-NTA, Clontech) and on StrepTactin affinity column (IBA). CWC27-iso2 (Uniprot Q6UX04-2) expressed as Sumo cleavable N-terminal His6 fusion protein was incubated with Senp2 protease overnight at 4°C and dialyzed against 50 mM Na_2HPO_4 pH 7.5, 150 mM NaCl. *In vitro* interaction assays were performed as previously described (20). For PTS pulldown, 12 μl of pre-blocked StrepTactin affinity beads (50% slurry, IBA) was used and precipitated proteins were eluted with 1× SDS loading buffer.

X-ray structure

For X-ray structure, recombinant proteins CWC22 and CWC27 were co-expressed in *E. coli* BL21 (DE3) grown in TB-medium at 37°C, as GST-3C-N-terminal His6 fusion or Sumo cleavable N-terminal His6 fusion proteins respectively. EIF4A3 was expressed as a Sumo cleavable N-terminal His6 fusion. Overexpression was induced at 18°C with 0.5 mM IPTG. Cells were lysed by sonication in 50 mM Na_2HPO_4 pH 7.5, 250 mM NaCl, 10 mM imidazole, 1 mM PMSF, and 25 mg/ml DNaseI, and the extract was cleared by centrifugation (4°C, 75 000 g, 30 min). In a first step, proteins were purified via a Ni^{2+}-NTA affinity column (5 ml, GE healthcare). In order to remove N-terminal His6-tags, proteins were incubated with 3C or Senp2 proteases overnight at 4°C and dialyzed against 50 mM Na_2HPO_4 pH 7.5, 150 mM NaCl for subsequent heparin chromatography (5 ml Heparin Q sepharose, GE Healthcare). Protein complexes were isolated by size exclusion chromatography (SEC) after concentrating to 20–30 mg/ml in a buffer containing 10 mM HEPES pH 7.5, 150 mM NaCl and 1 mM DTT using a HiLoad Superdex 75 column (GE Healthcare). The complex was stored at 80°C in SEC buffer.

Crystallization of CWC22-CWC27-EIF4A3(RecA2) complex. The complex was set up for crystallization at 20 mg/ml in SEC buffer by sitting-drop vapor diffusion in 0.2 ul drops obtained by mixture of equal volumes of protein and crystallization solution. Crystals appeared after 2 days at 4°C as monoclinic prism after mixing with 20% (w/v) PEG20000, 50 mM MES pH 6.5 and were cryoprotected in reservoir solution containing 33% (v/v) ethylene glycol prior to flash freezing in liquid nitrogen.

Data collection and crystal structure determination. Diffraction data were collected at the PXII beamline at the Swiss Light Source (SLS) in Villigen, Switzerland, and were processed with XDS (21) prior to scaling with Aimless of the CCP4 package (22). The structure of CWC22-CWC27-EIF4A3 (RecA2) was determined from selenomethionine substituted protein crystals. Single anomalous dispersion data were recorded at the Se peak wavelength, and AUTOSOL as part of the PHENIX package was used to locate Se sites. A combination of single anomalous dispersion and molecular replacement was used to solve the structure at 3.0 Å using known EIF4A3-CWC22 structure (PDB IDs: 4C9B) using the program Phaser (23). The asymmetric unit contained four molecules of the complex. The model was completed by iterative cycles of model building in COOT (24), followed by refinement in PHENIX (25) using NCS restraints.

mRNA-seq and data analysis

For mRNA-seq, h-TERT RPE-1 cells were transfected at 60–70% confluency with 9 μl of Lipofectamine and 1.5 μl of 20 μM DsiRNAs (Integrated DNA Technologies) CWC27, CWC22 or control. A mix of two different DsiRNA targeting different regions of the CWC27 and CWC22 genes was used. Transfections for replicates, were performed independently. Forty eight hours after transfection, RNAs were extracted using Monarch Total RNA Miniprep Kit (New England Biolabs). DsiRNA efficiency was checked by WB. RNA-seq was performed by Fasteris on paired-end libraries run on Illumina HiSeq using 2 × 150 bp.

Bioinformatic analysis. After trimming of the adapters using cutadapt (26), the reads were mapped on hg38 (Gencode GRCh38.p12.genome.fa) using STAR (version 020201,(27)) with default parameters, and adding the -quantMode GeneCounts option to generate gene count files. Gencode hg38 V29 gtf annotations were used. BamCoverage from deepTools (28) was used to generate bigwig file for quick visualization of the read counts on IGV

(29). Principal component analysis was performed using the 'pcaMethods' R package directly on gene counts. Intron retention was assessed using iREAD 0.8.0 (30), adding a post-treatment to remove the regions where two genes overlap. Resulting intron count files of controls *vs* CWC22 or CWC27 siRNA triplicates were processed with DESeq2 to find significant changes associated with the Knock Downs (KDs). Introns with $|\log_2(FC)| > 1$ and P-value < 0.05 were considered as significantly retained. Jsplice (31) was run in junction mode to find differential splicing. Alternative splicing modules (ASMs) with $|FC| > 1.5$ and P-value < 0.05 were considered as significantly differentially spliced. We used JSplice classification for alternative splicing events. For differential expression, DESeq2 (32) was run on gene counts of controls versus CWC22 or CWC27 siRNA triplicates, with default parameters. Genes with $|\log_2(FC)| > 1$ and P-value < 0.05 were considered as significantly regulated. Gene Ontology (GO) analysis was performed with GOrilla (33,34) on genes upregulated with a P-value of 0.05. 'Process' was chosen as ontology and default parameters were used. Revigo (35) was used for summarizing GO categories and for generating the graphics. Default parameters were used. TreeMap was used as representation. Subcategories were removed from the original TreeMap graph but they are indicated in the table.

RESULTS

CWC27 associates with CWC22 and eIF4A3

To better characterize CWC27 function, we first looked for its protein partners by immunoprecipitation. To maximize immunoprecipitation specificity and efficiency, CWC27 was epitope tagged with a N-terminal 3xFLAG peptide. To minimize artefacts due to overexpression, the FLAG was inserted by CRISPR–Cas9 editing of all CWC27 alleles in HeLa cells (Supplementary Figure S1A). Indeed, a Western Blot using affinity purified anti-CWC27 polyclonal antibodies confirmed that expression level of FLAG-CWC27 is similar to that of wild-type CWC27 in the parental cell line (Supplementary Figure S1B). Triplicates of FLAG immunoprecipitation from FLAG-CWC27 cells were analysed by label-free quantitative mass spectrometry (LC–MS/MS) using the parental HeLa cells as a negative control. The splicing factor CWC22 is found as the most significant interacting protein (Figure 1A). CWC22 is also detected by Western blotting after FLAG immunoprecipitation from FLAG-CWC27 cells but not from the parental ones (Figure 1B, lanes 5–8). The interaction between CWC27 and CWC22 is not RNA dependent since co-precipitation is unaffected by RNase treatment (Figure 1B, lane 6). As a further confirmation, CWC27-specific antibodies immunoprecipitated CWC22 (Figure 1C, lane 3) and two distinct CWC22-specific antibodies immunoprecipitated CWC27 (Figure 1C, lanes 4 and 5). CWC22 is bound to the EJC inside the spliceosome after the first step of splicing (36,37) and it is important for the recruitment of eIF4A3 into spliceosome (9,10,38). eIF4A3 is co-immunoprecipitated by FLAG antibodies from FLAG-CWC27 cell lysates with or without RNase treatment (Figure 1B) and by affinity purified CWC22 or CWC27 antibodies from HeLa cell lysates (Figure 1C, lanes 3–5). The other EJC subunits Y14 and MAGOH are weakly immunoprecipitated by CWC22 antibodies and hardly detected following CWC27 immunoprecipitation (Figure 1C, lanes 3–5).

We next wanted to explore the CWC22 and eIF4A3 interacting network, notably with nuclear splicing factors. For this, we used sucrose density centrifugation to isolate HeLa cell nuclei (see Materials and Methods) before immunoprecipitation. Then, we performed triplicate immunoprecipitations coupled to label-free quantitative mass spectrometry (LC–MS/MS) using this time affinity-purified polyclonal anti-CWC22 and anti-eIF4A3 antibodies. 88 and 107 statistically significant proteins were identified with CWC22 and eIF4A3 antibodies, respectively (Figure 1D and Supplementary Tables S1 and S2). Among the 17 statistically significant proteins common to both CWC22 and eIF4A3 immunoprecipitates, 10 are splicing-related factors and the remaining ones are linked to other mRNA maturation steps (Supplementary Table S3). Several splicing factors such as SLU7, CWC15, CWC22, CWC27 and CDC5L (the human orthologue of Cef1) belong to B^{act} spliceosome and subsequent complexes. Noteworthy, CWC27 is one of the most enriched proteins. Taken together these results indicate that CWC27, CWC22 and eIF4A3 interact with each other in spliceosomes.

A CWC22/CWC27/eIF4A3 ternary association requiring the CWC27 C-terminal domain

To map the interaction domains, we transiently expressed FLAG-tagged versions of the full-length (1–472) or truncated CWC27 proteins. All proteins were correctly expressed and localized in the nucleus (Figure 2A and Supplementary Figure S2). CWC22 and eIF4A3 both co-precipitated with the full-length and a truncation lacking the N-terminal PPIase domain (170–472) (Figure 2A, lanes 5 and 7) while they did not co-precipitate with truncations lacking fragments of the unstructured C-terminal domain (1–306) and (1–388) (Figure 2A, lane 6 and Supplementary Figure S2A, lanes 9 and 10), despite the truncated proteins remaining localized in the nucleus (Supplementary Figure S2B). These results indicate that the last 84 amino acids of CWC27 are required to interact with both CWC22 and eIF4A3.

Conversely, transfected FLAG-CWC22 and FLAG-eIF4A3 proteins co-immunoprecipitate CWC27 (Figure 2B, lanes 6 and 7). We next investigated two mutated versions of eIF4A3 known to affect its binding to CWC22. A quadruple mutation of the 298–301 sequence at the surface of eIF4A3 (REAN>HARD), called eIF4A3-mutG mutation, had been shown to strongly reduce eIF4A3-CWC22 interaction *in vitro* (9). The eIF4A3 D270G mutation is associated with the Richieri Costa Pereira syndrome (39). eIF4A3 D270 directly contacts lysine K174 of CWC22 (40) but its impact on eIF4A3-CWC22 interaction had not been investigated. We observed that both mutations not only reduce the interaction between CWC22 and eIF4A3 but also their interaction with CWC27 (Figure 2B and C). These observations strongly suggest that in live cells CWC27 forms a ternary complex with CWC22 and eIF4A3, requiring an intact CWC22/eIF4A3 interaction.

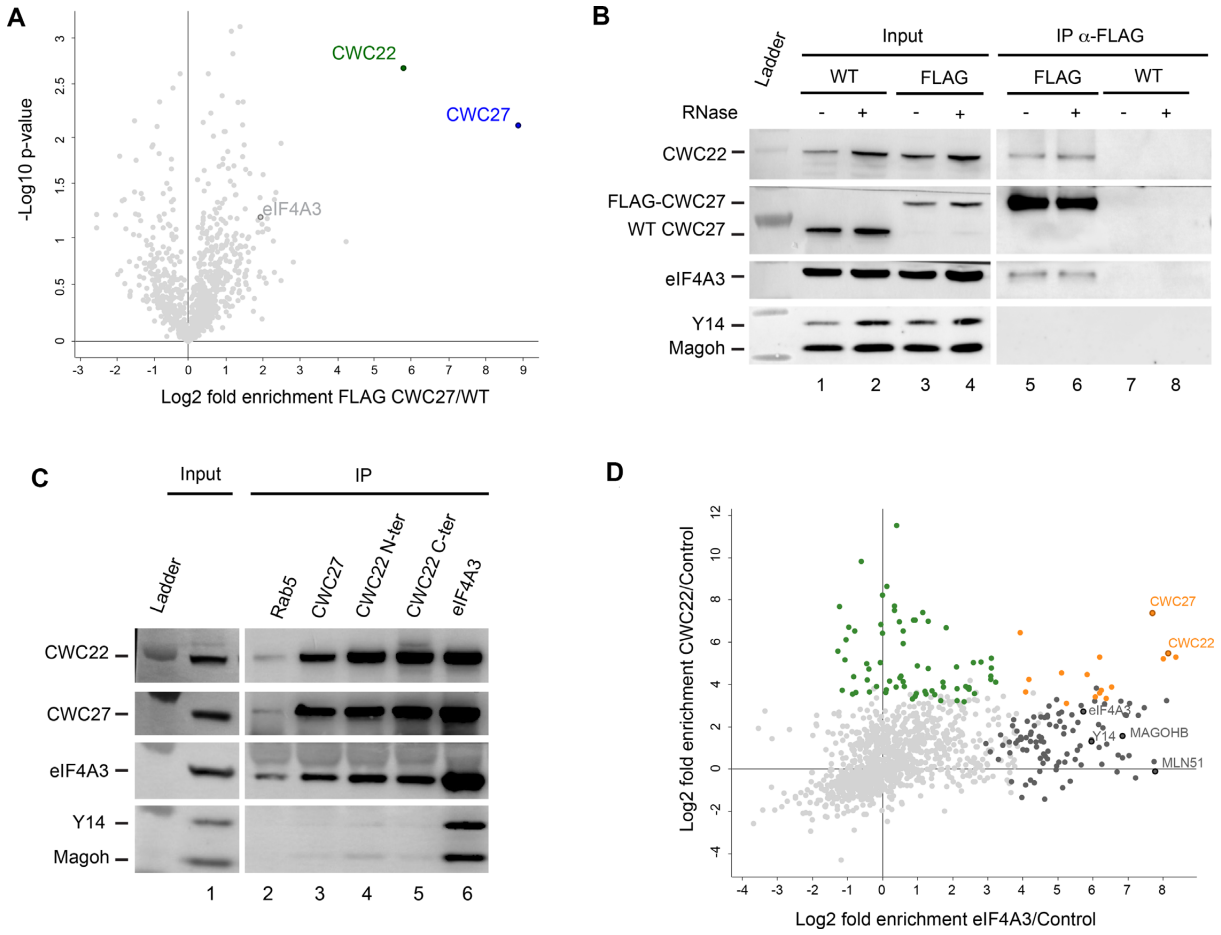

Figure 1. CWC27 is associated with CWC22 and eIF4A3. (**A**) Volcano plot of protein enrichment for FLAG-CWC27 immunoprecipitation versus control purification. Not-significant proteins (FDR > 0.01) are indicated in light grey. (**B**) Western blots of proteins coimmunoprecipitated with FLAG-CWC27. Protein extracts from WT and FLAG-CWC27 (FLAG) cells were treated with (+) or without (−) RNase. Detected proteins are indicated on the left. (**C**) Same as (B), with affinity purified antibodies anti-Rab5, anti-CWC27, anti-CWC22 (N-terminus), anti-CWC22 (C-terminus) and anti-eIF4A3 used for immunoprecipitations with protein extracts from HeLa cells treated with RNase. (**D**) Enrichment plot of proteins from eIF4A3 and CWC22 purifications. These enrichments were computed using the same control purifications without antibodies during the immunoprecipitation step. Proteins significantly (FDR ≤ 0.01) enriched in anti-CWC22 immunoprecipitation (green), in anti-eIF4A3 (dark gray) or in both (orange) are indicated on the plot. Not-significant proteins are indicated in light gray.

In vitro reconstitution of a CWC27/CWC22/eIF4A3 ternary complex

To better characterize the CWC27, CWC22 and eIF4A3 association, we used *in vitro* reconstitution experiments with recombinant proteins purified from bacteria (Figure 3A). All proteins were fused with a C-terminal His6 tag and when indicated, with a Calmodulin Binding Peptide (CBP) or a tandem Protein A-Twin Strep (PTS) N-terminal tag. Full-length CWC22 does not express well in bacteria, therefore we used a shorter version (CWC22-S; residues 100–665) more suitable for *in vitro* binding studies (9). The proteins were mixed, incubated with calmodulin beads, and after extensive washes, calmodulin bound protein(s) were fractionated by SDS-PAGE and visualized by Coomassie staining. CBP-CWC22-S co-retains PTS-CWC27 while CBP-eIF4A3 does not co-retain more PTS-CWC27 than control (Figure 3B, lanes 1, 3 and 4). As previously described (9), CBP-CWC22-S co-retains some eIF4A3 above control (Figure 3B, lanes 2 and 5). eIF4A3 co-retained with CWC22-S increases significantly in the presence of PTS-CWC27 (Figure 3B, lane 6). We next used a preformed CBP-CWC22-S/CWC27 heterodimer obtained by co-expression in bacteria and mixed it with eIF4A3. Again, the heterodimer co-retains more eIF4A3 than CBP-CWC22-S alone (Supplementary Figure S3). Conversely, PTS-CWC27 efficiently retains CBP-CWC22-S on StrepTactin beads whether eIF4A3 is added or not (Figure 3C, lanes 4 and 6). In contrast, addition of CBP-CWC22-S is an absolute requirement to retain eIF4A3 on the beads (Figure 3C, lanes 5 and 6). Taken together, these experiments suggest that eIF4A3 binds primarily to CWC22 and that CWC27 stabilizes this interaction.

In an attempt to define interaction domains, we performed the reconstitution experiments with protein fragments. The above described transfection experiments indicated that the last 84 aa of CWC27 isoform 1 (Uniprot Q6UX04-1) are required to interact with both CWC22 and

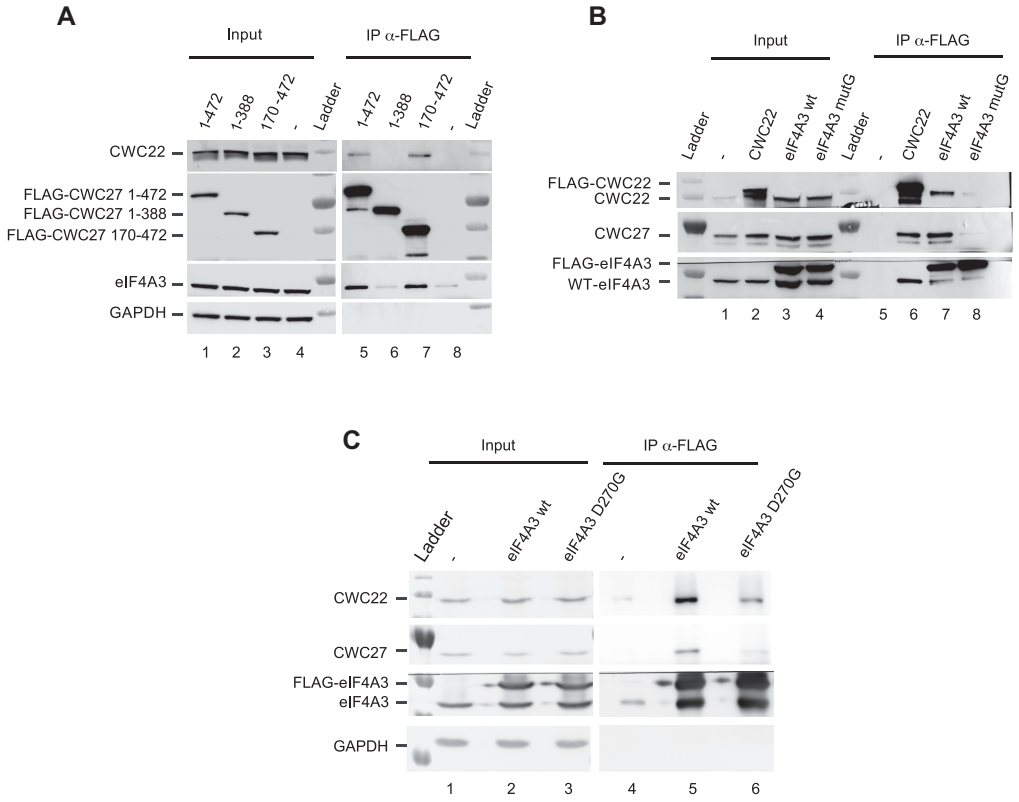

Figure 2. CWC27 C-terminal region mediates interaction with CWC22 and eIF4A3 (**A**) Western blots of proteins coimmunoprecipitated with FLAG-CWC27 isoforms or FLAG empty vector (−) transiently expressed in HeLa cells. Detected proteins are indicated on the left. (**B**) Same as (A), with transiently expressed FLAG-CWC22, FLAG-eIF4A3 WT or mutG. (**C**) Same as (A), with transiently expressed FLAG-eIF4A3 WT or the mutant FLAG-eIF4A3 D270G.

eIF4A3 in live cells. Therefore, we first repeated the *in vitro* binding assays using CWC27 isoform 2 (Uniprot Q6UX04-2) in which, due to an alternative splicing, the last 88 aa are replaced by 6 aa. Indeed, this isoform (CWC27-iso2) is not retained by CBP-CWC22-S whether or not eIF4A3 is added (Figure 3D, lanes 5 and 7). Conversely, a CWC27 C-terminal fragment (354–472) is retained by CBP-CWC22-S whether or not eIF4A3 is added (Figure 3D, lanes 6 and 8). This result demonstrates that CWC22 binds to the unstructured CWC27 C-terminal domain and not the N-terminal PPIase domain. eIF4A3 has been previously reported to interact with the MIF4G domain of CWC22 (9). A CWC22 fragment (119–431) containing this domain (CBP-CWC22-N) indeed retains eIF4A3 (Figure 3D, lane 9). This fragment also retains the CWC27 C-terminal fragment (Figure 3D, lane 11). It is thus possible to reconstitute a minimal complex (Figure 3D, lane 13) with the CWC22 MIF4G domain, the last 118 aa of CWC27 and eIF4A3 that might be suitable for structural studies.

