Biais de composition nucléotidique des gènes et épissage alternatif

Sébastien Lemaire


Numéro National de Thèse : 2019LYSEN005

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée par
l'Ecole Normale Supérieure de Lyon

Ecole Doctorale N° 340
Biologie Moléculaire Intégrative et Cellulaire

Spécialité de doctorat : Bio-informatique
Discipline : Biologie

Soutenue publiquement le 15/03/2019, par :
Sébastien LEMAIRE

Biais de composition nucléotidique des gènes et épissage alternatif

---

Devant le jury composé de :

VIEIRA-HEDDI Christina	Pr. d'Univ.	Université Lyon 1	Examinatrice
FERNADEZ DE LUCO Reini	CR	Institut de Génétique Humaine	Rapporteur
DUTERTRE Martin	CR	Institut Curie	Rapporteur
WEIL Dominique	DR	Institut de Biologie Paris Seine	Examinatrice
AUBOEUF Didier	DR	ENS de Lyon	Directeur de thèse

Table des matières

Résumé

L’épissage, une étape majeure de l’expression des gènes, consiste en l’élimination des introns et la production de transcrits matures ou ARNm. La régulation ou des perturbations de l’épissage sont impliquées dans de nombreuses situations physiopathologiques. Dans ce travail, j’ai utilisé et analysé par des approches de bio-informatiques un grand nombre de données générées à large échelle afin de mieux définir les règles gouvernant la reconnaissance des exons au cours de l’épissage. Je montre que les mécanismes de reconnaissance des exons dépendent du biais de la composition nucléotidique des gènes qui les hébergent. Ainsi, la reconnaissance des exons hébergés par des gènes enrichis en guanine et cytosine dépend essentiellement de leur site 5’ d’épissage qui peut être masqué par des structures secondaires. La reconnaissance des exons hébergés par des gènes enrichis en thymine et adénine dépend essentiellement des signaux d’épissage situés en amont des exons. Je montre également que l’organisation chromatinienne est différente selon les biais de composition nucléotidique des gènes et que cela a un impact spécifique sur la reconnaissance des exons. De nombreuses études démontrent que les gènes ne sont pas organisés de façon aléatoire dans un génome et que l’architecture des gènes et des chromosomes dépend de leur composition nucléotidique. Par conséquent, mes travaux suggèrent qu’il existe un lien direct entre composition nucléotidique d’une région du génome, architecture de la chromatine et sélection des exons au cours de l’épissage.

Introduction

1. Epissage des ARN pré-messagers

L'expression d'un gène codant se divise en deux étapes consécutives que sont la transcription et la traduction. La transcription est la synthèse par les ARN polymérases d'un ARN qui est la copie de l'ADN. Certains ARNs, les ARNm sont traduits en protéines. Chez l'homme, la majorité des gènes codants (95%) sont constitués d'exons et d'introns, ces derniers étant éliminés lors de l'épissage (Berk, 2016). En moyenne, chaque gène codant contient de 10 à 11 introns.

La définition des introns repose sur des séquences situées à leurs extrémités. Les introns commencent la plupart du temps par le dinucléotide "GU" qui définit le site d'épissage 5' (ou 5'SS) et se terminent par trois éléments : un "point de branchement" (BP) qui est souvent une adénine, une séquence en aval enrichie en base de type pyrimidine ou "poly-pyrimidine tract" (PPT) et le site d'épissage en 3' (ou 3'SS) qui correspond la plupart du temps au dinucléotide "AG". La réaction d'épissage est une double trans-estérification (Hang et al., 2015). La première trans-estérification raccorde le 5'SS au BP et la seconde raccorde les deux exons consécutifs (Fig.1). L'intron est ensuite dégradé alors que l'ARNm contenant les exons est exporté dans le cytoplasme où il est traduit.

Tous les éléments décrits ci-dessus sont nécessaires pour permettre l'assemblage du spliceosome qui est le complexe réalisant la réaction d'épissage (Matera and Wang, 2014). Le spliceosome est composé de petits ARNs (les snRNAs) et de protéines associées aux snRNAs (Hang et al., 2015). Ces protéines et les snRNAs composent les snRNPs. Le 5'SS est reconnu par la U1 snRNP et nécessite l'hybridation du U1 snRNA sur le pré-ARNm. Cette hybridation peut être aidée par des

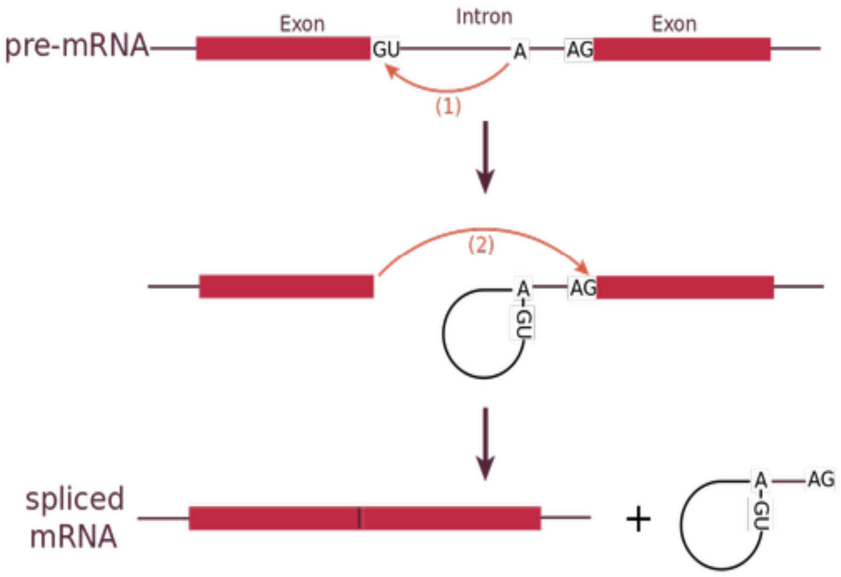


Figure 1 : Mécanisme d'épissage d'un intron. Une première trans-estérification relie le 5'SS (GU) au BP (A). La seconde trans-estérification relie les extrémités des deux exons.