3D structure of the CWC27/CWC22/eIF4A3 ternary complex

In order to obtain the 3D structure of the CWC27/CWC22/eIF4A3 ternary complex by X-ray crystallography, we expressed and purified recombinant MIF4G domain of CWC22 (residues 119–359), a fragment of the C-terminal region of CWC27 (320–431) and eIF4A3. No crystals were obtained upon large crystallization screening. A limited proteolysis experiment allowed us to identify a shorter CWC27 construct (378–431) still interacting with CWC22. A combinatory crystallization screening approach led us to crystallize the ternary complex CWC22 (119–359) / CWC27 (378–431)/eIF4A3 RecA2 domain (246–411). The crystal structure was solved by a combination of single-wavelength anomalous dispersion (SAD) using selenomethionine substitution (CWC22/CWC27) and molecular replacement (See methods). The structure is refined at 3.0 Å resolution, with a free *R* factor of 27%, a working *R* factor of 23% and good stereochemistry (Supplementary Table S4).

The final model encompasses the MIF4G domain of CWC22 (130–401), the RecA2 domain of eIF4A3 (246–411) and residues 378–426 of CWC27 (Figure 4A). The last five residues of CWC27 (427–431) as well as the N-terminal residues and a loop of CWC22 (116–122 and 142–148) are not visible in the electron density map. The RecA2 domain of eIF4A3 contacts CWC22 MIF4G domain, as observed in the previously published CWC22-eIF4A3 crystal structure. On the opposite side of the MIF4G domain, residues 378–402 of CWC27 form an extended helix that packs against a groove formed by a three alpha helices bundle of CWC22 (Figure 4A). The CWC27 C-terminal domain residues in contact with CWC22 are evolutionary conserved from yeast

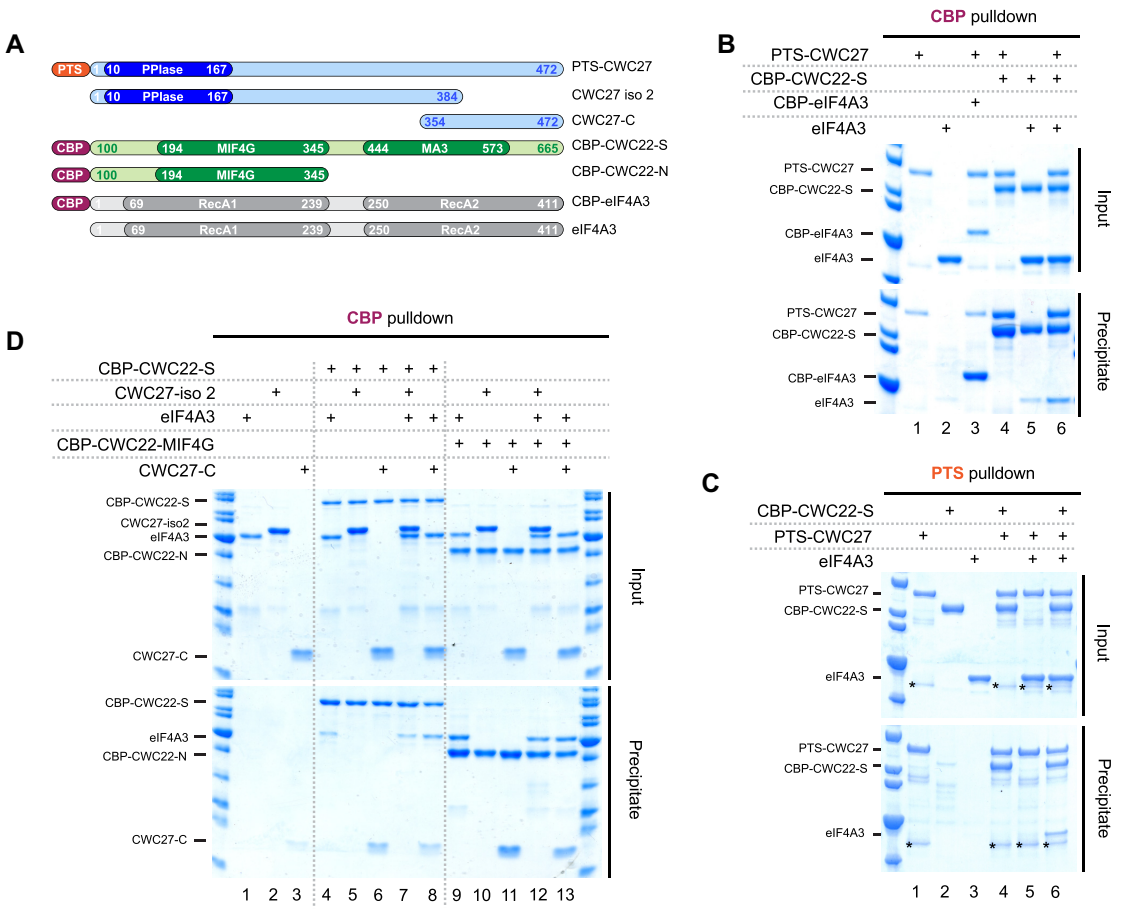

Figure 3. Direct interaction of CWC27 and eIF4A3 with CWC22 (**A**) Schematic representation of the different purified recombinant proteins (CWC27 in blue, CWC22 in green and eIF4A3 in gray) and the fused affinity tags (PTS in orange, CBP in purple). (**B**) Interaction of CWC27 with CWC22 and eIF4A3. Mixed recombinant proteins (input: 35% of total) and proteins retained with CBP-eIF4A3 (lane 3), CBP-CWC22-S (lanes 4 to 6) or not (lanes 1 and 2) on calmodulin beads (precipitate) were analyzed by SDS-PAGE. Proteins are indicated on the left. (**C**) Same as (B), with the indicated proteins retained with PTS-CWC27 (lanes 1, 4, 5 and 6) or not (lanes 2 and 3) on streptavidin beads. An unspecific contaminating protein co-purifying with PTS-CWC27 is marked (*). (**D**) Same as (b), with proteins retained with CBP-CWC22-S (lanes 4–8), with CBP-CWC22-MIF4G (lanes 9–13) or not (lanes 1–3) on calmodulin beads.

to mammals (Supplementary Figures S4A and S4b). The long CWC27 helix is followed by a loop (402–426) that folds around one side of MIF4G domain of CWC22 (Figure 4A). No direct contacts between eIF4A3 and CWC27 are detected.

CWC22 MIF4G and MA3 domains, as well as CWC27 PPIase N-terminal domain are clearly observed in cryo-EM structures of human B^{act} spliceosomes (3,5) (Figure 4B). However, neither the C-terminal region of CWC27 (427–472) nor eIF4A3 are visible in these structures. The MIF4G domain in our new structure was perfectly aligned to the one in the B^{act} spliceosome structure (3), and docking of the entire CWC27/CWC22/eIF4A3 new structure shows no particular clashes (Figure 4B). We then wanted to investigate what conformation could assume eIF4A3 in B^{act} spliceosomes. We aligned the CWC22 MIF4G domain of the CWC22/eIF4A3 crystal structure (40) (PDB: 4C9B) to the one present in the B^{act} spliceosome cryo-EM structure (3) (PDB: 6FF7). The RecA1 domain of eIF4A3 clashes with the spliceosomal factor EFTUD2 (Figure 4C), indicating that eIF4A3 must adopt in the spliceosome a closer conformation than the open conformation observed in the crystal structure of CWC22/eIF4A3.

In the C complex, after the first catalytic splicing reaction, CWC27 is no longer present and the MIF4G domain of CWC22 contacts the EJC, which is assembled onto mRNAs around 27 nt upstream the exon-exon junction (41). We docked our structure to that of the C spliceosome (PDB: 5YZG) and found that the loop of CWC27 clashes with the Y14 EJC subunit (Figure 4D). This is consistent with the fact that CWC27 leaves the spliceosome before EJC assembly. Our new data indicate that CWC27 is another player in eIF4A3 recruitment, with our structure illustrating the early contacts of eIF4A3 with B^{act} spliceosomes.

CWC27 and CWC22 knock-down in retinal cells impact common pathways linked to inflammation

Patients with genetic mutations in *CWC27* are prone to retinal degeneration (8). To explore the impact of CWC27 depletion on gene expression, we performed siRNA knockdown (KD) on immortalized hTERT RPE-1 cells from

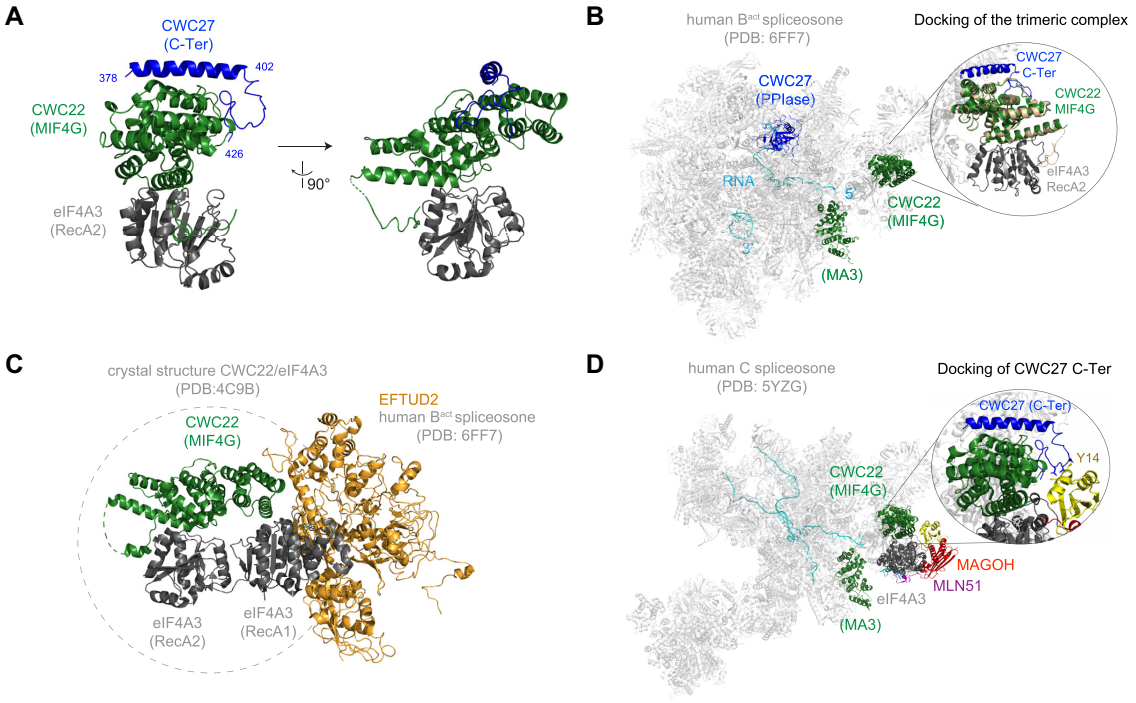

Figure 4. 3D structure of the complex CWC27/CWC22/eIF4A3 (**A**) Cartoon representations of the complex, shown in two orientations. Human CWC27 is in blue, human CWC22 MIF4G domain is in green and human eIF4A3 RecA2 domain is in dark gray. This and all other cartoon drawings were generated using PyMOL (http://www.pymol.org/). (**B**) Cartoon representation of human B^{act} spliceosome (PDB: 6FF7). The PPIase domain of CWC27 is in blue, the MIF4G and MA3 domains of CWC22 in green, the spliceosome in light gray and pre-mRNA in cyan. Zoom-in shows the docking of CWC27/CWC22/eIF4A3 structure in which CWC22 MIF4G domain is in wheat, while CWC27 and eIF4A3 RecA2 domain are colored as in (A). (**C**) Cartoon representation of CWC22/eIF4A3 (PDB: 4C9B) docked onto the structure of the B^{act} spliceosome (PDB: 6FF7). eIF4A3 RecA1 domain clashes with the EFTUD2 spliceosomal protein. EFTUD2 is in orange, CWC22 in green, eIF4A3 in dark gray. (**D**) Cartoon representation of human C spliceosome (PDB: 5YZG). Zoom-in shows the docking of CWC27, highlighting the clash between CWC27 loop and the EJC core protein Y14. MLN51 is in purple, MAGOH in red, Y14 in yellow, eIF4A3 in dark gray, CWC27 in blue and CWC22 in green.

human retinal pigment epithelium. These cells express CWC22, CWC27 and eIF4A3 correctly, and both EJC and the ternary complex CWC22/CWC27/eIF4A3 are detected as shown by co-immunoprecipitation of endogenous proteins (Supplementary Figure S5A). We performed separate KD of CWC27 and CWC22 followed by large-scale sequencing of mRNAs. eIF4A3 KD was not investigated as it resulted in rapid cell death within a few hours of treatment. The efficiency of CWC27 and CWC22 down-regulation was checked by RT-qPCR (Supplementary Figure S5B) and Western Blot analysis (Figure 5A). Interestingly, CWC27 KD reduces CWC22 protein levels and *vice versa,* while their respective transcripts are not affected. This shows that each protein stabilizes its partner, further supporting that the two proteins interact together *in vivo*. We sequenced KDs and control samples in triplicate with paired-end Illumina sequencing and obtained between 33 and 40 million reads per sample. After read mapping on hg38 with STAR (27), principal component analysis on gene counts showed clustering of the samples according to their experimental conditions (Supplementary Figure S5C). This, as well as visual inspection of read counts on IGV (29), validated sample reproducibility as well as the quality of library preparation and sequencing.

Since CWC22 and CWC27 are spliceosomal proteins, we first examined the impact of KD on splicing. 2385 and 1268 introns were significantly (P-value < 0.05, fold change $>$ 1.5) more retained in CWC22 and CWC27 KD respectively (Figure 5B). About half of the retained introns (619) after CWC27 KD were also retained after CWC22 KD (Figure 5C). CWC22 and CWC27 KD affected 500 and 290 alternative splicing events respectively (fold change $>$ 1.5; figure 5B and C). Of these events, 40% (132) were common to both CWC27 and CWC22 KD (Figure 5C). These results show that both proteins are involved in common splicing events.

The expression of 2040 and 1701 genes increased, and that of 1176 and 1526 decreased significantly (P-value<0.05, fold change $>$ 2) in CWC27 and CWC22 KD, respectively. Changes in gene expression in CWC27 KD were highly correlated to changes in CWC22 KD (Pearson $= 0.83$, P-value < 0.001) (Figure 5D). To annotate gene function, we performed a GO analysis on significantly up-regulated genes (P-value < 0.05) from both samples. Interestingly, they show enrichment in genes related to inflammation (Supplementary Table S5 and S6). Among the 10 most up-regulated genes in CWC27 KD ($>$30-fold), eight are linked to inflammation (Supplementary Table S7). Moreover, half of the 122 genes up-regulated $>$10-fold have a pro-inflammatory function, they correspond to cytokines (interferons, chemokines, members of the tumor necrosis factor super family), chemokine receptors, adhesion molecules, interferon-inducible transcripts, inflamma-

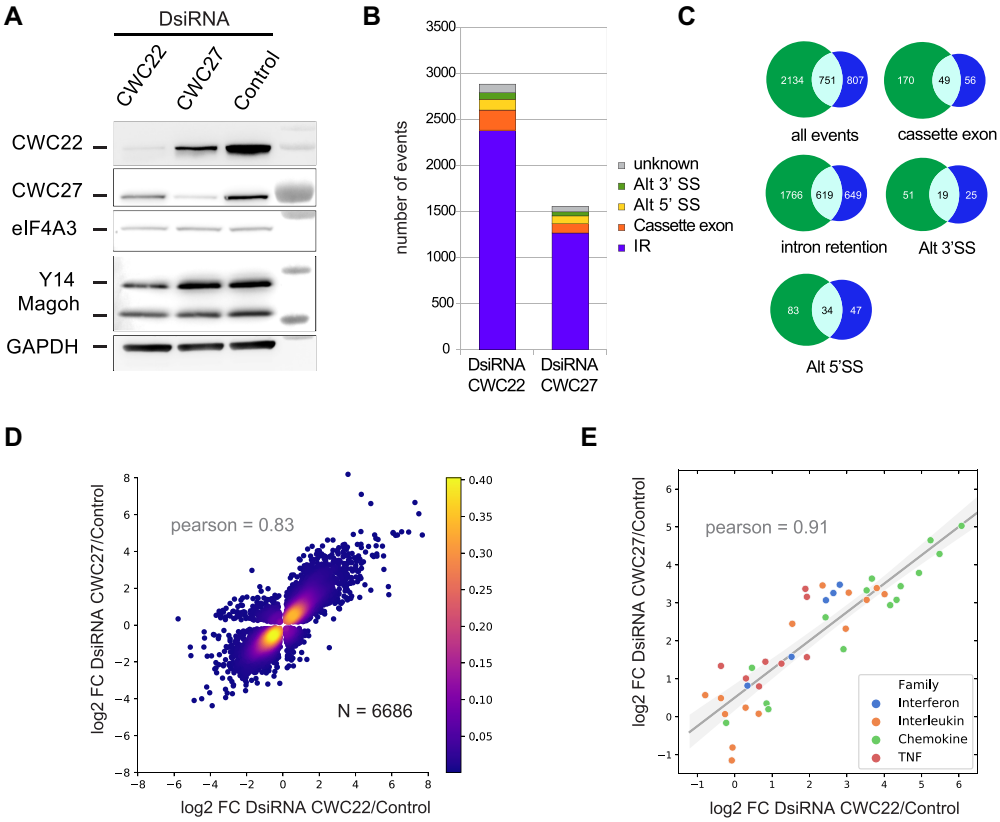

Figure 5. Impact of CWC27 and CWC22 KD on gene expression in RPE1 cells. (**A**) Western blots of CWC22 or CWC27 from hTERT RPE-1 cells treated with the corresponding DsiRNA. Detected proteins are indicated on the left. Molecular size markers are on the right. (**B**) Classification of alternative splicing events induced by CWC22 and CWC27 KDs. Intron retention (IR) events were identified with iREAD (30), using a minimum fold change of 2 (P-value < 0.05). Other splicing events were detected with JSplice (31), using a minimum fold change of 1.5 (P-value < 0.05). Alternative 3′- and 5′-splice sites: alt 3′ SS and alt 5′ SS, respectively. *Unknown* corresponds to complex splicing defects involving several junctions classified as unknown by JSplice. (**C**) Venn diagrams showing the overlap between significant splicing events changes in CWC22 (green) and CWC27 (blue) KDs. (**D**) Logarithmic fold change ($\log_2$ FC) of genes significantly up- or down- regulated ($N = 6686$, P-value < 0.05) in both CWC22 and CWC27 KD compared to control cells. Pearson correlation coefficient is 0.83 (P-value < 0.001). (**E**) Logarithmic fold change ($\log_2$ FC) of for expression of all detected cytokines in CWC22 and CWC27 KD compared to control cells. Linear regression (gray line) with 95% confidence interval of squared residuals (gray shade). Pearson correlation coefficient is 0.91 (P-value $< 10^{-16}$).

some components, inflammatory pathway modulators and actors of antigen presentation (Supplementary Table S7). The same genes are also up-regulated following CWC22 KD. For instance, all detected cytokines are up-regulated in both CWC27 and CWC22 KD with a fairly good correlation (Pearson coefficient of 0.91 and an associated P-value < 0.001) (Figure 5E). Among the 67 genes down-regulated >5-fold following CWC27 KD (Supplementary Table S8), some belong to the transforming growth factor beta signalling cascade, others are linked to the actin cytoskeleton, others are mitochondrial encoded transcripts for oxidative phosphorylation enzymes. The majority of these genes are also down-regulated following CWC22 KD (Supplementary Table S8). Together, our results show that CWC27 KD has a wide impact on gene expression. Moreover, down-regulated and up-regulated genes as well as alternative splicing events follow the same trend after CWC22 KD, indicating that common pathways are targeted by both proteins and that both proteins are physically and functionally linked.