protéines comme U1-70k, une protéine du U1 snRNP. Celle-ci interagit avec la tige-boucle I du U1 snRNA. A l'extrémité 3' de l'intron, la protéine SF1 se fixe sur des motifs UNA qui est par ailleurs contenue dans la séquence définissant le BP. La protéine U2AF65 qui se fixe sur des motifs riches en U, interagit avec le PPT en aval du BP. La protéine U2AF35 quant à elle, reconnait le 3'SS. SF1, U2AF65 et U2AF35 aident au recrutement de la U2 snRNP composée d'un snRNA, U2 snRNA qui s'hybride sur le BP. Lorsque les U1 et U2 snRNPs sont fixées sur l'ARN, la tri-snRNP U4/U6.U5 est recrutée. Ensuite, U1 snRNP et U4 snRNP sont détachées et ne restent que les snRNPs U2/U5/U6 qui rapproche le 5'SS, le BP et le 3'SS (Fig.2) (Lee and Rio, 2015). La double trans-estérification décrite ci-dessus est catalysée par ce complexe.

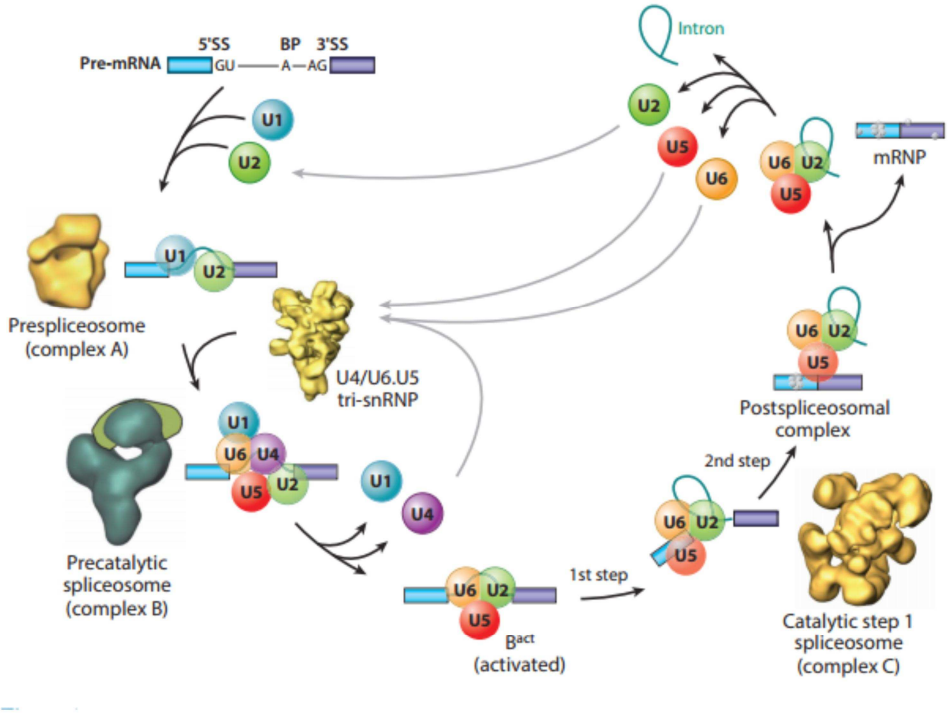


Figure 2 : Cycle d'assemblage et désassemblage du spliceosome lors de l'épissage d'un intron. (Lee and Rio, 2015)

Il a été proposé que les introns et les exons peuvent être détectés de deux façons différentes selon la taille des introns encadrant les exons (Berget, 1995; Conti et al., 2013; Fox-Walsh et al., 2005; Hertel, 2008; Talerico and Berget, 1994). Chez l'humain, les gènes contiennent de grands introns faisant de 1 à 4 kb (mais peuvent atteindre plus de 100 kb), tandis que les exons ont une taille moyenne de 150 b (entre 50 et 250). Il a été proposé que dans cette configuration, le recrutement des U1 et U2 snRNPs s'opèrent de façon synergique au travers de l'exon (Berget, 1995; Conti et al., 2013; Hertel, 2008; Robberson et al., 1990; Sterner et al., 1996). Dans ce modèle, appelé « exon definition » des protéines se liant aux exons des ARN pré-messagers, comme les facteurs d'épissage de la famille SR (voir ci-dessous), permettraient le recrutement des U2 et U1 snRNPs de part et d'autre de l'exon (Fig.3a). Un autre mode d'épissage, dit « intron definition », a été proposé lorsque les exons sont

encadrés de petits introns (<300 bp) (Conti et al., 2013; Sterner et al., 1996; Talerico and Berget, 1994). Dans cette configuration, les sites 5’ et 3’ aux extrémités des introns seraient suffisamment proche pour permettre le recrutement plus ou moins simultanément des U1 et U2 snRNPs à chaque extrémité des introns à éliminer (Fig.3b).

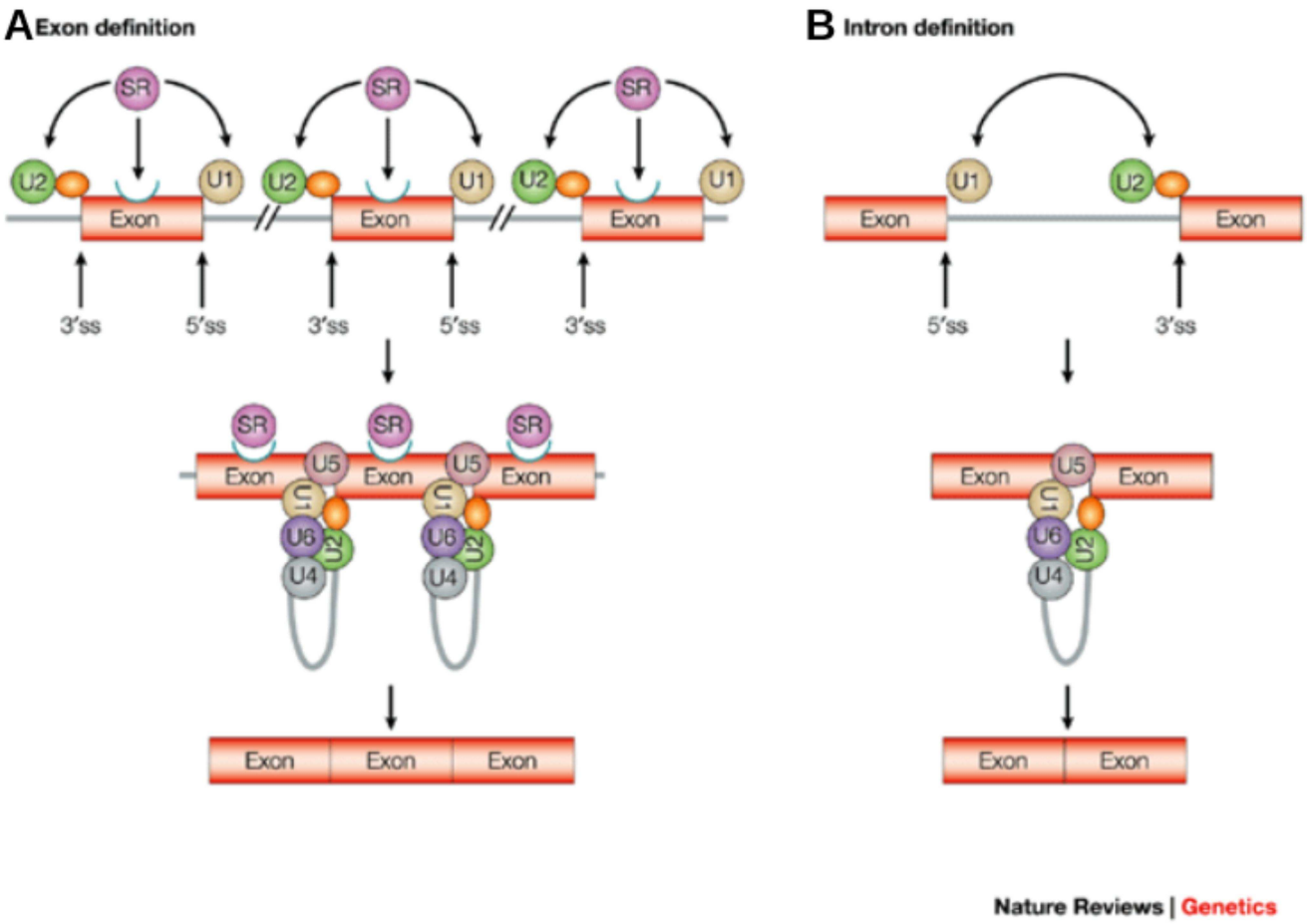


Figure 3 : A, Modèle « Exon definition » : des protéines de fixation à l’ARN telles que les protéines SRs se fixent sur l’exon, recrutant U1 snRNP et U2AF65/U2AF35 recrutant à leur tour U2 snRNP. B. Modèle « Intron definition » : la fixation de U1 snRNP au 5’SS, et celle de U2AF et U2 snRNP au PPT et BP, respectivement, en aval sur le même intron. La paire de sites d’épissage définit l’intron qui sera épissé. (Ast, 2004)

2. Epissage alternatif

Les séquences qui définissent les frontières entre introns et exons sont composées de quelques nucléotides et sont souvent présentes en très grand nombre dans chaque intron. Par ailleurs, ces séquences peuvent être dégénérées, c'est-à-dire plus ou moins divergentes de séquences consensuelles reconnues par les snRNAs ou les protéines associées (Gao et al., 2008; Taggart et al., 2017; Tan et al., 2016). Ces caractéristiques sont au centre de la notion d'épissage alternatif. En effet, si un site d'épissage diffère des séquences « canoniques », il est alors dit « faible » et sera difficilement détecté par le spliceosome. Ainsi, des exons flanqués de sites d'épissage faibles ont une plus grande probabilité d'être éliminés avec leurs introns flanquant. Par conséquent, un exon peut être exclu au moment de l'épissage et donc être absent du messager mature. Comme il le sera décrit ci-dessous, différents paramètres peuvent moduler la reconnaissance des exons, et conduisent à l'épissage alternatif, c'est-à-dire à la reconnaissance ou non, de façon régulée, d'exons ou de parties d'exons (Fig.4).