DISCUSSION

In this study, we provide new structural and functional insights into the splicing factor CWC27. In both yeast and human, CWC27 is composed of an inactive PPIase domain followed by a long and disordered region. The PPIase domain of CWC27 is the only part of CWC27 visible in spliceosome cryo-EM structures. It plays a conserved role during spliceosome assembly, as it is positioned identically in yeast (4) and human (3,5) B^{act} spliceosomes, and it is released concomitantly with the RNF113A protein (CWC24 in yeast) during the B to B* spliceosome conversion (5,6). Prior to our work, little was known about the long C-terminal region of CWC27. In humans, truncations of this region are associated with retinitis pigmentosa and developmental defects (8). Here, we show that the C-terminus of human CWC27 directly contacts its splicing partner CWC22 and together, these proteins offer a landing platform for the EJC core component eIF4A3.

We find that CWC22 is the main protein interacting with CWC27 in cell lysates. By biochemical and structural

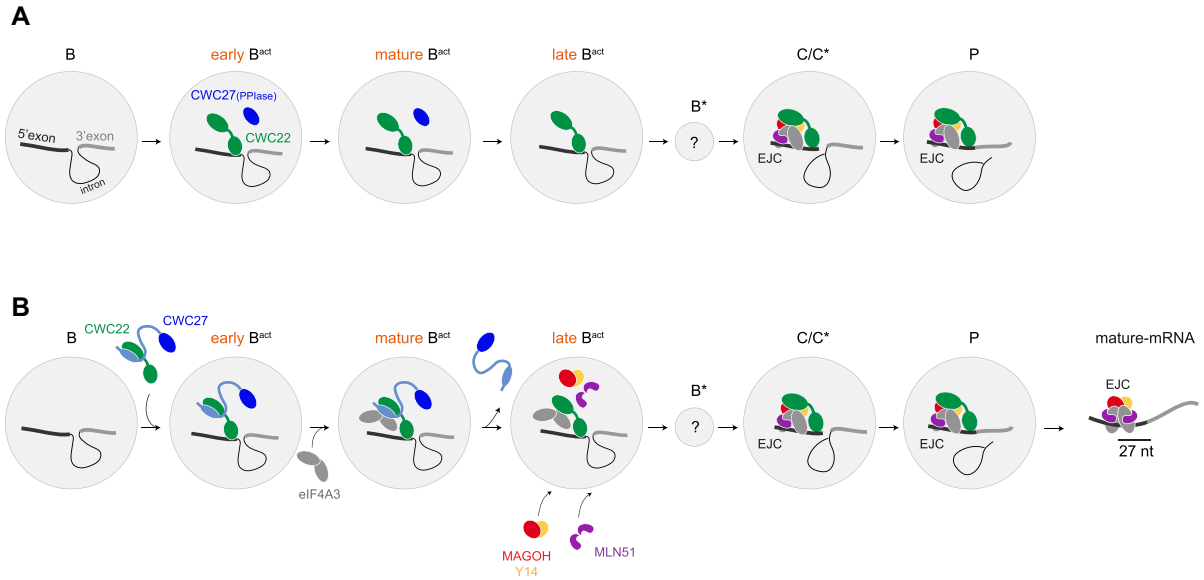

Figure 6. Model of EJC assembly by spliceosome (**A**) Cartoon representation of published cryo-EM structures of human spliceosomal complexes. Spliceosome complexes are represented by gray disks. Proteins involved in EJC core and assembly are colored. CWC27 PPIase domain is in dark blue, CWC22 in green. (**B**) Model of stepwise EJC assembly onto mRNA. The CWC27 C-terminal region is in light blue, with an ellipse representing the CWC27 alpha-helix interacting with CWC22. See discussion for description.

approaches, we showed that the interaction between the two proteins is direct and mediated by the C-terminus of CWC27 and the MIF4G domain of CWC22. Our evidences strongly suggest that the CWC22/CWC27 heterodimer can be considered as a building block because it exists independently of the spliceosome: (i) CWC22 is by far the major protein enriched in CWC27 immunoprecipitations, (ii) each protein stabilizes the other and (iii) both proteins are stably integrated in the spliceosome during the transition from B to B^{act} (3,5). Thus, we propose that CWC22 and CWC27 are recruited to B^{act} spliceosomes as a heterodimer.

CWC22 had previously been proposed to bind the EJC subunit eIF4A3 and escort it to the spliceosome (9,10,38). Here, we show that a small proportion of cellular CWC27 is bound to both CWC22 and eIF4A3. We were able to identify the interacting domains and to reconstitute a ternary complex containing CWC22 MIF4G domain, CWC27 C-terminal domain and eIF4A3. We solved the 3D structure of this complex and found that contacts between the CWC22 MIF4G and the eIF4A3 RecA2 domains are kept almost identical to those previously seen in our CWC22/eIF4A3 structure (40). The CWC27 C-terminal sequence (378–402) folds into a helix that tightly binds the MIF4G domain on the side opposite to eIF4A3 RecA2 domain. This C-terminal helix is followed by a loop (402–426) that packs into a groove on the CWC22 surface. Both eIF4A3 RecA1 domain and the last C-terminal 46 amino acids of CWC27 are not present in our structure. In the absence of direct contact between CWC27 and eIF4A3, how CWC27/CWC22 stabilizes the binding of eIF4A3 remains an open question.

The cryo-EM structures of multiple splicing complexes have been solved and notably the human pre-B, B, early B^{act}, mature B^{act}, late B^{act}, C, C*, P and ILS complexes (2). Neither CWC22, CWC27 nor eIF4A3 are found in B complexes (42,43) (Figure 6A). CWC22 and CWC27 are found in 'early' and 'mature' (3,5) B^{act} complex structures. eIF4A3 is not visible in these structures but it was identified by mass spectrometry in B^{act} spliceosomes isolated by Haselbach and colleagues (Supplementary Figure S4 in Haselbach *et al.* (3)). Noteworthy, eIF4A3 is the sole EJC subunit copurifying with the B^{act} complex. Loose interactions with spliceosome complexes might prevent its co-detection in 3D structures. While CWC27 is released before conversion of the 'mature' B^{act} into 'late' B^{act} complex (5), CWC22 remains within the spliceosome from the 'late' B^{act} to the P complex at the end of splicing (Figure 6A). In the structure of C (41), C* (37) and P (44,45) complexes, CWC22 is bound to eIF4A3 within an assembled EJC on its target RNA (Figure 6A).

Based on these data, we propose a new step-wise pathway for EJC core assembly by the splicing machinery (Figure 6B). CWC27 and CWC22 are bound together outside of the spliceosome before being integrated together in the B^{act} spliceosome in which the heterodimer forms a stable landing platform for eIF4A3. We propose that the ternary complex we found exists as part of the B^{act} complex because it is the only moment where CWC27, CWC22 and eIF4A3 are present according to the structures of different spliceosomal complexes. We suppose that in the B^{act} spliceosome eIF4A3 adopts a semi-closed conformation when bound to CWC27 and CWC22, since its open conformation is not compatible with the spliceosome and that the closed formation necessitates of the presence of the other EJC components (46,47). The structure of spliceosomes in which this entire ternary complex is visible, would tell us whether eIF4A3 directly contact both CWC27 and CWC22 and whether other spliceosome components participate to

eIF4A3 attachment. When docked on the C spliceosome our structure shows that the loop of CWC27 clashes with the Y14 EJC subunit (Figure 4C). This finding accounts for the release of CWC27 during the transition between B^{act} to B*, thus allowing the concomitant association of MAGOH/Y14 and MLN51 with eIF4A3 already bound to the spliceosome through the CWC22 MIF4G domain. The binding of CWC22 MA3 domain to mRNA 5′ exon serves to position eIF4A3 clamping to RNA around 27 nt upstream 5′-exon extremity. This pre-EJC is maintained until spliceosome disassembly after P complex (44,45) during which the release of CWC22 allows the complete folding of MLN51 around eIF4A3 and binding to RNA. This overall picture of EJC assembly still contains loopholes as we ignore when exactly MAGOH/Y14 and MLN51 are recruited. Obtaining the missing cryo-EM structure of the B* complex may also help characterize the mechanistic aspects of EJC recruitment and assembly. A complete picture of the recruitment of EJC core subunits to the spliceosome is essential to understand the mechanisms potentially modulating EJC assembly and thus EJC-dependent mRNA destiny.

Changes in gene expression generated in hTERT RPE1 cells by CWC27 KD were highly correlated to those generated by CWC22 KD. These observations strongly support a functional link between the two proteins. hTERT RPE1 cells are immortalized retinal pigmented epithelium cells that are known to contribute to immune and inflammatory responses in the eye (48). The major changes in gene expression following CWC27 KD or CWC22 KD relate to the activation of a pro-inflammatory state. Half of the most up-regulated genes correspond to interleukins that activate all categories of leukocytes as well as adhesion molecules that mediate the migration and adhesion of leukocytes (Supplementary Table S7). Conversely, TGFβ2 (Transforming Growth Factor Beta), one of the most down-regulated genes (Supplementary Table S8), is known to prevent inflammation in the eye (49). Among the most down-regulated genes, we found genes coding for actin, myosin, tropomyosin and transgelin that are associated with the actin cytoskeletal network (Supplementary Table S8). The expression of many down-regulated genes, including some of those associated with the actin cytoskeleton relies upon TGFβ (50). Down-regulation of TGFβ2 can lead to disruption of the actin networks (51). Integrity of the actin cytoskeleton contributes to mitochondrial DNA (mtDNA) maintenance (52) and knocking-out β-actin results in mitochondrial dysfunction characterized by mtDNA accumulation and aggregation of TFAM, a nuclear encoded mitochondrial transcription factor (53). Remarkably, TFAM as well as all 13 mitochondrial-encoded proteins involved in oxidative phosphorylation are down-regulated as a consequence of CWC27 KD or CWC22 KD. A decrease in mitochondrial enzymes involved in oxidative phosphorylation most likely results in decreased ATP production, leading to a reduced protein synthesis capacity and thus a general decrease in ribosomal protein transcripts. Mitochondrial dysfunction can trigger a pro-inflammatory state (54). It generates reactive oxygen species (ROS) and mtDNA activates the AIM2 inflammasome. As a result, mitochondrial dysfunction activates genes in the oxidative stress and pro-inflammatory pathways. We speculate that TGFβ2, actin cytoskeleton associated and mitochondrial-transcript down-regulation contributes to the activation of a pro-inflammatory state.

Mutations in human *CWC27* gene have been associated primarily to retinal degeneration and to a spectrum of other phenotypes with various degrees of severity, such as brachydactyly, craniofacial abnormalities, short stature, and neurological defects (8). TGFβ pathway deficiency in the retinal microglia induces inflammatory contributions to retinal degeneration (55). In particular, deficiencies in CTGF and GDF6 downstream the TGF cascade, are associated with retinal dystrophies (56,57). GDF6 is also involved in early mouse cranial development (58). A target knock-out of TFAM in mouse retinal pigment epithelia leads to retinal degeneration (59) and mutations in the MT-ATP6 gene cause the Neuropathy Ataxia Retinitis Pigmentosa (NARP) syndrome (60). Furthermore, patients with defective oxidative phosphorylation are subject to craniofacial anomalies and brachydactyly (61). Given the results of our knock-down experiments and the evidence mentioned above, we hypothesize that TGFβ2 and MT-gene down-regulation contributes to patient phenotypes associated with CWC27 deficiencies.

Spliceosomopathies are genetic disorders associated with mutations in constitutive splicing factors (8,62). Several of them share common phenotypes, including Retinitis Pigmentosa or craniofacial development defects. Our findings suggest inflammation as a possible link to the retinal degeneration associated with CWC27 deficiencies. Future work should investigate how the knock-down of spliceosomal proteins related to Retinitis Pigmentosa compared those related to craniofacial disorders impact transcriptomes. We can suppose that in some specific cell types, the transcriptome is more sensitive to constitutive splicing defects. Much remains to be done to identify precursor mRNAs which processing defects contribute to these pathologies.

DATA AVAILABILITY

Sequencing raw data has been deposited in GEO. The accession number is GSE145872. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145872.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank E. Bertrand, M.J. Moore and X. Morin for plasmids, E. Del Nery for RPE-1 cells and lab members for fruitful discussions. We thank D. Loew, F. Dingli and G. Arras of the Mass Spectrometry laboratory (Institut Curie, Paris, France) and, T. Léger, B. Morlet and C. Garcia of the Mass Spectrometry facility of Institut Jacques Monod (CNRS, Paris, France).

FUNDING

ANR differEnJCe grant [ANR-13-BSV8-0023]; ANR spEJCificity [ANR-17-CE12-0021 to H.L-H.] from the French

Agence Nationale de la Recherche; program ≪ Investissements d'Avenir ≫ launched by the French Government and implemented by ANR [ANR-10-LABX-54 MEMOLIFE and ANR-10-IDEX-0001-02 PSL* Research University to V.B. and H.L.H.]; Labex Memolife and the Foundation LNCC (Ligue Nationale Contre le Cancer to V.B.); Centre National de Recherche Scientifique, the Ecole Normale Supérieure and the Institut National de la Santé et de la Recherche Médicale, France. Funding for open access charge: CNRS.
Conflict of interest statement. None declared.

REFERENCES

1. Wahl,M.C., Will,C.L. and Luhrmann,R. (2009) The spliceosome: design principles of a dynamic RNP machine. *Cell*, **136**, 701–718.
2. Wilkinson,M.E., Charenton,C. and Nagai,K. (2019) RNA splicing by the spliceosome. *Annu. Rev. Biochem.*, doi:10.1146/annurev-biochem-091719-064225.
3. Haselbach,D., Komarov,I., Agafonov,D.E., Hartmuth,K., Graf,B., Dybkov,O., Urlaub,H., Kastner,B., Luhrmann,R. and Stark,H. (2018) Structure and conformational dynamics of the human spliceosomal B(act) complex. *Cell*, **172**, 454–464.
4. Yan,C., Wan,R., Bai,R., Huang,G. and Shi,Y. (2016) Structure of a yeast activated spliceosome at 3.5 A resolution. *Science*, **353**, 904–911.
5. Zhang,X., Yan,C., Zhan,X., Li,L., Lei,J. and Shi,Y. (2018) Structure of the human activated spliceosome in three conformational states. *Cell Res.*, **28**, 307–322.
6. Wan,R., Bai,R., Yan,C., Lei,J. and Shi,Y. (2019) Structures of the catalytically activated yeast spliceosome reveal the mechanism of branching. *Cell*, **177**, 339–351.
7. Ulrich,A. and Wahl,M.C. (2014) Structure and evolution of the spliceosomal peptidyl-prolyl cis-trans isomerase Cwc27. *Acta Crystallogr. D. Biol. Crystallogr.*, **70**, 3110–3123.
8. Xu,M., Xie,Y.A., Abouzeid,H., Gordon,C.T., Fiorentino,A., Sun,Z., Lehman,A., Osman,I.S., Dharmat,R., Riveiro-Alvarez,R. *et al.* (2017) Mutations in the spliceosome component CWC27 cause retinal degeneration with or without additional developmental anomalies. *Am. J. Hum. Genet.*, **100**, 592–604.
9. Barbosa,I., Haque,N., Fiorini,F., Barrandon,C., Tomasetto,C., Blanchette,M. and Le Hir,H. (2012) Human CWC22 escorts the helicase eIF4AIII to spliceosomes and promotes exon junction complex assembly. *Nat. Struct. Mol. Biol.*, **19**, 983–990.
10. Steckelberg,A.L., Boehm,V., Gromadzka,A.M. and Gehring,N.H. (2012) CWC22 connects pre-mRNA splicing and exon junction complex assembly. *Cell Rep.*, **2**, 454–461.
11. Hocq,R., Paternina,J., Alasseur,Q., Genovesio,A. and Le Hir,H. (2018) Monitored eCLIP: high accuracy mapping of RNA-protein interactions. *Nucleic Acids Res.*, **46**, 11553–11565.
12. Le Hir,H., Izaurralde,E., Maquat,L.E. and Moore,M.J. (2000) The spliceosome deposits multiple proteins 20–24 nucleotides upstream of mRNA exon-exon junctions. *EMBO J.*, **19**, 6860–6869.
13. Le Hir,H., Sauliere,J. and Wang,Z. (2016) The exon junction complex as a node of post-transcriptional networks. *Nat. Rev. Mol. Cell Biol.*, **17**, 41–54.
14. Leung,C.S. and Johnson,T.L. (2018) The exon junction complex: a multitasking guardian of the transcriptome. *Mol. Cell*, **72**, 799–801.
15. Merz,C., Urlaub,H., Will,C.L. and Luhrmann,R. (2007) Protein composition of human mRNPs spliced in vitro and differential requirements for mRNP protein recruitment. *RNA*, **13**, 116–128.
16. Reichert,V.L., Le Hir,H., Jurica,M.S. and Moore,M.J. (2002) 5′ exon interactions within the human spliceosome establish a framework for exon junction complex structure and assembly. *Genes Dev.*, **16**, 2778–2791.
17. Andersen,J.S., Lyon,C.E., Fox,A.H., Leung,A.K., Lam,Y.W., Steen,H., Mann,M. and Lamond,A.I. (2002) Directed proteomic analysis of the human nucleolus. *Curr. Biol.*, **12**, 1–11.
18. Tyanova,S., Temu,T. and Cox,J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.*, **11**, 2301–2319.
19. Tyanova,S. and Cox,J. (2018) Perseus: A bioinformatics platform for integrative analysis of proteomics data in cancer research. *Methods Mol. Biol.*, **1711**, 133–148.
20. Ballut,L., Marchadier,B., Baguet,A., Tomasetto,C., Seraphin,B. and Le Hir,H. (2005) The exon junction core complex is locked onto RNA by inhibition of eIF4AIII ATPase activity. *Nat. Struct. Mol. Biol.*, **12**, 861–869.
21. Kabsch,W. (2010) Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr. D. Biol. Crystallogr.*, **66**, 133–144.
22. Winn,M.D., Ballard,C.C., Cowtan,K.D., Dodson,E.J., Emsley,P., Evans,P.R., Keegan,R.M., Krissinel,E.B., Leslie,A.G., McCoy,A. *et al.* (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr. D. Biol. Crystallogr.*, **67**, 235–242.
23. Storoni,L.C., McCoy,A.J. and Read,R.J. (2004) Likelihood-enhanced fast rotation functions. *Acta Crystallogr. D. Biol. Crystallogr.*, **60**, 432–438.
24. Emsley,P., Lohkamp,B., Scott,W.G. and Cowtan,K. (2010) Features and development of Coot. *Acta Crystallogr. D. Biol. Crystallogr.*, **66**, 486–501.
25. Adams,P.D., Afonine,P.V., Bunkoczi,G., Chen,V.B., Davis,I.W., Echols,N., Headd,J.J., Hung,L.W., Kapral,G.J., Grosse-Kunstleve,R.W. *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D. Biol. Crystallogr.*, **66**, 213–221.
26. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, doi.org/10.14806/ej.17.1.200.
27. 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.
28. Ramirez,F., Ryan,D.P., Gruning,B., Bhardwaj,V., Kilpert,F., Richter,A.S., Heyne,S., Dundar,F. and Manke,T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, **44**, W160–W165.
29. Robinson,J.T., Thorvaldsdottir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.
30. Li,H.D., Funk,C.C. and Price,N.D. (2020) iREAD: a tool for intron retention detection from RNA-seq data. *BMC Genomics*, **21**, 128.
31. Christinat,Y., Pawlowski,R. and Krek,W. (2016) jSplice: a high-performance method for accurate prediction of alternative splicing events and its application to large-scale renal cancer transcriptome data. *Bioinformatics*, **32**, 2111–2119.
32. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
33. Eden,E., Lipson,D., Yogev,S. and Yakhini,Z. (2007) Discovering motifs in ranked lists of DNA sequences. *PLoS Comput. Biol.*, **3**, e39.
34. Eden,E., Navon,R., Steinfeld,I., Lipson,D. and Yakhini,Z. (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.
35. Supek,F., Bosnjak,M., Skunca,N. and Smuc,T. (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*, **6**, e21800.
36. Bertram,K., Agafonov,D.E., Liu,W.T., Dybkov,O., Will,C.L., Hartmuth,K., Urlaub,H., Kastner,B., Stark,H. and Luhrmann,R. (2017) Cryo-EM structure of a human spliceosome activated for step 2 of splicing. *Nature*, **542**, 318–323.
37. Zhang,X., Yan,C., Hang,J., Finci,L.I., Lei,J. and Shi,Y. (2017) An atomic structure of the human spliceosome. *Cell*, **169**, 918–929.
38. Alexandrov,A., Colognori,D., Shu,M.D. and Steitz,J.A. (2012) Human spliceosomal protein CWC22 plays a role in coupling splicing to exon junction complex deposition and nonsense-mediated decay. *PNAS*, **109**, 21313–21318.
39. Favaro,F.P., Alvizi,L., Zechi-Ceide,R.M., Bertola,D., Felix,T.M., de Souza,J., Raskin,S., Twigg,S.R., Weiner,A.M., Armas,P. *et al.* (2014) A noncoding expansion in EIF4A3 causes Richieri-Costa-Pereira syndrome, a craniofacial disorder associated with limb defects. *Am. J. Hum. Genet.*, **94**, 120–128.
40. Buchwald,G., Schussler,S., Basquin,C., Le Hir,H. and Conti,E. (2013) Crystal structure of the human eIF4AIII-CWC22 complex shows how a DEAD-box protein is inhibited by a MIF4G domain. *PNAS*, **110**, E4611–E4618.