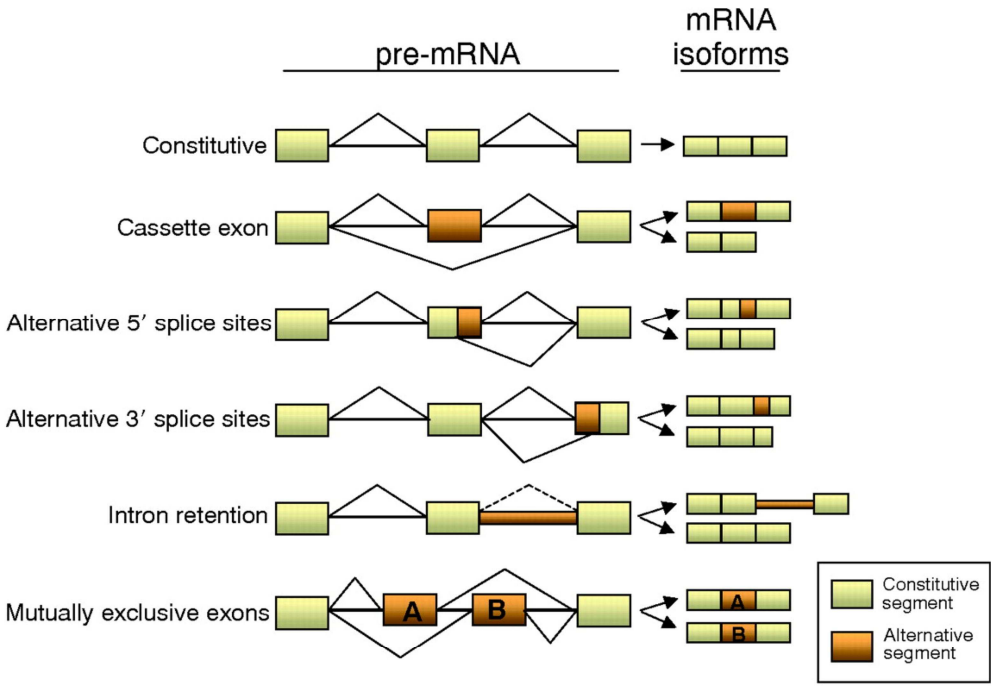


Figure 4 : Différents types d'évènement d'épissage alternatif (Srebrow and Kornblihtt, 2006)

L’épissage alternatif est la règle puisque 95% des gènes humains codants contenant plusieurs exons sont soumis à l’épissage alternatif (Pan et al., 2008; Wang et al., 2008). Ces gènes peuvent donc générer des ARNm différents (Fig.5). Les ARNm peuvent différer dans leur séquence codante ou dans les régions non-traduites. La stabilité de l’ARNm, sa localisation ou sa traduction peuvent en être alors affectées. Il a été estimé qu’au moins 37 % des ~20.000 gènes codants humains produit plusieurs isoformes protéiques (Kim et al., 2014).

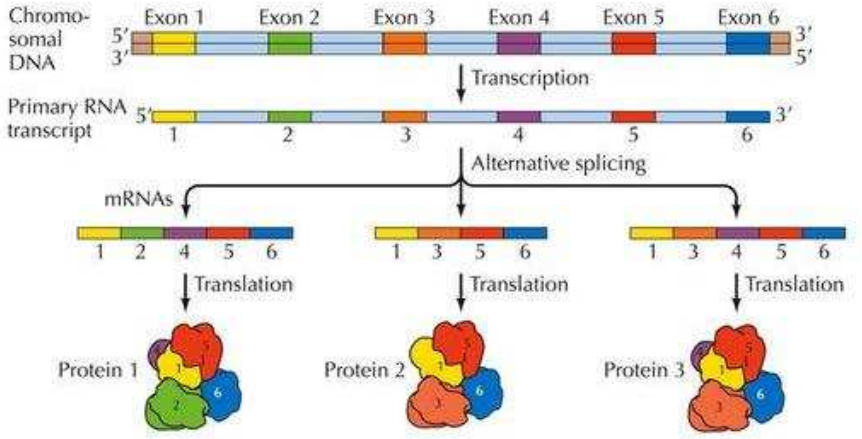


Figure 5 : Schéma représentant l’épissage alternatif. Plusieurs combinaisons d’exons peuvent former l’ARNm. A gauche, tous les exons sont inclus sauf l’exon 3, au centre, les exons 2 et 4 ne sont pas inclus, à droite, l’exon 2 n’est pas inclus. Il en résulte autant d’isoformes protéiques que d’ARNm différents. (« THE CELL, Fourth Edition »)

Ce mécanisme est un niveau de régulation biologiquement important car il est impliqué dans la différentiation des cellules au cours du développement (Baralle and Giudice, 2017; Furlanis and Scheiffele, 2018). Par exemple, l’inclusion de l’exon EDA du gène FN1 contribue à l’intégration de la protéine correspondante dans la matrice extracellulaire dans les tissues conjonctifs. L’élimination de l’exon EDA conduit à la synthèse d’une isoforme protéique qui ne s’intègre pas dans la matrice extracellulaire mais qui est soluble dans le plasma sanguin. Les cellules hépatiques sécrètent cette

isoforme dans le sang (Fig.6a) (Baralle and Giudice, 2017). L'exon 10 de EHMT2 est un autre exemple montrant l'importance de l'épissage dans les phénomènes de différentiation (Fig.6b). Il a en effet été montré que l'isoforme protéique EHMT2 ne contenant pas l'exon 10 est exprimée dans les progéniteurs neuronaux et est majoritairement cytoplasmique. De ce fait, cette isoforme n'assure pas une fonction bien caractérisée de EHMT2 qui est la méthylation de la lysine 9 de l'histone 3 (H3K9). Lors de la différenciation des progéniteurs en neurones, l'exon 10 est inclus et la protéine EHMT2 produite est nucléaire et méthyle H3K9. L'inclusion de l'exon 10 permettrait la répression de gènes qui maintiennent l'état de progéniteur et faciliterait ainsi la différenciation en neurone (Fiszbein et al., 2016). Enfin, soulignant la pertinence biologique de ce processus, la dérégulation de l'épissage ou de