41. Zhan,X., Yan,C., Zhang,X., Lei,J. and Shi,Y. (2018) Structure of a human catalytic step I spliceosome. *Science*, **359**, 537–545.
42. Bertram,K., Agafonov,D.E., Dybkov,O., Haselbach,D., Leelaram,M.N., Will,C.L., Urlaub,H., Kastner,B., Luhrmann,R. and Stark,H. (2017) Cryo-EM structure of a pre-catalytic human spliceosome primed for activation. *Cell*, **170**, 701–713.
43. Zhan,X., Yan,C., Zhang,X., Lei,J. and Shi,Y. (2018) Structures of the human pre-catalytic spliceosome and its precursor spliceosome. *Cell Res.*, **28**, 1129–1140.
44. Fica,S.M., Oubridge,C., Wilkinson,M.E., Newman,A.J. and Nagai,K. (2019) A human postcatalytic spliceosome structure reveals essential roles of metazoan factors for exon ligation. *Science*, **363**, 710–714.
45. Zhang,X., Zhan,X., Yan,C., Zhang,W., Liu,D., Lei,J. and Shi,Y. (2019) Structures of the human spliceosomes before and after release of the ligated exon. *Cell Res.*, **29**, 274–285.
46. Andersen,C.B., Ballut,L., Johansen,J.S., Chamieh,H., Nielsen,K.H., Oliveira,C.L., Pedersen,J.S., Seraphin,B., Le Hir,H. and Andersen,G.R. (2006) Structure of the exon junction core complex with a trapped DEAD-box ATPase bound to RNA. *Science*, **313**, 1968–1972.
47. Bono,F., Ebert,J., Lorentzen,E. and Conti,E. (2006) The crystal structure of the exon junction complex reveals how it maintains a stable grip on mRNA. *Cell*, **126**, 713–725.
48. Holtkamp,G.M., Kijlstra,A., Peek,R. and de Vos,A.F. (2001) Retinal pigment epithelium-immune system interactions: cytokine production and cytokine-induced changes. *Prog. Retin. Eye Res.*, **20**, 29–48.
49. Zamiri,P., Sugita,S. and Streilein,J.W. (2007) Immunosuppressive properties of the pigmented epithelial cells and the subretinal space. *Chem. Immunol. Allergy*, **92**, 86–93.
50. Kubo,E., Shibata,S., Shibata,T., Kiyokawa,E., Sasaki,H. and Singh,D.P. (2017) FGF2 antagonizes aberrant TGFbeta regulation of tropomyosin: role for posterior capsule opacity. *J. Cell. Mol. Med.*, **21**, 916–928.
51. Montecchi-Palmer,M., Bermudez,J.Y., Webber,H.C., Patel,G.C., Clark,A.F. and Mao,W. (2017) TGFbeta2 induces the formation of cross-linked actin networks (CLANs) in human trabecular meshwork cells through the smad and non-smad dependent pathways. *Invest. Ophthalmol. Vis. Sci.*, **58**, 1288–1295.
52. Reyes,A., He,J., Mao,C.C., Bailey,L.J., Di Re,M., Sembongi,H., Kazak,L., Dzionek,K., Holmes,J.B., Cluett,T.J. *et al.* (2011) Actin and myosin contribute to mammalian mitochondrial DNA maintenance. *Nucleic Acids Res.*, **39**, 5098–5108.
53. Kang,I., Chu,C.T. and Kaufman,B.A. (2018) The mitochondrial transcription factor TFAM in neurodegeneration: emerging evidence and mechanisms. *FEBS Lett.*, **592**, 793–811.
54. West,A.P. (2017) Mitochondrial dysfunction as a trigger of innate immune responses and inflammation. *Toxicology*, **391**, 54–63.
55. Ma,W., Silverman,S.M., Zhao,L., Villasmil,R., Campos,M.M., Amaral,J. and Wong,W.T. (2019) Absence of TGFbeta signaling in retinal microglia induces retinal degeneration and exacerbates choroidal neovascularization. *Elife*, **8**, e42049.
56. Asai-Coakwell,M., March,L., Dai,X.H., Duval,M., Lopez,I., French,C.R., Famulski,J., De Baere,E., Francis,P.J., Sundaresan,P. *et al.* (2013) Contribution of growth differentiation factor 6-dependent cell survival to early-onset retinal dystrophies. *Hum. Mol. Genet.*, **22**, 1432–1442.
57. Ma,T., Dong,L.J., Du,X.L., Niu,R. and Hu,B.J. (2018) Research progress on the role of connective tissue growth factor in fibrosis of diabetic retinopathy. *Int. J. Ophthalmol.*, **11**, 1550–1554.
58. Clendenning,D.E. and Mortlock,D.P. (2012) The BMP ligand Gdf6 prevents differentiation of coronal suture mesenchyme in early cranial development. *PLoS One*, **7**, e36789.
59. Zhao,C., Yasumura,D., Li,X., Matthes,M., Lloyd,M., Nielsen,G., Ahern,K., Snyder,M., Bok,D., Dunaief,J.L. *et al.* (2011) mTOR-mediated dedifferentiation of the retinal pigment epithelium initiates photoreceptor degeneration in mice. *J. Clin. Invest.*, **121**, 369–383.
60. Lefevere,E., Toft-Kehler,A.K., Vohra,R., Kolko,M., Moons,L. and Van Hove,I. (2017) Mitochondrial dysfunction underlying outer retinal diseases. *Mitochondrion*, **36**, 66–76.
61. Cormier-Daire,V., Rustin,P., Rotig,A., Chretien,D., Le Merrer,M., Belli,D., Le Goff,A., Hubert,P., Ricour,C. and Munnich,A. (1996) Craniofacial anomalies and malformations in respiratory chain deficiency. *Am. J. Med. Genet.*, **66**, 457–463.
62. Lehalle,D., Wieczorek,D., Zechi-Ceide,R.M., Passos-Bueno,M.R., Lyonnet,S., Amiel,J. and Gordon,C.T. (2015) A review of craniofacial disorders caused by spliceosomal defects. *Clin. Genet.*, **88**, 405–415.

Chapter 10

Résumé détaillé en français

10.1 Introduction

Le dogme central de la biologie repose sur trois piliers : l'ADN comme l'entrepôt de l'information génétique ; l'ARN messager (ARNm) comme moyen de transmission de cet information, et la protéine comme la machinerie exécutant les fonctions vitales de la cellule. Décrite ainsi, la synthèse des protéines à partir de l'ARN messager, qui est lui-même produit à partir de l'ADN, paraît un processus simple. Or, l'expression génique est l'aboutissement d'un réseau complexe d'interactions moléculaires extrêmement régulées.

La régulation post-transcriptionnelle de l'expression des gènes (PTGR) est un mécanisme cellulaire qui contrôle l'expression génique au niveau de l'ARNm à plusieurs étapes : la transformation (épissage, modification chimique, rognage), l'export nucléaire, la localisation subcellulaire, la traduction, et la dégradation. Plus de 1500 protéines de liaison à l'ARN (RBPs) sont orchestrées par la cellule afin de mettre en place le réseau de la PTGR. Les RBPs sont des protéines qui interagissent directement ou indirectement avec l'ARN. Ces interactions sont fondamentales pour une PTGR effective, ce qui permet l'homéostasie, la différentiation cellulaire, et le développement embryonnaire. La défaillance de ce système de régulation cause des multiples pathologies (Castello et al. 2013). Malgré leur importance, les fonctions cellulaires de plusieurs RBPs n'ont pas été caractérisées.

Le complexe EJC (*Exon Junction Complex*) est un complexe comportant quatre RBPs centrales (eIF4A3, MAGOH, Y14, and CASC3) et des multiples facteurs périphériques (Le Hir, Saulière, and Z. Wang 2016). L'EJC est déposé pendant l'épissage environ 24 nucléotides en amont de la jonction exonique (Le Hir, Izaurralde, et al. 2000). Ses particularités structurales lui permettent d'agripper l'ARN de manière stable indépendamment de sa séquence (Bono, Ebert, et al. 2006). Il accompagne, donc, les ARNm vers le cytoplasme et intervient à plusieurs étapes de leur vie via l'interaction avec de nombreux facteurs. Sa versatilité lui confère une grande importance dans le réseau de la PTGR. Ceci est illustré par des syndromes morphologiques et neurologiques causés par des mutations de certains de ses composants (Favaro et al. 2014; Albers et al. 2012), ce qui indique que l'EJC est essentiel pendant le développement

embryonnaire et la neurogénèse. Malgré son rôle central dans la régulation de l'expression des gènes, de nombreuses questions sur l'impact fonctionnel de l'EJC restent à résoudre.

10.2 Problématique

La polyvalence de l'EJC entrave l'étude de ses multiples fonctions. Les connaissances actuelles sur son assemblage et ses rôles ont été élucidées grâce à des techniques de biologie moléculaire limitées à un faible débit. Ceci restreint l'étude de l'impact de la liaison de l'EJC sur la régulation spécifique des gènes. En effet, ces approches ignorent son rôle précis à différentes étapes de la vie de différents transcrits. Par exemple, il est incertain si la présence de l'EJC sur tous les exons d'un ARNm est nécessaire pour une correcte expression du gène. Au contraire, dans le cas où l'assemblage de l'EJC est ciblé sur des exons spécifiques, au sein de gènes spécifiques, les facteurs qui déterminent son dépôt restent inconnus. Les modalités de liaison de l'EJC, ainsi que son impact fonctionnel, ne peuvent être découvertes qu'avec une carte des sites de liaison à l'échelle du transcriptome.

Les méthodes de CLIP (Cross-Linking and Immunoprécipitation) associée au séquençage à haut-débit (CLIP-seq) visent à découvrir les sites de liaison d'une RBP à l'échelle du transcriptome. Cette méthode consiste à induire des liaisons covalentes entre les protéines et l'ARN interagissant in vivo avec de la lumière UV, ce qui permet d'effectuer une immunoprécipitation (IP) sous des conditions sévères. Pendant l'IP, les anticorps contre la RBP d'intérêt purifient la protéine ainsi que les fragments d'ARN qui y sont liés. L'ensemble de ces fragments constitue une banque. Le séquençage de la banque, suivi par l'alignement des lectures sur le génome, entraîne des enrichissements locaux de signal, qui correspondent aux sites de liaison de la RBP. Cependant, l'efficacité du photopontage et de l'IP limitent considérablement l'obtention de bibliothèques de fragments d'ARN qui représentent tous les sites de liaison spécifiques de la RBP. Par ailleurs, l'analyse des données CLIP-seq est une tâche difficile en termes de sensibilité, de spécificité et de reproductibilité.

Au fil des années, plusieurs tentatives ont été faites pour établir la carte de liaison de l'EJC. Les résultats suggèrent que l'EJC n'est pas assemblé de manière homogène sur tous les exons d'un ARNm, mais de manière spécifique sur quelques exons. Cependant, ces études ont généralement contourné la reproductibilité de la détection des sites. Au fur et à mesure que les méthodes de CLIP ont gagné en résolution, la reproductibilité des différents sites de liaison est devenu plus difficile à évaluer. Ainsi, malgré les avancées expérimentales, une carte à haute résolution, et suffisamment reproductible, à l'échelle du transcriptome des sites de liaison de l'EJC n'a pas encore été établi.

10.3 Résultats

10.3.1 Obtention de données CLIP à haute résolution

L'obtention d'une banque de sites de liaison d'une RBP requiert plusieurs étapes expérimentales. Après l'IP, les protéines liées sont dégradées avec la protéinase K, laissant un court peptide lié à des fragments d'ARN. Ces fragments sont transcrits en ADN complémentaire (ADNc) par une rétrotranscriptase (RTase) dans le sens de 3' $\rightarrow$ 5'. Souvent, le peptide, encore accroché à l'ARN, entraîne l'arrêt de la RTase au niveau du site de liaison protéine-ARN. Cela permet d'utiliser l'extrémité 5' des lectures issues du séquençage pour cartographier les sites de liaison au nucléotide près. Cependant, la RTase est capable de dépasser le peptide et de transcrire le fragment ARN entier. Étant donné que la longueur des fragments peut être supérieure à 100 nucléotides, l'utilisation de l'extrémité 5' de ces lectures, dites read-through, introduit un biais dans l'attribution des sites de liaison de la RBP en question.

Pour détecter ces événements, un oligonucléotide de 13 nt de longueur (linker), dont la séquence est connue, est ligaturé à l'extrémité 5' des fragments d'ARN, en amont du peptide résiduel. Ainsi, les ADNc (et les lectures qu'ils génèrent) incluant la séquence du linker correspondent aux événements de read-through. La séparation de ces lectures après le séquençage permet une plus grande précision dans l'attribution des sites de liaison. L'addition de cet étape au protocol d'eCLIP (Eric L. Van Nostrand et al. 2016) permet de surveiller la proportion de read-through et d'augmenter la précision des site des liaison. Nous avons donc nommé cette stratégie meCLIP (monitored eCLIP).

10.3.2 Le prétraitement de données

Les données issues du séquençage à haut débit doivent être mises au point, ou prétraitées, avant d'effectuer l'analyse en soi. En général, cela consiste à vérifier la qualité des données, à rogner les bases de mauvaise qualité et les adaptateurs de séquençage, à éliminer les doublons de PCR, et à aligner les lectures sur le génome. Pour détecter les doublons de PCR dans le CLIP, un identificateur moléculaire unique (UMI) est ligaturé avant l'amplification de l'ADNc par PCR. Dans notre stratégie de prétraitement, les doublons de PCR étaient éliminés avant la l'alignement sur le génome. Les lectures avec un UMI et une séquence d'ADNc identiques étaient marquées comme des doublons et écartées. Cependant, les doublons de PCR avec des erreurs de séquençage contournaient ce filtre, ce qui entraînait une sous-estimation de la duplication PCR et un signal biaisé. Afin de surmonter ce problème, nous avons intégré l'utilisation de UMI-tools dans notre stratégie de prétraitement (T. Smith, Heger, and Sudbery 2017). Ce programme évalue uniquement les UMI des lectures ayant des coordonnées génomiques identiques ; l'élimination de doublons de PCR s'effectue donc après l'alignement sur le génome. En autorisant quelques discordances au sein de ces séquences—potentiellement causées par des erreurs de séquençage—UMI-tools estime avec une meilleure précision les doublons de PCR.

Nous avions généré 8 banques meCLIP de la protéine eIF4A3 de l'EJC. Or, après la déduplication de PCR avec UMI-tool, ces jeux de données ont fini par avoir

une faible couverture. Nous avons donc décidé de les fusionner en deux pseudo-jeux de données. En parallèle, nous avons travaillé pour obtenir davantage de jeux de données CLIP pour l'EJC. Pour accélérer la production de banques, nous avons adopté une stratégie de pré-séquençage suivie d'une évaluation de la complexité. De cette manière, nous avons réussi à optimiser la génération de données CLIP et obtenir deux banques additionnelles : eCLIP1 et eCLIP2, qui ont chacune été séquencées deux fois. Comme eCLIP2 s'est avérée plus complexe que eCLIP1, il a généré deux grands jeux de données (eCLIP2-1 et eCLIP2-2). Afin d'obtenir des résultats plus comparables, nous avons décidé de réduire aléatoirement eCLIP2-1 et eCLIP2-2 au nombre exact de lectures de eCLIP1-1 et eCLIP1-2, respectivement. Ceci a été effectué deux fois pour chaque eCLIP2, ce qui a donné eCLIP2-S1, -S2, -S3 et -S4. Au total, nous avons généré huit jeux de données CLIP eIF4A3 pour les analyses ultérieures.

10.3.3 La détection de sites liaison et sa reproductibilité

Le signal CLIP dans les jeux de données de l'EJC peut être agrégé et représenté autour de la jonction exonique. Cette représentation est appelée méta-exon, et elle offre une notion globale et qualitative du positionnement du complexe. On a observé un fort enrichissement de signal autour d'une fenêtre étroite autour du 27e nucléotide en amont de la jonction exonique. Cette région correspond au site de liaison canonique de l'EJC, tel qu'elle a été définie lors de la découverte du complexe. Concrètement, nous avons défini la région canonique des exons comme la section de 11 nt autour du 27e nucléotide à partir de l'extrémité 3' (entre le 32e et le 22e nucléotides).

Parmi les outils de détection de sites de liaison—également appelée détection de pics— disponibles pour CLIP, CITS (Shah et al. 2017) et PureCLIP (Krakau, Richard, and Marsico 2017) sont capables de détecter des sites de liaison au nucléotide près. Nous avons donc effectué la détection de pics sur les jeux de données CLIP de l'EJC avec ces deux outils, afin d'obtenir une carte de sites de liaison à haute résolution. Les deux algorithmes détectent environ 1000 pics à l'intérieur de la région canonique, malgré le fort enrichissement observé dans le méta-exon. Cette contradiction suggère que ces outils fournissent des résultats avec une faible sensibilité. Par la suite, nous avons utilisé les pics PureCLIP pour les analyses ultérieures puisqu'ils avaient un signal plus fort que les pics CITS (un plus grand nombre de lectures au sein des pics), indiquant une spécificité plus élevée.

Ensuite, nous avons évalué la reproductibilité des résultats en suivant la recommandation du consortium ENCODE pour la détection des pics. La valeur IDR (Irreproducible Discovery Rate) est la probabilité qu'une paire de pics détectés au même endroit dans deux échantillons différents soient irreproductible. Les pics ayant une valeur IDR inférieure à 0,05 peuvent être considérés comme significativement reproductibles. Or, nous avons trouvé un faible nombre de pics reproductibles parmi nos jeux de données (rarement supérieur à 50). Étant donné que la valeur IDR est calculée sur les pics détectés dans les deux échantillons, nous avons calculé l'indice Jaccard des pics (c.f. Équation 10.1). Cet indice correspond à la fraction des éléments

communs entre deux ensembles, ce qui permet de quantifier la reproductibilité de la détection des pics. Dans ce contexte, l'indice Jaccard des pics PureCLIP de nos ensembles de données CLIP se situait entre 0,03 et 0,19. Cela indiquait que le nombre de pics communs entre les échantillons était faible.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{10.1}$$

Equation 10.1: L'indice de Jaccard (J) entre deux ensembles (A et B) correspond au rapport entre la taille de l'intersection ($|A \cap B|$) sur la taille de l'union ($|A \cup B|$).