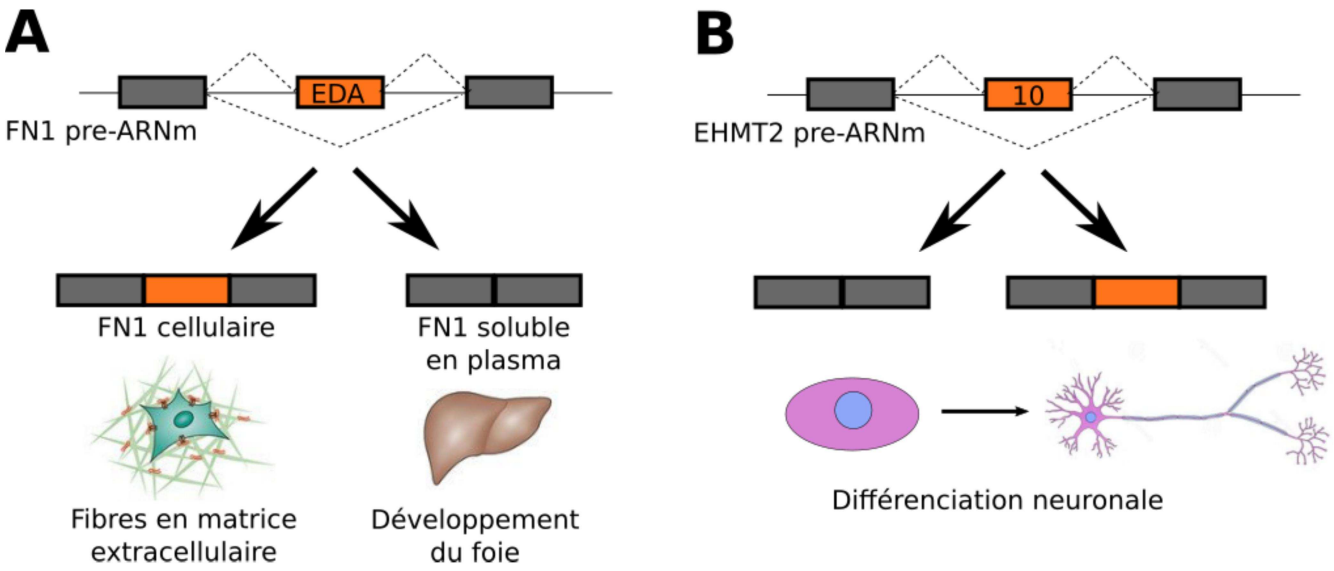


Figure 6 : A. L'exon EDA (orange) du gène codant la fibronectine (FN1) est un exon cassette différentiellement inclus en comparant des fibroblastes et des cellules hépatiques. Sa présence joue un rôle dans l'intégration de la fibronectine dans la matrice extracellulaire [Baralle 2017]. B. Le taux d'inclusion de l'exon 10 de EHMT2 augmente au cours de la différentiation des progéniteurs en neurones. Cet exon complète le signal NLS d'adressage au noyau de EHMT2, permettant à cette protéine de méthyler les lysine 9 de l'histone 3 dans le noyau (Fiszbein et al., 2016).

l'épissage alternatif est impliquée dans de nombreuses maladies, y compris le cancer (Cáceres and Kornblihtt, 2002; Scotti and Swanson, 2016 ; Srebrow and Kornblihtt, 2006).

3. Mécanismes de régulation de l'épissage alternatif au niveau de l'ARN

Les exemples décrits ci-dessus soulignent l'importance de l'épissage alternatif en termes de conséquences fonctionnelles. Dans ce contexte, il est clairement établi que ce phénomène est régulé à de multiples niveaux.

3.1. Facteurs d'épissage

De multiples protéines intéragissent avec le pre-ARNm. Cette intéraction peut passer par des séquences consensuelles de fixation que des manipulations *in vitro* sur des exons cibles ou des expériences de CLIP ont permises d'établir. Probablement plus d'une centaine de protéines intéragissant avec le pre-ARNm (en plus des protéines directement associées au spliceosome) contribuent à réguler l'épissage alternatif. Ces protéines sont des protéines qui se fixent sur les ARNs au niveau de séquences régulatrices de l'épissage (ou SREs) (Baralle et al., 2006; Wang and Burge, 2008; Wang et al., 2012). Les modes d'action de ces protéines sont multiples. Certaines protéines facilitent le recrutement des snRNPs aux sites d'épissage alors que d'autres peuvent masquer ces sites d'épissage. On peut regrouper ces facteurs d'épissage en trois ensembles : les protéines SRs, les hnRNPs, et les autres facteurs.

Les protéines SRs ou SRSFs correspondent à une douzaine de protéines qui partagent une même organisation en domaines et se fixant sur des SREs via des domaines RRMs (Änkö, 2014; Graveley, 2000; Kanopka et al., 1996; Sahebi et al., 2016). Les SRSFs peuvent favoriser la détection

des sites d'épissage en se fixant sur une SRE au sein d'un exon (Graveley, 2000). Ces facteurs peuvent favoriser le recrutement d'éléments du spliceosome au travers d'interactions avec U1-70K et U2AF35, par exemple. Ceci augmente alors la probabilité de détection des exons auxquels les protéines SRSFs se fixent. Par exemple, l'exon EDA de la fibronectine décrit ci-dessus est activé par des SRSFs se fixant sur les exons (Lavigueur et al., 1993). Certaines études ont également montré que la fixation des protéines SRSFs sur des séquences introniques peut réprimer la détection d'un site d'épissage proximal. Par exemple, l'épissage du pre-ARNm de l'adénovirus L1 peut produire deux ARNm contenant ou pas l'exon IIIa, qui dépend de la détection du 3'SS. En amont, se trouve une région contenant trois sites de fixation pour des SRSFs. Lorsque les SRSFs se fixent sur ces SREs, ils empêchent le recrutement de U2 snRNP et donc la détection du 3'SS de l'exon IIIa (Kanopka et al., 1996).