Afin de comparer la reproductibilité de nos données de CLIP de l'EJC, nous avons effectué une détection de pics à grande échelle sur les données eCLIP de 72 RBP différentes du consortium ENCODE. Après avoir détecté des pics sur deux ensembles de données différents de chaque protéine, nous avons calculé l'indice Jaccard des pics couramment détectés. Nous avons obtenu des valeurs situées entre 0,01 et 0,40, avec une valeur médiane de 0,18. Par conséquent, nous avons constaté que la reproductibilité des pics à haute résolution était faible de manière globale pour les données CLIP.

Nous avons donc été confrontés à un dilemme concernant l'analyse des données CLIP. D'une part, les représentations types méta-exon de l'EJC montrent un enrichissement reproductible du signal au niveau du 27ème nucléotide en amont de la jonction. Cependant , elles ne donnent pas d'informations sur les différents sites de liaison du complexe. En revanche, la détection des pics a le potentiel d'identifier les cibles de l'EJC et leur position précise. Pourtant, la reproductibilité de la détection de pics s'est avérée faible. Nous avons donc exploré des alternatives afin d'exploiter les données de l'EJC en donnant la priorité à la reproductibilité des résultats. Une méthode pour exploiter le signal spécifique de l'EJC Comme nous l'avons montré jusqu'à présent, la détection de sites de liaison à haute résolution a un faible taux de reproductibilité. Pour surmonter cette limitation, nous proposons un protocol d'analyse à la fois spécifique à l'EJC et centré sur la reproductibilité. Notre approche a donné de meilleurs résultats sur nos données CLIP que les outils conventionnels plus généralistes.

Afin de détecter les exons enrichis en signal de l'EJC, nous avons calculé le score d'enrichissement EJC (EES). Comme nous l'avons observé dans la représentation type méta-exons, les nucléotides autour du 27e nucléotide en amont de la jonctions présentent un enrichissement élevé alors que d'autres régions présentent un signal plat correspondant au bruit de fond. Par conséquent, nous avons défini deux régions au sein de chaque exon : la région canonique, une fenêtre de 11 nucléotides autour du 27e nucléotide, du 32e au 22e nucléotide en amont de la jonction exonique, et la région non canonique, qui couvre une fenêtre de 11 nucléotides entre le 15e et le 5e nucléotides en amont de la jonction. Nous avons ensuite compté les extrémités 5' des lectures situées dans chaque région, profitant ainsi de la haute résolution des données. La valeur EES correspond au rapport entre les lectures de la région canonique et les lectures de la région non canonique. Si le signal de la région canonique est 2 fois supérieure à celui de la région canonique (EES ¿ 2), nous considérons un exon comme enrichi en EJC.

Nous avons calculé les valeurs d'EES sur l'ensembles de nos données CLIP, ainsi que dans les contrôles input et les données RNA-seq. Nous avons d'abord constaté que le nombre d'exons enrichis était de 3,5 à 7 fois plus élevé dans les données CLIP que dans l'input, ce qui indique une grande spécificité de la stratégie EES. Nous avons également observé que le nombre d'exons enrichis était environ 4 fois plus élevé que le nombre d'exons avec des pics PureCLIP dans la région canonique. Cela suggère que la stratégie EES a une sensibilité plus élevée que la méthode PureCLIP. En outre, bien que le nombre d'exons enrichis dans les données CLIP n'ait été que 1,3 à 1,7 fois supérieur par rapport aux données RNA-seq, la distribution d'EES a montré des valeurs significativement plus élevées dans le premier. Cela indique que l'≪ enrichissement ≫ apparent dans les données RNA-seq est plus faible que dans les données CLIP de l'EJC. Ces résultats montrent que la méthode d'EES a à la fois une spécificité et une sensibilité élevées pour détecter les exons enrichis en signal de l'EJC.

Pour comparer davantage les résultats de PureCLIP, nous avons évalué la reproductibilité des exons détectés avec la méthode EES. Nous avons donc calculé l'indice Jaccard des exons communément détectés dans chaque paire possible de jeux de données CLIP. Nous avons trouvé une valeur Jaccard moyenne de 0,28 dans les résultats de la méthode EES, ce qui était significativement plus élevé que la valeur moyenne de 0,24 des résultats PureCLIP (P ¡ 10-4). Il est à noter que ces dernières valeurs correspondent à des exons avec au moins un pic de PureCLIP dans la région canonique. Bien que la reproductibilité des exons détectés soit plus élevée que celle des pics PureCLIP (c.f. la section précédente), plus de 70% des exons détectés par la méthode EES se trouvent dans un seul réplicat et non dans l'autre. Pour cette raison, nous avons exploré un autre niveau de mesure du signal EJC où la reproductibilité serait plus élevée.

Nous avons quantifié le nombre d'exons enrichis par gène, et avons désigné cette valeur comme la Loaded Fraction (LF). Premièrement, nous avons trouvé un indice Jaccard élevé des gènes avec une LF ¿ 0 (d'environ 0.85). Deuxièmement, afin de trouver des gènes avec des LF similaires, nous avons calculé le ratio de LF entre deux réplicats et sélectionné les gènes dont les ratios issus de toutes les comparaisons entre réplicats étaient compris entre 0,66 et 1,5. Nous avons ainsi obtenu une liste de 151 gènes dont le taux d'occupation était reproductible (reproducibly loaded genes RLG).

10.4 Discussion

La création de bibliothèques CLIP-seq est une tâche difficile, surtout lorsqu'il s'agit de détecter des sites de reliure reproductibles. D'après les expériences de HeLa RNA-seq, nous avons estimé à environ 10 000 gènes exprimés (¿2,3 RPKM), ce qui se traduit par plus de 118 000 gènes exprimés exons. Les ensembles de données eCLIP eIF4A3 que nous avons obtenus avec succès détectent entre 5.000 à 8.000 exons enrichis par l'EJC. Cela correspond à un taux de détection de 4 à 7%, qui semble relativement faible malgré le signal minier provenant de la région canonique.

Cela peut s'expliquer par le mécanisme contraignant de l'EJC. Comme indiqué dans section 2.2.2, le eIF4A3 adopte une conformation fermée sur l'ARN et interagit avec le ribose-phosphate plutôt qu'avec les bases nucléotidiques. Ainsi, il est probablement plus difficile de capturer une quantité considérable d'interactions EJC-ARN, car l'irradiation par la lumière UV est plus susceptible de créer des liens entre les résidus d'ARN. En conséquence, nos résultats suggèrent que le contenu en uracile favorise des exons robustes chargés en EJC. Là encore, l'irradiation aux UV est plus susceptible de créer des liaisons covalentes entre les pyrimidines et les résidus (principalement) aromatiques. Ainsi, combinée à l'interaction entre l'eIF4A3 et l'ARN, la chimie de la réticulation UV peut expliquer la faible sensibilité globale de la détection du site de liaison.

Pourtant, eIF4A3 n'est pas la seule protéine EJC qui entre en contact avec l'ARN. Comme CASC3 directement entre en contact avec la base d'un nucléotide (Andersen et al. 2006), on peut considérer qu'il s'agit d'une candidat CLIP plus efficace que le eIF4A3. Toutefois, des rapports récents suggèrent que Le CASC3 peut ne pas être un élément constitutif de base sur tous les gènes, avec notamment une implication fonctionnelle dans la sensibilité à la NMD Mabin et al. 2018. Ainsi, le CLIPping CASC3 pourrait révéler une sous-population de EJC avec des rôles spécifiques, plutôt qu'un panorama plus large de la régulation par les EJC.

Dans ces travaux de thèse, nous avons mis en place une stratégie de prétraitement qui accélère l'obtention de la bibliothèque CLIP de l'EJC. Nous estimons maintenant la complexité et la spécificité des signaux des bibliothèques dans un pré-séquençage. Cela nous permet de nous concentrer sur les ensembles de données susceptibles de représenter la majorité des exons chargés par le EJC, tout en éliminant les bibliothèques CLIP de mauvaise qualité ou ayant échoué au début du processus. C'est donc une question de temps avant que nous obtenions de meilleures bibliothèques CLIP de l'EJC qui révèlent un plus grand nombre de gènes détectés avec robustesse, que ce soit chez l'homme, la drosophile ou d'autres espèces.

10.5 Conclusion

Les techniques CLIP-seq sont une approche permettant de découvrir les modalités de liaison des RBP à l'échelle du transcriptome. Cependant, elles représentent un défi sur le plan expérimental et de l'analyse des données.

Les protocoles CLIP comportent de multiples étapes qui s'étendent sur plusieurs jours. L'obtention d'une bibliothèque de haute qualité une fois que toutes les étapes ont été réalisées dépend de la nature de la protéine d'intérêt et de nombreux facteurs. Cela nécessite d'investir du temps et de l'énergie dans l'optimisation de nombreuses conditions expérimentales, spécifiquement pour la RBP ciblée. Ici, nous avons mis en œuvre une stratégie de prétraitement des données qui a accéléré le processus d'optimisation pour eIF4A3. Cependant, il reste beaucoup de travail à faire pour optimiser CLIP pour d'autres protéines et dans d'autres organismes.

Les résultats expérimentaux ont un impact direct sur la qualité des données à analyser. Cependant, quelle que soit la qualité, l'analyse des données n'est pas simple. Les principaux La limitation dans la détection des sites de liaison d'un seul nucléotide est la reproductibilité. Nous avons montré que les sites de liaison détectés avec les pics d'appel CLIP ont une reproductibilité limitée. Nous avons donc développé une stratégie spécifique pour exploiter les données CLIP de l'EJC de manière reproductible, qui peut être extrapolée à d'autres RBP avec des modalités de liaison similaires.

Notre méthode nous a permis de réévaluer avec une grande confiance le taux de chargement de l'EJC sur les transcriptions humaines. Nous avons constaté une présence de EJC plus faible qu'auparavant estimé. L'absence d'EJC à certaines jonctions soulève des questions sur la la régulation différentielle des événements d'épissage spécifiques. Des analyses plus approfondies pourraient révéler les facteurs qui sous-tendent la charge et la détection des exons chargés par l'EJC, tant chez les humains et la drosophile, ainsi que ses implications fonctionnelles dans la régulation de l'expression des gènes.

List of Figures

Bibliography

Alachkar, A. et al. (June 2013). *An EJC Factor RBM8a Regulates Anxiety Behaviors*. en. Issue: 6 Pages: 887-899 Volume: 13. URL: https://www.eurekaselect.com/111636/article (visited on 08/31/2020).

Albers, Cornelis A. et al. (Feb. 2012). "Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome". eng. In: *Nature Genetics* 44.4, 435–439, S1–2. ISSN: 1546-1718. DOI: 10.1038/ng.1083.

Amrani, Nadia et al. (Nov. 2004). "A faux 3-UTR promotes aberrant termination and triggers nonsense- mediated mRNA decay". en. In: *Nature* 432.7013. Number: 7013 Publisher: Nature Publishing Group, pp. 112–118. ISSN: 1476-4687. DOI: 10.1038/nature03060. URL: https://www.nature.com/articles/nature03060 (visited on 09/21/2020).

Andersen, Christian B. F. et al. (Sept. 2006). "Structure of the Exon Junction Core Complex with a Trapped DEAD-Box ATPase Bound to RNA". en. In: *Science* 313.5795. Publisher: American Association for the Advancement of Science Section: Report, pp. 1968–1972. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1131981. URL: https://science-sciencemag-org.insb.bib.cnrs.fr/content/313/5795/1968 (visited on 03/18/2020).

Andrews, S. (2010). *Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data*. URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (visited on 04/15/2020).

Änkö, Minna-Liisa et al. (2012). "The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes". In: *Genome Biology* 13.3. ISBN: 1465-6914 (Electronic) 1465-6906 (Linking), R17. ISSN: 1465-6906. DOI: 10.1186/gb-2012-13-3-r17. URL: http://genomebiology.com/2012/13/3/R17.

Ashburner, Michael et al. (May 2000). "Gene Ontology: tool for the unification of biology". In: *Nature genetics* 25.1, pp. 25–29. ISSN: 1061-4036. DOI: 10.1038/75556. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3037419/ (visited on 06/28/2020).

Ashton-Beaucage, Dariel et al. (Oct. 2010). "The Exon Junction Complex Controls the Splicing of mapk and Other Long Intron-Containing Transcripts in Drosophila". In: *Cell* 143.2, pp. 251–262. ISSN: 0092-8674. DOI: 10.1016/j.cell.2010.09.014. URL: http://www.sciencedirect.com/science/article/pii/S0092867410010627 (visited on 04/24/2019).

Bailey, Timothy L. (June 2011). "DREME: motif discovery in transcription factor ChIP-seq data". en. In: *Bioinformatics* 27.12. Publisher: Oxford Academic, pp. 1653–1659. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr261. URL: https://academic.oup.com/bioinformatics/article/27/12/1653/257754 (visited on 06/10/2020).

Ballut, Lionel et al. (Oct. 2005). "The exon junction core complex is locked onto RNA by inhibition of eIF4AIII ATPase activity". eng. In: *Nature Structural & Molecular Biology* 12.10, pp. 861–869. ISSN: 1545-9993. DOI: 10.1038/nsmb990.

Barbosa, Isabelle et al. (Oct. 2012). "Human CWC22 escorts the helicase eIF4AIII to spliceosomes and promotes exon junction complex assembly". en. In: *Nature Structural & Molecular Biology* 19.10. Number: 10 Publisher: Nature Publishing Group, pp. 983–990. ISSN: 1545-9985. DOI: 10.1038/nsmb.2380. URL: http://www.nature.com/articles/nsmb.2380 (visited on 06/19/2020).

Bartkowska, Katarzyna et al. (2018). "Roles of the exon junction complex components in the central nervous system: a mini review". In: *Reviews in the Neurosciences* 29.8, pp. 817–824. ISSN: 0334-1763. DOI: 10.1515/revneuro-2017-0113. URL: http://www.degruyter.com/view/j/revneuro.2018.29.issue-8/revneuro-2017-0113/revneuro-2017-0113.xml (visited on 03/18/2020).

Blazquez, Lorea et al. (Nov. 2018). "Exon Junction Complex Shapes the Transcriptome by Repressing Recursive Splicing". In: *Molecular Cell* 72.3, 496–509.e9. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2018.09.033. URL: http://www.sciencedirect.com/science/article/pii/S1097276518308323 (visited on 01/18/2019).

Bobola, Nicoletta et al. (Mar. 1996). "Asymmetric Accumulation of Ash1p in Postanaphase Nuclei Depends on a Myosin and Restricts Yeast Mating-Type Switching to Mother Cells". English. In: *Cell* 84.5. Publisher: Elsevier, pp. 699–709. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/S0092-8674(00)81048-X. URL: https://www.cell.com/cell/abstract/S0092-8674(00)81048-X (visited on 07/29/2020).

Boehm, Volker et al. (Nov. 2018). "Exon Junction Complexes Suppress Spurious Splice Sites to Safeguard Transcriptome Integrity". In: *Molecular Cell* 72.3, 482–495.e7. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2018.08.030. URL: http://www.sciencedirect.com/science/article/pii/S1097276518306907 (visited on 01/18/2019).

Böhl, Florian et al. (Oct. 2000). "She2p, a novel RNA-binding protein tethers ASH1 mRNA to the Myo4p myosin motor via She3p". In: *The EMBO Journal* 19.20. Publisher: John Wiley & Sons, Ltd, pp. 5514–5524. ISSN: 0261-4189. DOI: 10.1093/emboj/19.20.5514. URL: https://www.embopress.org/doi/10.1093/emboj/19.20.5514 (visited on 07/29/2020).

Bono, Fulvia, Atlanta G. Cook, et al. (Jan. 2010). "Nuclear Import Mechanism of the EJC Component Mago-Y14 Revealed by Structural Studies of Importin 13". en. In: *Molecular Cell* 37.2, pp. 211–222. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2010.01.007. URL: http://www.sciencedirect.com/science/article/pii/S1097276510000328 (visited on 06/20/2020).

Bono, Fulvia, Judith Ebert, et al. (Aug. 2006). "The crystal structure of the exon junction complex reveals how it maintains a stable grip on mRNA". eng. In: *Cell* 126.4, pp. 713–725. ISSN: 0092-8674. DOI: 10.1016/j.cell.2006.08.006.

Bottini, Silvia, Nedra Hamouda-Tekaya, et al. (2017). "From benchmarking HITS-CLIP peak detection programs to a new method for identification of miRNA-binding sites from Ago2-CLIP data". In: *Nucleic Acids Research* 45.9, e71. ISSN: 13624962. DOI: 10.1093/nar/gkx007.

Bottini, Silvia, David Pratella, et al. (2018). "Recent computational developments on CLIP-seq data analysis and microRNA targeting implications". eng. In: *Briefings in Bioinformatics* 19.6, pp. 1290–1301. ISSN: 1477-4054. DOI: 10.1093/bib/bbx063.

Brennan, Christopher M., Imed-Eddine Gallouzi, and Joan A. Steitz (Oct. 2000). "Protein Ligands to Hur Modulate Its Interaction with Target Mrnas in Vivo". en. In: *Journal of Cell Biology* 151.1. Publisher: The Rockefeller University Press, pp. 1–14. ISSN: 0021-9525. DOI: 10.1083/jcb.151.1.1. URL: https://rupress.org/jcb/article/151/1/1/54257/Protein-Ligands-to-Hur-Modulate-Its-Interaction (visited on 07/29/2020).

Brogna, Saverio and Jikai Wen (Feb. 2009). "Nonsense-mediated mRNA decay (NMD) mechanisms". en. In: *Nature Structural & Molecular Biology* 16.2. Number: 2 Publisher: Nature Publishing Group, pp. 107–113. ISSN: 1545-9985. DOI: 10.1038/nsmb.1550. URL: https://www.nature.com/articles/nsmb.1550 (visited on 09/21/2020).

Brown, James B. et al. (Aug. 2014). "Diversity and dynamics of the *Drosophila* transcriptome". en. In: *Nature* 512.7515, pp. 393–399. ISSN: 1476-4687. DOI: 10.1038/nature12962. URL: https://www.nature.com/articles/nature12962 (visited on 07/17/2019).

Buchwald, Gretel et al. (Nov. 2013). "Crystal structure of the human eIF4AIII–CWC22 complex shows how a DEAD-box protein is inhibited by a MIF4G domain". en. In: *Proceedings of the National Academy of Sciences* 110.48. Publisher: National Academy of Sciences Section: PNAS Plus, E4611–E4618. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1314684110. URL: https://www.pnas.org/content/110/48/E4611 (visited on 06/19/2020).

Bühler, Marc et al. (May 2006). "EJC-independent degradation of nonsense immunoglobulin- mRNA depends on 3 UTR length". en. In: *Nature Structural & Molecular Biology* 13.5. Number: 5 Publisher: Nature Publishing Group, pp. 462–464. ISSN: 1545-9985. DOI: 10.1038/nsmb1081. URL: https://www.nature.com/articles/nsmb1081 (visited on 09/21/2020).

Busetto, Virginia et al. (June 2020). "Structural and functional insights into CWC27/CWC22 heterodimer linking the exon junction complex to spliceosomes". en. In: *Nucleic Acids Research* 48.10. Publisher: Oxford Academic, pp. 5670–5683. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa267. URL: https://academic.oup.com/nar/article/48/10/5670/5824608 (visited on 06/07/2020).

Callis, J., M. Fromm, and V. Walbot (Dec. 1987). "Introns increase gene expression in cultured maize cells". eng. In: *Genes & Development* 1.10, pp. 1183–1200. ISSN: 0890-9369. DOI: 10.1101/gad.1.10.1183.

Carmody, Sean R. and Susan R. Wente (June 2009). "mRNA nuclear export at a glance". In: *Journal of Cell Science* 122.12, pp. 1933–1937. ISSN: 0021-9533. DOI: 10.1242/jcs.041236. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2723150/ (visited on 06/17/2020).