Les hnRNPs forment un autre ensemble de protéines qui se fixent à l'ARN via des domaines RRM, qRRM, KH, ou des boîtes RGG (Geuens et al., 2016; Han et al., 2010). L'action des hnRNPs sur la régulation de l'épissage est souvent différente de celle des SRSFs. Ainsi, contrairement aux SRSFs, les hnRNPs répriment souvent l'inclusion d'exons auxquels ils se fixent. Par exemple, les protéines hnRNPA1/A2 se fixent sur l'exon 18 de BRCA1, causant l'exclusion de cet exon (Goina et al., 2008). Par ailleurs, la fixation de protéines hnRNPs au sein d'un intron peut activer l'inclusion de l'exon adjacent (Expert-Bezançon et al., 2002). Par exemple, hnRNPK reconnaît une séquence intronique du pré-ARNm de la β-Tropomyosine activant l'inclusion de l'exon 6. Toutefois, ces règles générales sont à prendre avec précaution car il existe des exemples contradictoires. Par exemple, hnRNPA1/A2 est un répresseur de l'exon 7 de SMN2 lorsqu'il se fixe sur des séquences introniques (Hua et al., 2008).

Les facteurs d'épissage appartenant aux familles décrites ci-dessus, ont, comme l'ensemble des gènes codant, leur expression régulée. Ainsi, différents types cellulaires ou des cellules exposées à différentes variations de leur environnement expriment différents facteurs d'épissage. La régulation de l'expression des facteurs d'épissage par des facteurs de transcription est un niveau supplémentaire de régulation de l'épissage. Par exemple, l'épissage du gène PKM, codant pour la « pyruvate kinase » est régulé par les facteurs d'épissage hnRNPA1 et A2 dont l'expression est régulée par le facteur de transcription c-Myc (David et al., 2010). De plus, les facteurs d'épissage sont fréquemment impliqués dans des boucles de rétrocontrôle qui les régulent (Dardenne et al., 2014; Lareau and Brenner, 2015; Nasim et al., 2002). Par exemple, les facteurs d'épissage DDX5 et DDX17 sont régulés par une boucle de rétro-contrôle au cours de la différenciation musculaire. En effet, il a été montré que DDX5 et DDX17 contrôlent l'activité transcriptionelle du facteur de différenciation Myod, qui activent l'expression des micro-ARN miR-1 et miR-206, qui eux-mêmes répriment l'expression de DDX5 et DDX17 (Dardenne et al., 2014). Il a également été montré que le facteur d'épissage hnRNPA1 se régule par rétrocontrôle négatif en inhibant l'épissage de l'intron 10 de son propre pré-ARNm. L'absence de ce rétrocontrôle conduirait à la surexpression de hnRNPA1, et à la mort cellulaire par apoptose (Suzuki and Matsuoka, 2017).

3.2. Structures secondaires des ARNs

Au fur et à mesure qu'un pré-ARNm émerge de l'ARN polymérase II, celui-ci interagit avec des protéines et/ou des interactions entre les nucléotides le composant conduisent à la formation de structures secondaires (Fig.7). Ces structures secondaires peuvent interférer avec la régulation de l'épissage. En effet, des éléments dans la séquence du pré-ARNm, comme les sites d'épissage ou les

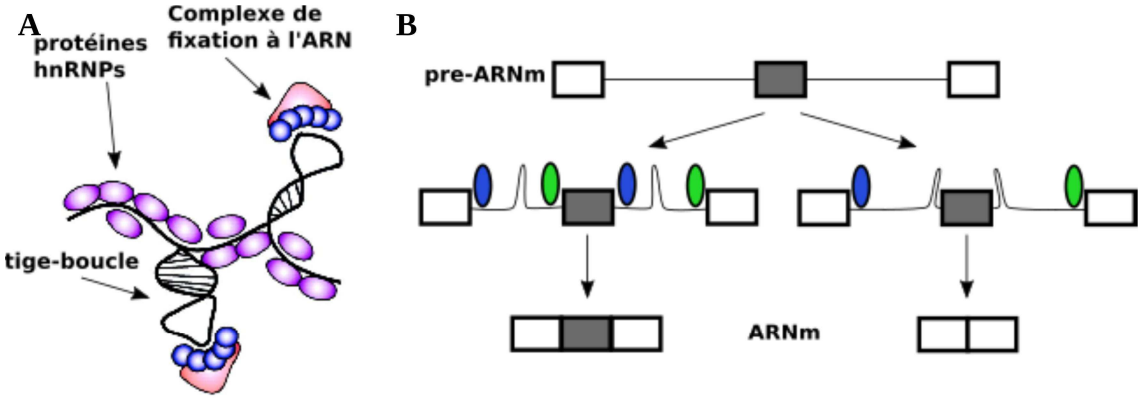


Figure 7 : A. Interactions de l'ARN pré-messager avec des protéines et formation de structures secondaires de type tige-bouche (Buratti and Baralle, 2004). B. La structure secondaire de l'ARN est favorable à l'épissage d'un intron si elle rapproche les sites d'épissage, ou inhibe l'épissage en séquestrant l'un des sites d'épissage, c'est-à-dire en empêchant leur reconnaissance par des protéines associées au spliceosome (cercles vert et bleu).

SREs, qui seraient séparés par une grande séquence peuvent être rapprochés dans l'espace grâce à la formation de structures secondaires. Dans le cas des sites d'épissage, leur rapprochement favorise l'assemblage du spliceosome sur le pré-ARNm. Notamment, une étude (Pervouchine et al., 2012) a relevé que de long introns peuvent avoir à leurs extrémités deux séquences complémentaires. En s'hybridant ces séquences complémentaires rapprocheraient les sites d'épissage. Cependant, des structures secondaires peuvent, à l'inverse, séquestrer des SREs, des sites d'épissages ou des exons entiers. Dans ce cas, les structures secondaires peuvent inhiber la reconnaissance d'un exon. Par exemple, il a été montré qu'une SRE située dans l'intron en aval de l'exon IIIb de FGFR2 en inhibe la détection (Baraniak et al., 2003). Cependant, des séquences complémentaires présentes dans le même intron forment une tige-boucle qui rapproche un autre SRE du premier et qui l'inhibe. L'exon IIIb est alors inclus (Fig.8).

Une tige-boucle peut réprimer la détection d'un intron en séquestrant un site d'épissage. Par exemple, le 5'SS de l'exon 10 du gène Tau est séquestré au sein d'une tige-boucle. Celle-ci empêche

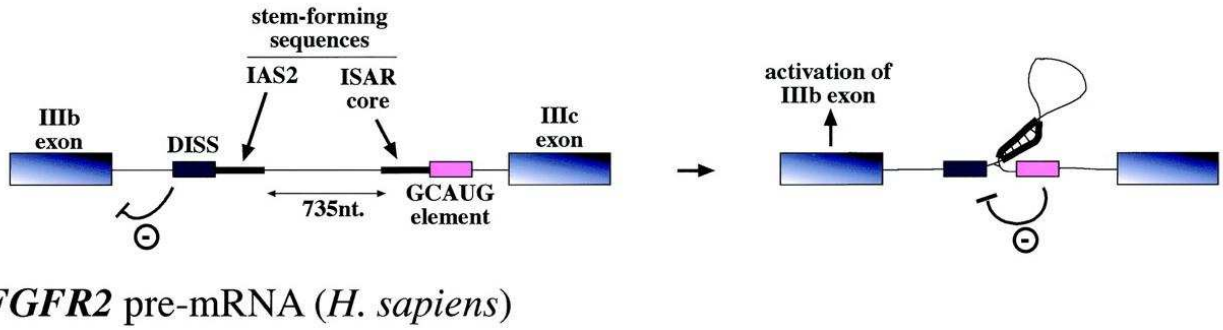


Figure 8 : Interaction entre deux SREs via leur rapprochement par la formation d'une tige-boucle. Les séquences IAS2 et ISAR (lignes noires épaisses) forment une tige boucle qui rapproche l'élément DISS (rectangle noir) de l'élément GCAUG (rectangle rose), résultant en l'activation de l'exon IIIb de FGFR2 (Buratti and Baralle, 2004).

l'hybridation de U1 snRNA sur le 5'SS et donc induit l'exclusion de l'exon (Fig.9). Le recrutement de l'hélicase à ARN DDX5 (ou « p68 ») par RBM4 permet de déstructurer la tige-boucle séquestrant le 5'SS, qui peut alors s'hybrider avec U1 snRNA (Kar et al., 2011; Liu, 2002).

Le pré-ARNm peut également former des structures secondaires via son interaction avec des facteurs d'épissage. Par exemple, la fixation de U2AF65 en elle-même forcerait le PPT à se courber (Kent et al., 2003), rapprochant le BP avec le 3'SS qui sont de part et d'autre (Fig.10). Ou encore, les introns flanquant l'exon 7B de hnRNPA1-like contiennent chacun un SRE de fixation de hnRNPA1. Deux hnRNPA1 peuvent former un dimère qui, dans le cas de l'exon 7B, fait le pontage physique entre les deux introns. Une boucle est ainsi formée et séquestre l'exon qui n'est alors pas détecté (Fig.11) (Nasim et al., 2002).

Le pre-ARNm peut aussi former des structures secondaires comme les G-quadruplexes et les I-motifs (Fig.12) qui n'impliquent pas une hybridation canonique entre nucléotides (Millevoi et al., 2012;

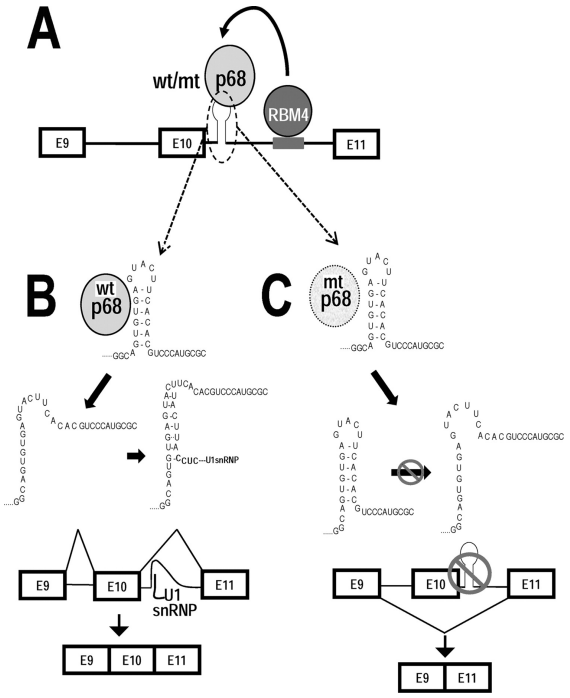


Figure 9 : Séquestration d'un site d'épissage par une tige-boucle inhibant l'épissage. Une tige-boucle est en compétition avec U1 snRNP pour l'interaction avec le 5'SS de l'exon 10 de Tau. L'hélicase à ARN p68 (ou DDX5), en déstructurant la tige-boucle, favorise l'interaction U1 snRNA avec le 5'ss et donc la reconnaissance de l'exon (Kar et al., 2011).

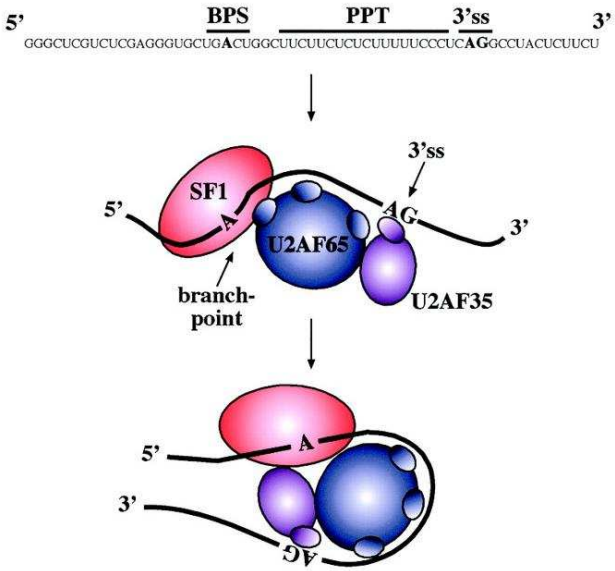


Figure 10 : Conformation de l'ARN induite par son interaction avec une protéine. La figure montre comment l'interaction d'un ARN avec les protéines UAF35, U2AF65 et SF1 peut induire sa courbure (Buratti and Baralle, 2004).