Castello, Alfredo et al. (May 2013). "RNA-binding proteins in Mendelian disease". en. In: *Trends in Genetics* 29.5, pp. 318–327. ISSN: 0168-9525. DOI: 10.1016/j.tig.2013.01.004. URL: http://www.sciencedirect.com/science/article/pii/S0168952513000164 (visited on 07/17/2020).

Cenik, Can et al. (2011). "Genome Analysis Reveals Interplay between 5UTR Introns and Nuclear mRNA Export for Secretory and Mitochondrial Genes". en. In: *PLOS Genetics* 7.4. Publisher: Public Library of Science, e1001366. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1001366. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1001366 (visited on 07/29/2020).

Chakrabarti, Anob M. et al. (2018). "Data Science Issues in Studying Protein–RNA Interactions with CLIP Technologies". In: *Annual Review of Biomedical Data Science* 1.1, annurev–biodatasci–080917–013525. ISSN: 2574-3414. DOI: 10.1146/annurev-biodatasci-080917-013525. URL: http://www.annualreviews.org/doi/10.1146/annurev-biodatasci-080917-013525.

Chazal, Pierre-Etienne et al. (Apr. 2013). "EJC core component MLN51 interacts with eIF3 and activates translation". en. In: *Proceedings of the National Academy of Sciences* 110.15. Publisher: National Academy of Sciences Section: Biological Sciences, pp. 5903–5908. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1218732110. URL: https://www.pnas.org/content/110/15/5903 (visited on 06/19/2020).

Chen, Beibei et al. (Jan. 2014). "PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis". In: *Genome Biology* 15.1, R18. ISSN: 1474-760X. DOI: 10.1186/gb-2014-15-1-r18. URL: https://doi.org/10.1186/gb-2014-15-1-r18 (visited on 06/09/2020).

Cléry, Antoine, Markus Blatter, and Frédéric H-T Allain (June 2008). "RNA recognition motifs: boring? Not quite". en. In: *Current Opinion in Structural Biology*. Nucleic acids / Sequences and topology 18.3, pp. 290–298. ISSN: 0959-440X. DOI: 10.1016/j.sbi.2008.04.002. URL: http://www.sciencedirect.com/science/article/pii/S0959440X08000584 (visited on 06/16/2020).

Conti, Elena and Elisa Izaurralde (June 2005). "Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species". en. In: *Current Opinion in Cell Biology*. Nucleus and gene expression 17.3, pp. 316–325. ISSN: 0955-0674. DOI: 10.1016/j.ceb.2005.04.005. URL: http://www.sciencedirect.com/science/article/pii/S0955067405000475 (visited on 07/29/2020).

Corcoran, David L et al. (2011). "PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data". In: *Genome Biology* 12.8, R79. ISSN: 1465-6906. DOI: 10.1186/gb-2011-12-8-r79. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3302668/ (visited on 06/09/2020).

Danan, Charles, Sudhir Manickavel, and Markus Hafner (2016). "PAR-CLIP: A Method for Transcriptome-Wide Identification of RNA Binding Protein Interaction Sites". In: *Methods in molecular biology (Clifton, N.J.)* 1358, pp. 153–173. ISSN: 1064-3745. DOI: 10.1007/978-1-4939-3067-8_10. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5142217/ (visited on 05/31/2020).

Darnell, Robert B. (2010). "HITS-CLIP: panoramic views of protein–RNA regulation in living cells". en. In: *WIREs RNA* 1.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wrna.31, pp. 266–286. ISSN: 1757-7012. DOI: 10.1002/wrna.31. URL: http://onlinelibrary.wiley.com/doi/abs/10.1002/wrna.31 (visited on 06/08/2020).

Dassi, Erik (2017). "Handshakes and Fights: The Regulatory Interplay of RNA-Binding Proteins". In: *Frontiers in Molecular Biosciences* 4.September, pp. 1–8. ISSN: 2296-889X. DOI: 10.3389/fmolb.2017.00067. URL: http://journal.frontiersin.org/article/10.3389/fmolb.2017.00067/full.

Davis, Carrie A et al. (Jan. 2018). "The Encyclopedia of DNA elements (ENCODE): data portal update". In: *Nucleic Acids Research* 46.Database issue, pp. D794–D801. ISSN: 0305-1048. DOI: 10.1093/nar/gkx1081. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5753278/ (visited on 07/07/2020).

Davydova, E. K. et al. (July 1997). "Overexpression in COS cells of p50, the major core protein associated with mRNA, results in translation inhibition". eng. In: *Nucleic Acids Research* 25.14, pp. 2911–2916. ISSN: 0305-1048. DOI: 10.1093/nar/25.14.2911.

De, Nfi (2018). "Chapter 12". In: *October* 1751. ISBN: 9781493977109.

Degot, Sébastien, Hervé Le Hir, et al. (Aug. 2004). "Association of the breast cancer protein MLN51 with the exon junction complex via its speckle localizer and RNA binding module". eng. In: *The Journal of Biological Chemistry* 279.32, pp. 33702–33715. ISSN: 0021-9258. DOI: 10.1074/jbc.M402754200.

Degot, Sébastien, Catherine H. Régnier, et al. (June 2002). "Metastatic Lymph Node 51, a novel nucleo-cytoplasmic protein overexpressed in breast cancer". eng. In: *Oncogene* 21.28, pp. 4422–4434. ISSN: 0950-9232. DOI: 10.1038/sj.onc.1205611.

Doma, Meenakshi K. and Roy Parker (Mar. 2006). "Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation". en. In: *Nature* 440.7083. Number: 7083 Publisher: Nature Publishing Group, pp. 561–564. ISSN: 1476-4687. DOI: 10.

1038/nature04530. URL: https://www.nature.com/articles/nature04530 (visited on 07/29/2020).

Dominguez, Daniel et al. (June 2018). "Sequence, Structure, and Context Preferences of Human RNA Binding Proteins". In: *Molecular Cell* 70.5, 854–867.e9. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2018.05.001. URL: http://www.sciencedirect.com/science/article/pii/S1097276518303514 (visited on 01/18/2019).

Dostie, Josée and Gideon Dreyfuss (July 2002). "Translation Is Required to Remove Y14 from mRNAs in the Cytoplasm". English. In: *Current Biology* 12.13. Publisher: Elsevier, pp. 1060–1067. ISSN: 0960-9822. DOI: 10.1016/S0960-9822(02)00902-8. URL: https://www.cell.com/current-biology/abstract/S0960-9822(02)00902-8 (visited on 06/20/2020).

Drexler, Heather L., Karine Choquet, and L. Stirling Churchman (Mar. 2020). "Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores". en. In: *Molecular Cell* 77.5, 985–998.e8. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2019.11.017. URL: http://www.sciencedirect.com/science/article/pii/S1097276519308652 (visited on 07/21/2020).

Dvinge, Heidi (2018). "Regulation of alternative mRNA splicing: old players and new perspectives". en. In: *FEBS Letters* 592.17. _eprint: https://febs.onlinelibrary.wiley.com/doi/pdf/10.1002/1873-3468.13119, pp. 2987–3006. ISSN: 1873-3468. DOI: 10.1002/1873-3468.13119. URL: https://febs.onlinelibrary.wiley.com/doi/abs/10.1002/1873-3468.13119 (visited on 06/16/2020).

Fatscher, Tobias, Volker Boehm, and Niels H. Gehring (Dec. 2015). "Mechanism, factors, and physiological role of nonsense-mediated mRNA decay". en. In: *Cellular and Molecular Life Sciences* 72.23, pp. 4523–4544. ISSN: 1420-9071. DOI: 10.1007/s00018-015-2017-9. URL: https://doi.org/10.1007/s00018-015-2017-9 (visited on 06/03/2020).

Favaro, Francine P. et al. (Jan. 2014). "A noncoding expansion in EIF4A3 causes Richieri-Costa-Pereira syndrome, a craniofacial disorder associated with limb defects". eng. In: *American Journal of Human Genetics* 94.1, pp. 120–128. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2013.11.020.

Fejes, Anthony P. et al. (Aug. 2008). "FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology". In: *Bioinformatics* 24.15, pp. 1729–1730. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btn305. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2638869/ (visited on 06/30/2020).

Fontrodona, Nicolas et al. (Apr. 2019). "Interplay between coding and exonic splicing regulatory sequences". en. In: *Genome Research*, gr.241315.118. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.241315.118. URL: http://genome.cshlp.org.insb.bib.cnrs.fr/content/early/2019/04/08/gr.241315.118 (visited on 04/10/2019).

Frankish, Adam et al. (2019). "GENCODE reference annotation for the human and mouse genomes". eng. In: *Nucleic Acids Research* 47.D1, pp. D766–D773. ISSN: 1362-4962. DOI: 10.1093/nar/gky955.

Frischmeyer, Pamela A. et al. (Mar. 2002). "An mRNA Surveillance Mechanism That Eliminates Transcripts Lacking Termination Codons". en. In: *Science* 295.5563. Publisher: American Association for the Advancement of Science Section: Report, pp. 2258–2261. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1067338. URL: https://science.sciencemag.org/content/295/5563/2258 (visited on 07/29/2020).

Fu, Xiang-Dong and Manuel Ares (Oct. 2014). "Context-dependent control of alternative splicing by RNA-binding proteins". In: *Nature reviews. Genetics* 15.10, pp. 689–701. ISSN: 1471-0056. DOI: 10.1038/nrg3778. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4440546/ (visited on 04/05/2019).

Garneau, Nicole L., Jeffrey Wilusz, and Carol J. Wilusz (Feb. 2007). "The highways and byways of mRNA decay". en. In: *Nature Reviews Molecular Cell Biology* 8.2. Number: 2 Publisher: Nature Publishing Group, pp. 113–126. ISSN: 1471-0080. DOI: 10.1038/nrm2104. URL: http://www.nature.com/articles/nrm2104 (visited on 07/29/2020).

Gáspár, Imre and Anne Ephrussi (Sept. 2017). "RNA localization feeds translation". en. In: *Science* 357.6357. Publisher: American Association for the Advancement of Science Section: Perspective, pp. 1235–1236. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aao5796. URL: https://science.sciencemag.org/content/357/6357/1235 (visited on 07/29/2020).

Gatfield, David et al. (Aug. 2003). "Nonsense-mediated mRNA decay in Drosophila:at the intersection of the yeast and mammalian pathways". In: *The EMBO Journal* 22.15. Publisher: John Wiley & Sons, Ltd, pp. 3960–3970. ISSN: 0261-4189. DOI: 10.1093/emboj/cdg371. URL: https://www.embopress.org/doi/full/10.1093/emboj/cdg371 (visited on 09/21/2020).

Gehring, Niels H., Joachim B. Kunz, et al. (Oct. 2005). "Exon-Junction Complex Components Specify Distinct Routes of Nonsense-Mediated mRNA Decay with Differential Cofactor Requirements". en. In: *Molecular Cell* 20.1, pp. 65–75. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2005.08.012. URL: http://www.sciencedirect.com/science/article/pii/S1097276505015546 (visited on 07/29/2020).

Gehring, Niels H., Styliani Lamprinaki, Matthias W. Hentze, et al. (May 2009). "The hierarchy of exon-junction complex assembly by the spliceosome explains key features of mammalian nonsense-mediated mRNA decay". eng. In: *PLoS biology* 7.5, e1000120. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1000120.

Gehring, Niels H., Styliani Lamprinaki, Andreas E. Kulozik, et al. (May 2009). "Disassembly of exon junction complexes by PYM". eng. In: *Cell* 137.3, pp. 536–548. ISSN: 1097-4172. DOI: 10.1016/j.cell.2009.02.042.

Gerstberger, Stefanie, Markus Hafner, and Thomas Tuschl (2014). "A census of human RNA-binding proteins". In: *Nature Reviews Genetics* 15.12. arXiv: Figures, S., 2010. Supplementary information. Nature, 1(c), pp.1–7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3006164&tool=pmcentrez&rendertype=abs Publisher: Nature Publishing Group ISBN: 1471-0064 (Electronic)\r1471-0056 (Linking), pp. 829–845. ISSN: 1471-0056. DOI: 10.1038/nrg3813. URL: http://www.nature.com/doifinder/10.1038/nrg3813.

Ghosh, Sanjay, Virginie Marchand, et al. (Apr. 2012). "Control of RNP motility and localization by a splicing-dependent structure in oskar mRNA". en. In: *Nature Structural & Molecular Biology* 19.4. Number: 4 Publisher: Nature Publishing Group, pp. 441–449. ISSN: 1545-9985. DOI: 10.1038/nsmb.2257. URL: http://www.nature.com/articles/nsmb.2257 (visited on 06/05/2020).

Ghosh, Sanjay, Ales Obrdlik, et al. (2014). "The EJC Binding and Dissociating Activity of PYM Is Regulated in Drosophila". en. In: *PLOS Genetics* 10.6. Publisher: Public Library of Science, e1004455. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004455. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004455 (visited on 06/20/2020).

Glisovic, Tina et al. (June 2008). "RNA-binding proteins and post-transcriptional gene regulation". In: *FEBS letters* 582.14, pp. 1977–1986. ISSN: 0014-5793. DOI: 10.1016/j.febslet.2008.03.004. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2858862/ (visited on 06/11/2020).

Graveley, B. R., K. J. Hertel, and T. Maniatis (Nov. 1998). "A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers". eng. In: *The EMBO journal* 17.22, pp. 6747–6756. ISSN: 0261-4189. DOI: 10.1093/emboj/17.22.6747.

Gromadzka, Agnieszka M. et al. (Mar. 2016). "A short conserved motif in ALYREF directs cap- and EJC-dependent assembly of export complexes on spliced mRNAs". In: *Nucleic Acids Research* 44.5, pp. 2348–2361. ISSN: 0305-1048. DOI: 10.1093/nar/gkw009. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4797287/ (visited on 06/12/2020).

Günzl, Arthur and Albrecht Bindereif (1999). "Oligonucleotide-Targeted RNase H Protection Analysis of RNA-Protein Complexes". en. In: *RNA-Protein Interaction Protocols*. Ed. by Susan R. Haynes. Methods in Molecular Biology™. Totowa, NJ: Humana Press, pp. 93–103. ISBN: 978-1-59259-676-8. DOI: 10.1385/1-59259-676-2:93. URL: https://doi.org/10.1385/1-59259-676-2:93 (visited on 06/18/2020).

Haberman, Nejc et al. (Jan. 2017). "Insights into the design and interpretation of iCLIP experiments". In: *Genome Biology* 18.1, p. 7. ISSN: 1474-760X. DOI: 10.1186/s13059-016-1130-x. URL: https://doi.org/10.1186/s13059-016-1130-x (visited on 08/10/2020).

Hachet, Olivier and Anne Ephrussi (Apr. 2004). "Splicing of oskar RNA in the nucleus is coupled to its cytoplasmic localization". en. In: *Nature* 428.6986. Number: 6986 Publisher: Nature Publishing Group, pp. 959–963. ISSN: 1476-4687. DOI: 10.1038/nature02521. URL: http://www.nature.com/articles/nature02521 (visited on 06/05/2020).

Haselbach, David et al. (2018). "Structure and Conformational Dynamics of the Human Spliceosomal Bact Complex". eng. In: *Cell* 172.3, 454–464.e11. ISSN: 1097-4172. DOI: 10.1016/j.cell.2018.01.010.

Hauer, Christian, Tomaz Curk, et al. (2015). "Improved binding site assignment by high-resolution mapping of RNA–protein interactions using iCLIP". In: *Nature Communications* 6. Publisher: Nature Publishing Group ISBN: 2041-1723, p. 7921. ISSN: 2041-1723. DOI: 10.1038/ncomms8921. URL: http://dx.doi.org/10.1038/ncomms8921%5Cnhttp://www.nature.com/ncomms/2015/150811/ncomms8921/full/ncomms8921.html%5Cnhttp://www.nature.com/doifinder/10.1038/ncomms8921.

Hauer, Christian, Jana Sieber, et al. (2016). "Exon Junction Complexes Show a Distributional Bias toward Alternatively Spliced mRNAs and against mRNAs Coding for Ribosomal Proteins". In: *Cell Reports* 16.6. Publisher: The Author(s), pp. 1588–1603. ISSN: 22111247. DOI: 10.1016/j.celrep.2016.06.096. URL: http://dx.doi.org/10.1016/j.celrep.2016.06.096.

Hayashi, Rippei et al. (Aug. 2014). "The exon junction complex is required for definition and excision of neighboring introns in Drosophila". en. In: *Genes & Development* 28.16. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1772–1785. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.245738.114. URL: http://genesdev.cshlp.org/content/28/16/1772 (visited on 07/29/2020).

Heinz, Sven et al. (May 2010). "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities". eng. In: *Molecular Cell* 38.4, pp. 576–589. ISSN: 1097-4164. DOI: 10.1016/j.molcel.2010.05.004.

Hentze, Matthias W. et al. (2018). "A brave new world of RNA-binding proteins". In: *Nature Reviews Molecular Cell Biology* 19.5. Publisher: Nature Publishing Group, pp. 327–341. ISSN: 1471-0072. DOI: 10.1038/nrm.2017.130. URL: http://www.nature.com/doifinder/10.1038/nrm.2017.130.

Herzel, Lydia et al. (Oct. 2017). "Splicing and transcription touch base: co-transcriptional spliceosome assembly and function". en. In: *Nature Reviews Molecular Cell Biology* 18.10. Number: 10 Publisher: Nature Publishing Group, pp. 637–650. ISSN: 1471-

0080. DOI: 10.1038/nrm.2017.63. URL: http://www.nature.com/articles/nrm.2017.63 (visited on 06/21/2020).

Hocq, Rémi et al. (Sept. 2018). "Monitored eCLIP: high accuracy mapping of RNA-protein interactions". eng. In: *Nucleic Acids Research*. ISSN: 1362-4962. DOI: 10.1093/nar/gky858.

Hoof, Ambro van et al. (Mar. 2002). "Exosome-Mediated Recognition and Degradation of mRNAs Lacking a Termination Codon". en. In: *Science* 295.5563. Publisher: American Association for the Advancement of Science Section: Report, pp. 2262–2264. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1067272. URL: https://science-sciencemag-org.insb.bib.cnrs.fr/content/295/5563/2262 (visited on 07/29/2020).

Hopp, Thomas P. et al. (Oct. 1988). "A Short Polypeptide Marker Sequence Useful for Recombinant Protein Identification and Purification". en. In: *Bio/Technology* 6.10. Number: 10 Publisher: Nature Publishing Group, pp. 1204–1210. ISSN: 1546-1696. DOI: 10.1038/nbt1088-1204. URL: https://www.nature.com/articles/nbt1088-1204 (visited on 07/20/2020).

Hurt, Jessica a, Alex D Robertson, and Christopher B Burge (2013). "Global analyses of UPF1 binding and function reveals expanded scope of nonsense-mediated mRNA decay Global analyses of UPF1 binding and function reveals expanded scope of nonsense-mediated mRNA decay Department of Biology". In: ISBN: 1549-5469 (Electronic)\r1088-9051 (Linking), pp. 1636–1650. ISSN: 10889051. DOI: 10.1101/gr.157354.113.

Ingolia, Nicholas T. (Mar. 2014). "Ribosome profiling: new views of translation, from single codons to genome scale". en. In: *Nature Reviews Genetics* 15.3. Number: 3 Publisher: Nature Publishing Group, pp. 205–213. ISSN: 1471-0064. DOI: 10.1038/nrg3645. URL: http://www.nature.com/articles/nrg3645 (visited on 07/01/2020).

Jackson, Richard J., Christopher U. T. Hellen, and Tatyana V. Pestova (Feb. 2010). "The mechanism of eukaryotic translation initiation and principles of its regulation". en. In: *Nature Reviews Molecular Cell Biology* 11.2. Number: 2 Publisher: Nature Publishing Group, pp. 113–127. ISSN: 1471-0080. DOI: 10.1038/nrm2838. URL: https://www.nature.com/articles/nrm2838 (visited on 06/17/2020).

— (Jan. 2012). "Termination and post-termination events in eukaryotic translation". en. In: *Advances in Protein Chemistry and Structural Biology*. Ed. by Assen Marintchev. Vol. 86. Fidelity and Quality Control in Gene Expression. Academic Press, pp. 45–93. DOI: 10.1016/B978-0-12-386497-0.00002-5. URL: http://www.sciencedirect.com/science/article/pii/B9780123864970000025 (visited on 06/17/2020).

Jacobson, Allan (Jan. 1996). "16 Poly(A) Metabolism and Translation: The Closed-loop Model". en-US. In: *Cold Spring Harbor Monograph Archive* 30.0. Number: 0, pp. 451–480. DOI: 10.1101/0.451-480. URL: https://cshmonographs.org/index.php/monographs/article/view/3317 (visited on 06/17/2020).

Järvelin, Aino I. et al. (Apr. 2016). "The new (dis)order in RNA regulation". In: *Cell Communication and Signaling* 14.1, p. 9. ISSN: 1478-811X. DOI: 10.1186/s12964-016-0132-3. URL: https://doi.org/10.1186/s12964-016-0132-3 (visited on 06/16/2020).

Johnstone, Oona and Paul Lasko (2001). "Translational Regulation and RNA Localization in Drosophila Oocytes and Embryos". In: *Annual Review of Genetics* 35.1. _eprint: https://doi.org/10.1146/annurev.genet.35.102401.090756, pp. 365–406. DOI: 10.1146/annurev.genet.35.102401.090756. URL: https://doi.org/10.1146/annurev.genet.35.102401.090756 (visited on 06/17/2020).

Kataoka, N. et al. (Sept. 2000). "Pre-mRNA splicing imprints mRNA in the nucleus with a novel RNA-binding protein that persists in the cytoplasm". eng. In: *Molecular Cell* 6.3, pp. 673–682. ISSN: 1097-2765. DOI: 10.1016/s1097-2765(00)00065-4.

Keene, Jack D., Jordan M. Komisarow, and Matthew B. Friedersdorf (June 2006). "RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts". en. In: *Nature Protocols* 1.1, pp. 302–307. ISSN: 1750-2799. DOI: 10.1038/nprot.2006.47. URL: https://www.nature.com/articles/nprot.2006.47 (visited on 10/01/2019).

Keren, Hadas, Galit Lev-Maor, and Gil Ast (May 2010). "Alternative splicing and evolution: diversification, exon definition and function". eng. In: *Nature Reviews. Genetics* 11.5, pp. 345–355. ISSN: 1471-0064. DOI: 10.1038/nrg2776.

Kim-Ha, Jeongsil, Jeffrey L. Smith, and Paul M. Macdonald (July 1991). "oskar mRNA is localized to the posterior pole of the Drosophila oocyte". English. In: *Cell* 66.1. Publisher: Elsevier, pp. 23–35. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/0092-8674(91)90136-M. URL: https://www-cell-com.insb.bib.cnrs.fr/cell/abstract/0092-8674(91)90136-M (visited on 06/23/2020).

König, Julian et al. (2010). "ICLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution". In: *Nature Structural and Molecular Biology* 17.7. ISBN: 1545-9985 (Electronic)\r1545-9985 (Linking), pp. 909–915. ISSN: 15459993. DOI: 10.1038/nsmb.1838.

Krakau, Sabrina, Hugues Richard, and Annalisa Marsico (2017). "PureCLIP : capturing target-specific protein – RNA interaction footprints from single-nucleotide CLIP-seq data". In: Publisher: Genome Biology, pp. 1–17. DOI: 10.1186/s13059-017-1364-2.

Landt, Stephen G. et al. (2012). "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia". In: *Genome Research* 22.9. ISBN: 1549-5469 (Electronic)\r1088-9051 (Linking), pp. 1813–1831. ISSN: 10889051. DOI: 10.1101/gr.136184.111.

Lau, Chi-Kong et al. (May 2003). "Structure of the Y14-Magoh Core of the Exon Junction Complex". en. In: *Current Biology* 13.11, pp. 933–941. ISSN: 0960-9822. DOI: 10.1016/S0960-9822(03)00328-2. URL: http://www.sciencedirect.com/science/article/pii/S0960982203003282 (visited on 06/19/2020).

Le Hir, Hervé, David Gatfield, I. C. Braun, et al. (Dec. 2001). "The protein Mago provides a link between splicing and mRNA localization". eng. In: *EMBO reports* 2.12, pp. 1119–1124. ISSN: 1469-221X. DOI: 10.1093/embo-reports/kve245.

Le Hir, Hervé, David Gatfield, Elisa Izaurralde, et al. (Sept. 2001). "The exon–exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay". In: *The EMBO Journal* 20.17, pp. 4987–4997. ISSN: 0261-4189. DOI: 10.1093/emboj/20.17.4987. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC125616/ (visited on 06/19/2020).

Le Hir, Hervé, Elisa Izaurralde, et al. (2000). "The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions". In: *EMBO Journal* 19.24. ISBN: 0261-4189, pp. 6860–6869. ISSN: 02614189. DOI: 10.1093/emboj/19.24.6860.

Le Hir, Hervé, Jérôme Saulière, and Zhen Wang (Jan. 2016). "The exon junction complex as a node of post-transcriptional networks". en. In: *Nature Reviews Molecular Cell Biology* 17.1. Number: 1 Publisher: Nature Publishing Group, pp. 41–54. ISSN: 1471-0080. DOI: 10.1038/nrm.2015.7. URL: http://www.nature.com/articles/nrm.2015.7 (visited on 05/30/2020).

Lee, Flora C.Y. and Jernej Ule (2018). "Advances in CLIP Technologies for Studies of Protein-RNA Interactions". In: *Molecular Cell* 69.3. Publisher: Elsevier Inc., pp. 354–369. ISSN: 10974164. DOI: 10.1016/j.molcel.2018.01.005. URL: https://doi.org/10.1016/j.molcel.2018.01.005.

Legrand, Carine and Francesca Tuorto (Jan. 2020). "RiboVIEW: a computational framework for visualization, quality control and statistical analysis of ribosome profiling data". en. In: *Nucleic Acids Research* 48.2. Publisher: Oxford Academic, e7–e7. ISSN:

0305-1048. DOI: 10.1093/nar/gkz1074. URL: https://academic.oup.com/nar/article/48/2/e7/5645003 (visited on 07/29/2020).

Lejeune, Fabrice et al. (July 2002). "The exon junction complex is detected on CBP80-bound but not eIF4E-bound mRNA in mammalian cells: dynamics of mRNP remodeling". In: *The EMBO Journal* 21.13, pp. 3536–3545. ISSN: 0261-4189. DOI: 10.1093/emboj/cdf345. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC126094/ (visited on 06/20/2020).

Lemaire, Sébastien et al. (2019). "Characterizing the interplay between gene nucleotide composition bias and splicing". eng. In: *Genome Biology* 20.1, p. 259. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1869-y.

Li, Qunhua et al. (2011). "Measuring reproducibility of high-throughput experiments". In: *Annals of Applied Statistics* 5.3. arXiv: 1110.4705 ISBN: 1932-6157, pp. 1752–1779. ISSN: 19326157. DOI: 10.1214/11-AOAS466.

Liu, Qi et al. (2015). "Assessing Computational Steps for CLIP-Seq Data Analysis". In: *BioMed Research International* 2015, pp. 27–34. ISSN: 23146141. DOI: 10.1155/2015/196082.

Long, Jennifer C. and Javier F. Caceres (Jan. 2009). "The SR protein family of splicing factors: master regulators of gene expression". en. In: *Biochemical Journal* 417.1. Publisher: Portland Press, pp. 15–27. ISSN: 0264-6021. DOI: 10.1042/BJ20081501. URL: /biochemj/article/417/1/15/45140/The-SR-protein-family-of-splicing-factors-master (visited on 06/16/2020).

Losson, R and F Lacroute (Oct. 1979). "Interference of nonsense mutations with eukaryotic messenger RNA stability." In: *Proceedings of the National Academy of Sciences of the United States of America* 76.10, pp. 5134–5137. ISSN: 0027-8424. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC413094/ (visited on 07/29/2020).

Lovci, Michael T. et al. (Dec. 2013). "Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges". eng. In: *Nature Structural & Molecular Biology* 20.12, pp. 1434–1442. ISSN: 1545-9985. DOI: 10.1038/nsmb.2699.

Love, Michael I., Wolfgang Huber, and Simon Anders (Dec. 2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12, p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8. URL: https://doi.org/10.1186/s13059-014-0550-8 (visited on 07/16/2020).

Lunde, Bradley M., Claire Moore, and Gabriele Varani (June 2007). "RNA-binding proteins: modular design for efficient function". In: *Nature reviews. Molecular cell biology* 8.6, pp. 479–490. ISSN: 1471-0072. DOI: 10.1038/nrm2178. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5507177/ (visited on 06/16/2020).

Luo, M. J. and R. Reed (Dec. 1999). "Splicing is required for rapid and efficient mRNA export in metazoans". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 96.26, pp. 14937–14942. ISSN: 0027-8424. DOI: 10.1073/pnas.96.26.14937.

Lykke-Andersen, Jens, Mei-Di Shu, and Joan A. Steitz (Sept. 2001). "Communication of the Position of Exon-Exon Junctions to the mRNA Surveillance Machinery by the Protein RNPS1". en. In: *Science* 293.5536. Publisher: American Association for the Advancement of Science Section: Report, pp. 1836–1839. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1062786. URL: https://science.sciencemag.org/content/293/5536/1836 (visited on 07/29/2020).

Ma, Xiaoju Max et al. (Apr. 2008). "SKAR Links Pre-mRNA Splicing to mTOR/S6K1-Mediated Enhanced Translation Efficiency of Spliced mRNAs". English. In: *Cell* 133.2. Publisher: Elsevier, pp. 303–313. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2008.02.031. URL: https://www.cell.com/cell/abstract/S0092-8674(08)00282-1 (visited on 06/22/2020).

Mabin, Justin W. et al. (Nov. 2018). "The Exon Junction Complex Undergoes a Compositional Switch that Alters mRNP Structure and Nonsense-Mediated mRNA Decay Activity". In: *Cell Reports* 25.9, 2431–2446.e7. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2018.11.046. URL: http://www.sciencedirect.com/science/article/pii/S2211124718318096 (visited on 01/18/2019).

Majumdar, Romit, Amitabha Bandyopadhyay, and Umadas Maitra (Feb. 2003). "Mammalian Translation Initiation Factor eIF1 Functions with eIF1A and eIF3 in the Formation of a Stable 40 S Preinitiation Complex". en. In: *Journal of Biological Chemistry* 278.8. Publisher: American Society for Biochemistry and Molecular Biology, pp. 6580–6587. ISSN: 0021-9258, 1083-351X. DOI: 10.1074/jbc.M210357200. URL: http://www.jbc.org/content/278/8/6580 (visited on 06/17/2020).

Malone, Colin D. et al. (Aug. 2014). "The exon junction complex controls transposable element activity by ensuring faithful splicing of the piwi transcript". In: *Genes & Development* 28.16, pp. 1786–1799. ISSN: 0890-9369. DOI: 10.1101/gad.245829.114. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4197963/ (visited on 06/12/2020).

Maquat, L. E. (July 1995). "When cells stop making sense: effects of nonsense codons on RNA metabolism in vertebrate cells". eng. In: *RNA (New York, N.Y.)* 1.5, pp. 453–465. ISSN: 1355-8382.

Maquat, Lynne E. (May 2005). "Nonsense-mediated mRNA decay in mammals". en. In: *Journal of Cell Science* 118.9. Publisher: The Company of Biologists Ltd Section: Cell Science at a Glance, pp. 1773–1776. ISSN: 0021-9533, 1477-9137. DOI: 10.1242/jcs.01701. URL: https://jcs.biologists.org/content/118/9/1773 (visited on 06/04/2020).

Maquat, Lynne E. et al. (Dec. 1981). "Unstable -globin mRNA in mRNA-deficient 0 thalassemia". en. In: *Cell* 27.3, Part 2, pp. 543–553. ISSN: 0092-8674. DOI: 10.1016/0092-8674(81)90396-2. URL: http://www.sciencedirect.com/science/article/pii/0092867481903962 (visited on 07/29/2020).

Martin, Georges and Mihaela Zavolan (June 2016). "Redesigning CLIP for efficiency, accuracy and speed". en. In: *Nature Methods* 13.6. Number: 6 Publisher: Nature Publishing Group, pp. 482–483. ISSN: 1548-7105. DOI: 10.1038/nmeth.3870. URL: http://www.nature.com/articles/nmeth.3870 (visited on 06/08/2020).

Martin, Kelsey C. and Anne Ephrussi (Feb. 2009). "mRNA Localization: Gene Expression in the Spatial Dimension". In: *Cell* 136.4, p. 719. ISSN: 0092-8674. DOI: 10.1016/j.cell.2009.01.044. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2819924/ (visited on 06/17/2020).

Mayeda, Akila and Adrian R. Krainer (Feb. 2012). "In Vitro Splicing Assays". English. In: *Alternative pre-mRNA Splicing: Theory and Protocols*. Publisher: Wiley-VCH, pp. 320–329. DOI: 10.1002/9783527636778.ch30. URL: https://pure.fujita-hu.ac.jp/en/publications/in-vitro-splicing-assays (visited on 06/18/2020).

McMahon, J. J., E. E. Miller, and D. L. Silver (Dec. 2016). "The exon junction complex in neural development and neurodevelopmental disease". In: *International Journal of Developmental Neuroscience* 55, pp. 117–123. ISSN: 0736-5748. DOI: 10.1016/j.ijdevneu.2016.03.006. URL: http://www.sciencedirect.com/science/article/pii/S0736574816300478 (visited on 01/18/2019).

Metkar, Mihir et al. (Nov. 2018). "Higher-Order Organization Principles of Pretranslational mRNPs". en. In: *Molecular Cell* 72.4, 715–726.e3. ISSN: 10972765. DOI: 10.1016/j.molcel.2018.09.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S1097276518307834 (visited on 11/16/2018).

Mi, Huaiyu et al. (Jan. 2019). "PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools". en. In: *Nucleic Acids Research* 47.D1. Publisher: Oxford Academic, pp. D419–D426. ISSN: 0305-1048. DOI:

10.1093/nar/gky1038. URL: https://academic.oup.com/nar/article/47/D1/D419/5165346 (visited on 06/28/2020).

Mingot, José-Manuel et al. (July 2001). "Importin 13: a novel mediator of nuclear import and export". In: *The EMBO Journal* 20.14, pp. 3685–3694. ISSN: 0261-4189. DOI: 10.1093/emboj/20.14.3685. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC125545/ (visited on 06/20/2020).

Moore, Melissa J. and Nick J. Proudfoot (Feb. 2009). "Pre-mRNA Processing Reaches Back to Transcription and Ahead to Translation". en. In: *Cell* 136.4, pp. 688–700. ISSN: 0092-8674. DOI: 10.1016/j.cell.2009.02.001. URL: http://www.sciencedirect.com/science/article/pii/S0092867409001330 (visited on 07/28/2020).

Mortazavi, Ali et al. (July 2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". en. In: *Nature Methods* 5.7. Number: 7 Publisher: Nature Publishing Group, pp. 621–628. ISSN: 1548-7105. DOI: 10.1038/nmeth.1226. URL: http://www.nature.com/articles/nmeth.1226 (visited on 07/11/2020).

Newmark, P. A. and R. E. Boswell (May 1994). "The mago nashi locus encodes an essential product required for germ plasm assembly in Drosophila". en. In: *Development* 120.5. Publisher: The Company of Biologists Ltd Section: JOURNAL ARTICLES, pp. 1303–1313. ISSN: 0950-1991, 1477-9129. URL: https://dev.biologists.org/content/120/5/1303 (visited on 06/19/2020).

Nguyen, Nga Thi Thuy et al. (July 2018). "RSAT 2018: regulatory sequence analysis tools 20th anniversary". en. In: *Nucleic Acids Research* 46.W1. Publisher: Oxford Academic, W209–W214. ISSN: 0305-1048. DOI: 10.1093/nar/gky317. URL: https://academic.oup.com/nar/article/46/W1/W209/4990780 (visited on 03/05/2020).

Nielsen, Klaus H. et al. (Jan. 2009). "Mechanism of ATP turnover inhibition in the EJC". en. In: *RNA* 15.1. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 67–75. ISSN: 1355-8382, 1469-9001. DOI: 10.1261/rna.1283109. URL: http://rnajournal.cshlp.org/content/15/1/67 (visited on 07/29/2020).

NOTT, AJIT, SHLOMO H. MEISLIN, and MELISSA J. MOORE (May 2003). "A quantitative analysis of intron effects on mammalian gene expression". In: *RNA* 9.5, pp. 607–617. ISSN: 1355-8382. DOI: 10.1261/rna.5250403. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1370426/ (visited on 06/22/2020).

Nussbacher, Julia K. and Gene W. Yeo (2018). "Systematic Discovery of RNA Binding Proteins that Regulate MicroRNA Levels". In: *Molecular Cell* 69.6. Publisher: Elsevier Inc., 1005–1016.e7. ISSN: 10972765. DOI: 10.1016/j.molcel.2018.02.012. URL: http://linkinghub.elsevier.com/retrieve/pii/S1097276518301102.

Obrdlik, Ales et al. (July 2019). "The Transcriptome-wide Landscape and Modalities of EJC Binding in Adult Drosophila". en. In: *Cell Reports* 28.5, 1219–1236.e11. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2019.06.088. URL: http://www.sciencedirect.com/science/article/pii/S221112471930868X (visited on 03/18/2020).

Palacios, Isabel M. et al. (Feb. 2004). "An eIF4AIII-containing complex required for mRNA localization and nonsense-mediated mRNA decay". eng. In: *Nature* 427.6976, pp. 753–757. ISSN: 1476-4687. DOI: 10.1038/nature02351.

Palmiter, R D et al. (Jan. 1991). "Heterologous introns can enhance expression of transgenes in mice." In: *Proceedings of the National Academy of Sciences of the United States of America* 88.2, pp. 478–482. ISSN: 0027-8424. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC50834/ (visited on 06/22/2020).

Pan, Qun et al. (Dec. 2008). "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing". en. In: *Nature Genetics* 40.12. Number: 12 Publisher: Nature Publishing Group, pp. 1413–1415. ISSN: 1546-1718.

DOI: 10.1038/ng.259. URL: https://www.nature.com/articles/ng.259 (visited on 06/11/2020).

Prévôt, Déborah, Jean-Luc Darlix, and Théophile Ohlmann (2003). "Conducting the initiation of protein synthesis: the role of eIF4G". en. In: *Biology of the Cell* 95.3-4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1016/S0248-4900%2803%2900031-5, pp. 141–156. ISSN: 1768-322X. DOI: 10.1016/S0248-4900(03)00031-5. URL: http://onlinelibrary.wiley.com/doi/abs/10.1016/S0248-4900%2803%2900031-5 (visited on 07/28/2020).

Quinones-Valdez, Giovanni et al. (Jan. 2019). "Regulation of RNA editing by RNA-binding proteins in human cells". en. In: *Communications Biology* 2.1, pp. 1–14. ISSN: 2399-3642. DOI: 10.1038/s42003-018-0271-8. URL: https://www.nature.com/articles/s42003-018-0271-8 (visited on 01/28/2020).

Reyes-Herrera, Paula H and Elisa Ficarra (Oct. 2014). "Computational Methods for CLIP-seq Data Processing". In: *Bioinformatics and Biology Insights* 8, pp. 199–207. ISSN: 1177-9322. DOI: 10.4137/BBI.S16803. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4196881/ (visited on 06/01/2020).

Reyes-herrera, Paula H and Elisa Ficarra (2014). "Bioinformatics and Biology Insights Computational Methods for CLIP-seq Data Processing". In: pp. 199–207. ISSN: 1177-9322. DOI: 10.4137/BBI.S16803.Received.

Roignant, Jean-Yves and Jessica E. Treisman (Oct. 2010). "Exon junction complex subunits are required to splice Drosophila MAP kinase, a large heterochromatic gene". In: *Cell* 143.2, pp. 238–250. ISSN: 0092-8674. DOI: 10.1016/j.cell.2010.09.036. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2955985/ (visited on 06/23/2020).

Saulière, Jérôme, Nazmul Haque, et al. (Oct. 2010). "The exon junction complex differentially marks spliced junctions". en. In: *Nature Structural & Molecular Biology* 17.10, pp. 1269–1271. ISSN: 1545-9985. DOI: 10.1038/nsmb.1890. URL: https://www.nature.com/articles/nsmb.1890 (visited on 04/24/2019).

Saulière, Jérôme, Valentine Murigneux, et al. (Nov. 2012). "CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex". en. In: *Nature Structural & Molecular Biology* 19.11. Number: 11 Publisher: Nature Publishing Group, pp. 1124–1131. ISSN: 1545-9985. DOI: 10.1038/nsmb.2420. URL: http://www.nature.com/articles/nsmb.2420 (visited on 06/28/2020).

Schoenberg, Daniel R. and Lynne E. Maquat (Apr. 2012). "Regulation of cytoplasmic mRNA decay". en. In: *Nature Reviews Genetics* 13.4. Number: 4 Publisher: Nature Publishing Group, pp. 246–259. ISSN: 1471-0064. DOI: 10.1038/nrg3160. URL: https://www.nature.com/articles/nrg3160 (visited on 06/17/2020).

SEQC (Sept. 2014). "A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control consortium". In: *Nature biotechnology* 32.9, pp. 903–914. ISSN: 1087-0156. DOI: 10.1038/nbt.2957. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4321899/ (visited on 05/02/2019).

Shah, Ankeeta et al. (2017). "CLIP Tool Kit (CTK): A flexible and robust pipeline to analyze CLIP sequencing data". In: *Bioinformatics* 33.4, pp. 566–567. ISSN: 14602059. DOI: 10.1093/bioinformatics/btw653.

Shi, Hang and Rui-Ming Xu (Apr. 2003). "Crystal structure of the Drosophila Mago nashi–Y14 complex". en. In: *Genes & Development* 17.8. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 971–976. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.260403. URL: http://genesdev.cshlp.org/content/17/8/971 (visited on 06/19/2020).

Shibuya, Toshiharu, Thomas Tange, Nahum Sonenberg, et al. (Apr. 2004). "eIF4AIII binds spliced mRNA in the exon junction complex and is essential for nonsense-mediated decay". eng. In: *Nature Structural & Molecular Biology* 11.4, pp. 346–351. ISSN: 1545-9993. DOI: 10.1038/nsmb750.

Shibuya, Toshiharu, Thomas Tange, M. Elizabeth Stroupe, et al. (Mar. 2006). "Mutational analysis of human eIF4AIII identifies regions necessary for exon junction complex formation and nonsense-mediated mRNA decay". eng. In: *RNA (New York, N.Y.)* 12.3, pp. 360–374. ISSN: 1355-8382. DOI: 10.1261/rna.2190706.

Sievers, Cem et al. (Nov. 2012). "Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data". eng. In: *Nucleic Acids Research* 40.20, e160. ISSN: 1362-4962. DOI: 10.1093/nar/gks697.

Silver, Debra L. et al. (May 2010). "The exon junction complex component Magoh controls brain size by regulating neural stem cell division". eng. In: *Nature Neuroscience* 13.5, pp. 551–558. ISSN: 1546-1726. DOI: 10.1038/nn.2527.

Singh, Guramrit, Alper Kucukural, et al. (2012). "The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR protein nexus". In: *Cell* 151.4. arXiv: NIHMS150003 Publisher: Elsevier Inc. ISBN: 1097-4172 (Electronic)\n0092-8674 (Linking), pp. 750–764. ISSN: 00928674. DOI: 10.1016/j.cell.2012.10.007. URL: http://dx.doi.org/10.1016/j.cell.2012.10.007.

Singh, Guramrit, Gabriel Pratt, et al. (2015). "The Clothes Make the mRNA: Past and Present Trends in mRNP Fashion". In: *Annual Review of Biochemistry* 84.1. arXiv: 15334406 ISBN: 0066-4154, pp. 325–354. ISSN: 0066-4154. DOI: 10.1146/annurev-biochem-080111-092106. URL: http://www.annualreviews.org/doi/10.1146/annurev-biochem-080111-092106.

Singh, Kusum K et al. (Aug. 2013). "Two mammalian MAGOH genes contribute to exon junction complex composition and nonsense-mediated decay". In: *RNA Biology* 10.8, pp. 1291–1298. ISSN: 1547-6286. DOI: 10.4161/rna.25827. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3817150/ (visited on 06/19/2020).

Smith, Tom, Andreas Heger, and Ian Sudbery (Mar. 2017). "UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy". en. In: *Genome Research* 27.3, pp. 491–499. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.209601.116. URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.209601.116 (visited on 12/13/2018).

Sonenberg, Nahum and Alan G. Hinnebusch (Feb. 2009). "Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets". en. In: *Cell* 136.4, pp. 731–745. ISSN: 0092-8674. DOI: 10.1016/j.cell.2009.01.042. URL: http://www.sciencedirect.com/science/article/pii/S0092867409000907 (visited on 07/29/2020).

Sugimoto, Yoichiro et al. (Aug. 2012). "Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions". In: *Genome Biology* 13.8, R67. ISSN: 1474-760X. DOI: 10.1186/gb-2012-13-8-r67. URL: https://doi.org/10.1186/gb-2012-13-8-r67 (visited on 06/26/2020).

Tange, Thomas et al. (Dec. 2005). "Biochemical analysis of the EJC reveals two new factors and a stable tetrameric protein core". In: *RNA* 11.12, pp. 1869–1883. ISSN: 1355-8382. DOI: 10.1261/rna.2155905. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1370875/ (visited on 06/04/2020).

Tenenbaum, Scott A. et al. (Dec. 2000). "Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays". en. In: *Proceedings of the National Academy of Sciences* 97.26. Publisher: National Academy of Sciences Section: Biological Sciences, pp. 14085–14090. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.97.26.14085. URL: https://www.pnas.org/content/97/26/14085 (visited on 06/12/2020).

"The Gene Ontology Resource" (Jan. 2019). "The Gene Ontology Resource: 20 years and still GOing strong". en. In: *Nucleic Acids Research* 47.D1. Publisher: Oxford Academic, pp. D330–D338. ISSN: 0305-1048. DOI: 10.1093/nar/gky1055. URL: https://academic.oup.com/nar/article/47/D1/D330/5160994 (visited on 06/28/2020).

Theurkauf, W. E. et al. (2006). "rasiRNAs, DNA damage, and embryonic axis specification". eng. In: *Cold Spring Harbor Symposia on Quantitative Biology* 71, pp. 171–180. ISSN: 0091-7451. DOI: 10.1101/sqb.2006.71.066.

Uhl, Michael et al. (2017). "Computational analysis of CLIP-seq data". In: *Methods* 118-119.February, pp. 60–72. ISSN: 10959130. DOI: 10.1016/j.ymeth.2017.02.006.

Ule, Jernej, Hun-Way Hwang, and Robert B. Darnell (Aug. 2018). "The Future of Cross-Linking and Immunoprecipitation (CLIP)". eng. In: *Cold Spring Harbor Perspectives in Biology* 10.8. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a032243.

Ule, Jernej, Kirk B. Jensen, et al. (Nov. 2003). "CLIP identifies Nova-regulated RNA networks in the brain". eng. In: *Science (New York, N.Y.)* 302.5648, pp. 1212–1215. ISSN: 1095-9203. DOI: 10.1126/science.1090095.

Ule, Jernej, Kirk Jensen, et al. (Dec. 2005). "CLIP: A method for identifying protein–RNA interaction sites in living cells". en. In: *Methods*. Post-transcriptional Regulation of Gene Expression 37.4, pp. 376–386. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2005.07.018. URL: http://www.sciencedirect.com/science/article/pii/S1046202305001787 (visited on 06/24/2020).

Ule, Jernej, Giovanni Stefani, et al. (Nov. 2006). "An RNA map predicting Nova-dependent splicing regulation". en. In: *Nature* 444.7119. Number: 7119 Publisher: Nature Publishing Group, pp. 580–586. ISSN: 1476-4687. DOI: 10.1038/nature05304. URL: http://www.nature.com/articles/nature05304 (visited on 06/10/2020).

Uren, Philip J. et al. (2012). "Site identification in high-throughput RNA-protein interaction data". In: *Bioinformatics* 28.23. ISBN: 1367-4811 (Electronic)\r1367-4803 (Linking), pp. 3013–3020. ISSN: 13674803. DOI: 10.1093/bioinformatics/bts569.

Valente, Louis and Kazuko Nishikura (Jan. 2005). "ADAR Gene Family and A-to-I RNA Editing: Diverse Roles in Posttranscriptional Gene Regulation". en. In: *Progress in Nucleic Acid Research and Molecular Biology*. Vol. 79. Academic Press, pp. 299–338. DOI: 10.1016/S0079-6603(04)79006-6. URL: http://www.sciencedirect.com/science/article/pii/S0079660304790066 (visited on 06/16/2020).

Van Nostrand, Eric L. et al. (June 2016). "Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)". en. In: *Nature Methods* 13.6. Number: 6 Publisher: Nature Publishing Group, pp. 508–514. ISSN: 1548-7105. DOI: 10.1038/nmeth.3810. URL: http://www.nature.com/articles/nmeth.3810 (visited on 06/11/2020).

Van Nostrand, Eric L et al. (Oct. 2019). *Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins*. en. preprint. Genomics. DOI: 10.1101/807008. URL: http://biorxiv.org/lookup/doi/10.1101/807008 (visited on 11/05/2019).

Wahl, Markus C., Cindy L. Will, and Reinhard Lührmann (Feb. 2009). "The Spliceosome: Design Principles of a Dynamic RNP Machine". English. In: *Cell* 136.4. Publisher: Elsevier, pp. 701–718. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2009.02.009. URL: https://www.cell.com/cell/abstract/S0092-8674(09)00146-9 (visited on 06/12/2020).

Wang, Tao, Guanghua Xiao, et al. (2015). "Design and bioinformatics analysis of genome-wide CLIP experiments". In: *Nucleic Acids Research* 43.11. ISBN: 1362-4962, pp. 5263–5274. ISSN: 13624962. DOI: 10.1093/nar/gkv439.

Wang, Tao, Yang Xie, and Guanghua Xiao (2014). "dCLIP: a computational approach for comparative CLIP-seq analyses". In: *Genome Biology* 15.1, R11. ISSN: 1465-6906.

DOI: 10.1186/gb-2014-15-1-r11. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4054096/ (visited on 06/10/2020).

Wang, Yang et al. (Oct. 2012). "Intronic Splicing Enhancers, Cognate Splicing Factors and Context Dependent Regulation Rules". In: *Nature structural & molecular biology* 19.10, pp. 1044–1052. ISSN: 1545-9993. DOI: 10.1038/nsmb.2377. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3753194/ (visited on 07/29/2020).

Wang, Zhen, Lionel Ballut, et al. (June 2018). "Exon Junction Complexes can have distinct functional flavours to regulate specific splicing events". en. In: *Scientific Reports* 8.1. Number: 1 Publisher: Nature Publishing Group, pp. 1–8. ISSN: 2045-2322. DOI: 10.1038/s41598-018-27826-y. URL: http://www.nature.com/articles/s41598-018-27826-y (visited on 03/18/2020).

Wang, Zhen, Valentine Murigneux, and Hervé Le Hir (2014). "Transcriptome-wide modulation of splicing by the exon junction complex". In: *Genome Biology* 15.12. ISBN: 1474-760X (Electronic)\r1474-7596 (Linking), p. 551. ISSN: 1465-6906. DOI: 10.1186/s13059-014-0551-7. URL: http://genomebiology.com/2014/15/12/551.

Weyn-Vanhentenryck, Sebastien M et al. (Mar. 2014). "HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism". In: *Cell reports* 6.6, pp. 1139–1152. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2014.02.005. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3992522/ (visited on 05/31/2020).

Will, Cindy L. and Reinhard Lührmann (July 2011). "Spliceosome Structure and Function". In: *Cold Spring Harbor Perspectives in Biology* 3.7. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a003707. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3119917/ (visited on 07/22/2020).

Xu, Mingchu et al. (Apr. 2017). "Mutations in the Spliceosome Component CWC27 Cause Retinal Degeneration with or without Additional Developmental Anomalies". In: *American Journal of Human Genetics* 100.4, pp. 592–604. ISSN: 0002-9297. DOI: 10.1016/j.ajhg.2017.02.008. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5384039/ (visited on 06/20/2020).

Yee, Brian et al. (Nov. 2018). "RBP-Maps enables robust generation of splicing regulatory maps". en. In: *RNA*. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, rna.069237.118. ISSN: 1355-8382, 1469-9001. DOI: 10.1261/rna.069237.118. URL: http://rnajournal.cshlp.org/content/early/2018/11/09/rna.069237.118 (visited on 06/09/2020).

Yun, Jonghyun, Tao Wang, and Guanghua Xiao (2014). "Bayesian hidden Markov models to identify RNA-protein interaction sites in PAR-CLIP". In: *Biometrics* 70.2. arXiv: NIHMS150003 ISBN: 0006-341x, pp. 430–440. ISSN: 15410420. DOI: 10.1111/biom.12147.

Zarnack, Kathi et al. (Jan. 2013). "Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of Alu Elements". English. In: *Cell* 152.3. Publisher: Elsevier, pp. 453–466. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2012.12.023. URL: http://www.cell.com/cell/abstract/S0092-8674(12)01545-0 (visited on 03/18/2020).

Zarnegar, Brian J et al. (June 2016). "irCLIP platform for efficient characterization of protein—RNA interactions". In: *Nature methods* 13.6, pp. 489–492. ISSN: 1548-7091. DOI: 10.1038/nmeth.3840. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5477425/ (visited on 05/31/2020).

Zhang, Chaolin and Robert B. Darnell (2011). "Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data". In: *Nature Biotechnology* 29.7. ISBN: 1087-0156, pp. 607–614. ISSN: 10870156. DOI: 10.1038/nbt.1873.

Zhang, Xiaofeng et al. (2017). "An Atomic Structure of the Human Spliceosome". In: *Cell* 169.5. Publisher: Elsevier Inc. ISBN: 1097-4172 (Electronic) 0092-8674 (Linking), 918–929.e14. ISSN: 10974172. DOI: 10.1016/j.cell.2017.04.033.

Zhang, Yong et al. (Sept. 2008). "Model-based Analysis of ChIP-Seq (MACS)". In: *Genome Biology* 9.9. Publisher: BioMed Central, R137–R137. ISSN: 1465-6906. DOI: 10.1186/gb-2008-9-9-r137. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2592715/.

Zhao, Xian-Feng et al. (Jan. 2000). "MAGOH Interacts with a Novel RNA-Binding Protein". en. In: *Genomics* 63.1, pp. 145–148. ISSN: 0888-7543. DOI: 10.1006/geno.1999.6064. URL: http://www.sciencedirect.com/science/article/pii/S0888754399960640 (visited on 06/19/2020).

Zünd, David et al. (2013). "Translation-dependent displacement of UPF1 from coding sequences causes its enrichment in 3' UTRs". In: *Nature structural & molecular biology* 20.8. Publisher: Nature Publishing Group ISBN: 1545-9985 (Electronic)\n1545-9985 (Linking), pp. 936–943. ISSN: 1545-9985. DOI: 10.1038/nsmb.2635. URL: http://dx.doi.org/10.1038/nsmb.2635.

RÉSUMÉ

La régulation post-transcriptionnelle de l'expression des gènes est un réseau d'interactions impliquant de nombreuses protéines de liaison à l'ARN et des ARN non-codants afin d'orchestrer la vie complexe des ARN messagers (ARNm). Chez les métazoaires, le complexe EJC (*Exon Junction Complex*) est un complexe multiprotéique déposé sur la jonction exonique des ARNm pendant l'épissage. L'EJC interagit avec de nombreux facteurs et est important pour le couplage fonctionnel entre l'épissage et l'export du noyau, la localisation, la traduction et la dégradation des ARNm. Malgré son rôle central dans la régulation génique et le développement de l'organisme, aucune carte exhaustive des sites de liaison de l'EJC n'a encore été établie. La méthode de CLIP (Cross-Linking and Immunoprécipitation) associée au séquençage à haut-débit (CLIP-seq) permet d'identifier les sites de liaison protéine à l'ARN in vivo. Cependant, les analyses des données de CLIP-seq ont permettent aujourd'hui d'obtenir une vue globale plutôt qu'une caractérisation individuelle des sites de liaison d'une protéine. En effet, les détecteurs de pics conventionnels appliqués aux données de CLIP de l'EJC produisent des résultats dont la reproductibilité et la sensibilité sont limitées.
Durant ma thèse, nous avons développé une stratégie dédiée à la détection du signal de l'EJC au niveau exonique. En agrégeant les informations de différents réplicas, nous avons généré une liste de gènes reproductibles. Au sein de ces gènes, nous avons trouvé une forte corrélation entre la robustesse de détection des exons et le contenu en thymidine (T) au niveau des sites de liaison. Posant l'hypothèse que ceci est un effet du photopontage, nous avons corrigé le score de robustesse par le contenu en T et avons ainsi clairement montré que l'EJC est déposé sur certains exons et pas sur d'autres. Par conséquent, le complexe EJC est déposé de manière différentielle le long d'un même transcrit. Nous avons ainsi établi une carte des sites de liaisons de l'EJC sans précédent. L'intégration de données supplémentaires a montré que le dépôt de l'EJC est indépendant de l'abondance du transcrit et n'est pas expliqué par des annotations fonctionnelles connues du gène. Bien que ce travail n'a pas permis à ce stade d'identifier les raisons de ce dépôt différentiel, nous présentons une première méthode d'analyse spécifique et reproductible des exons liés à un EJC par CLIP-seq.
Les deux contributions principales de ce travail sont donc les suivantes. Premièrement, nous proposons une méthode robuste pour détecter l'enrichissement du signal de l'EJC à l'échelle de l'exon, en démontrant quantitativement que celle-ci est plus reproductible et plus sensible que les solutions offertes par les outils actuels. Deuxièmement, nous prouvons que, au sein d'un même transcrit, l'EJC peut être présent sur des exons, et absent d'autres, suggérant que le dépôt de l'EJC est un processus régulé suivant un code qui reste à découvrir.

MOTS CLÉS

régulation post-transcriptionnelle, transcriptomique, intéraction protéine-ARN, bioinformatique

ABSTRACT

Post-transcriptional Gene Expression Regulation is a complex network that involves RNA-binding proteins and non-coding RNAs to orchestrate the complex life of mRNAs. In metazoans, the Exon Junction Complex (EJC) is a multi-protein complex deposited onto mRNAs exon junctions during splicing. The EJC interacts with numerous factors and is important for coupling pre-mRNA splicing with mRNA nuclear export, localization, translation, and decay. Despite its central role in gene expression and in organism development, the comprehensive map of EJC binding sites is lacking. Crosslinking and immunoprecipitation coupled with high-throughput sequencing (CLIP-seq) aims to identify transcriptome-wide RNA-protein interactions in vivo. Yet, current trends in CLIP-seq data analysis gravitate towards painting a global landscape rather than characterizing individual binding sites. However, we observed that current peak callers applied to EJC CLIP data yield results with limited reproducibility and sensibility.
During my PhD, we developed a dedicated strategy to detect EJC signal enrichment at the exon level. By aggregating data from several replicates, we built a list of robust genes with reproducible EJC loading rate. Within robust genes, we assigned a robustness score to each exon according to frequency of detection across replicates. We found that the exon robustness score was correlated to the thymidine (T) content of EJC binding sites. Assuming this was due to cross-linking chemistry, we corrected the score for the T content and found exons with either high or low detection rates. The last suggests that EJC loading is not homogeneous along a transcript, but rather differential. Thus, we established an unprecedented binding site map of the EJC in living cells validated by statistical tools. Crossing this map with other information showed that EJC loading is independent of transcript expression levels or known gene functional annotations. Although the scope of this work does not include possible explanations for this differential loading, it presents a first reproducible and specific data analysis pipeline to detect EJC-loaded exons.
Altogether, our contribution is twofold. First, we proposed a robust way to detect EJC signal enrichment at the exon level and demonstrated quantitatively that our approach is more reproducible and more sensitive compared to conventional tools. Second, we proved that the EJC can be present on some, and absent on other exons of the same transcript suggesting that EJC loading is a regulated process following a code that remains to be discovered.

KEYWORDS

post-transcriptional regulation, transcriptomics, protein-RNA interactions, bioinformatics