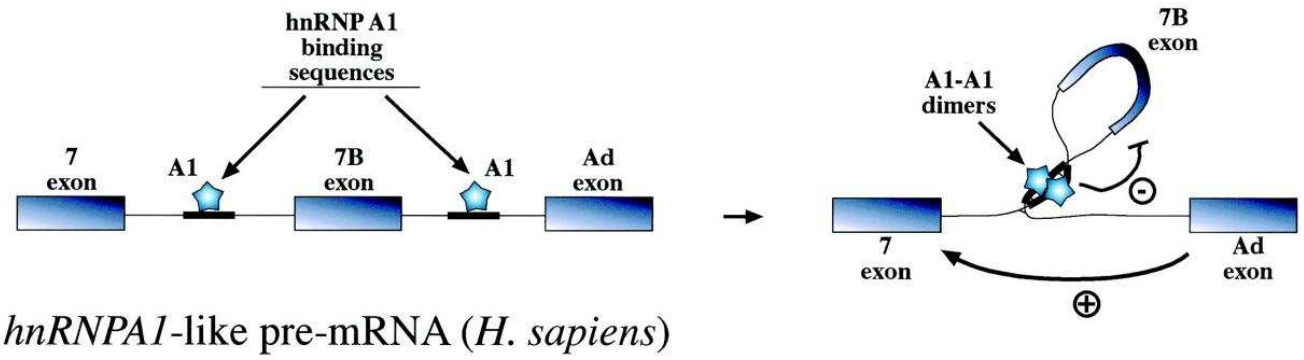


Figure 11 : Le pré-ARNm du gène hnRNPA1-like forme des boucles par interaction avec un dimère de facteurs d'épissage (A1-A1), ce qui a des conséquences sur l'épissage de l'exon 7B (Buratti and Baralle, 2004).

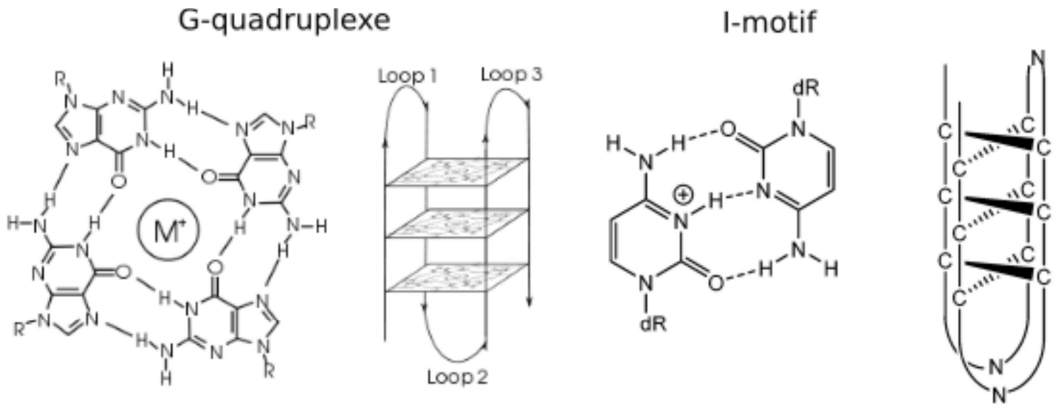


Figure 12 : Structures dites « G-quadruplex » (gauche) et « I-motif » (droite). Le G-quadruplex consiste en des « plateaux » de 4 guanines, formés par des interactions non-Watson/Crick entre guanines et stabilisés par un ion métallique comme le magnésium (Huppert and Balasubramanian, 2005). Le I-motif est formé par un empilement de paires de cytosines (Gurung et al., 2015).

Snoussi et al., 2001; Song et al., 2016). L'impact de ces structures secondaires sur la régulation de l'épissage a été mis en évidence notamment en étudiant les exons régulés par hnRNPF à l'échelle du génome (Huang et al., 2017). Des centaines d'exons sont régulés par hnRNPF qui se fixe sur une séquence pouvant former un G-quadruplexe. Dans cette étude, il est montré que la formation du G-quadruplexe, empêche la fixation de hnRNPF sur les pré-ARNm, modifiant l'épissage des exons à proximité.

3.3. Modifications chimiques des ARNs

Il est établi depuis longtemps que les nucléotides composant les ARNs peuvent être modifiés biochimiquement (Gilbert et al., 2016; Liu et al., 2014) . Par exemple, une base uridine peut être modifiée en pseudouridine, une base adénine peut être modifiée en inosine. Par ailleurs, les bases adénine et cytosine peuvent être méthylées. Il a longtemps été postulé que ces modifications impactaient essentiellement des ARN non-codants, comme les tRNAs ou les rRNAS. Cependant, des études récentes montrent que les pré-ARNm sont souvent modifiés par « edition » (modification d'une base Adenine en une base inosine) ou méthylation de l'adénosine (m6A). De façon très intéressante, ces modifications chimiques peuvent avoir un impact sur l'épissage. Un des mécanismes proposé est que la modification chimique d'une base (par exemple sa méthylation) peut modifier la fixation de protéines et/ou la formation de structures secondaires des ARNs (Liu et al., 2015; Ping et al., 2014; Zhao et al., 2014).

Ainsi, Liu *et al.* ont étudié un site de fixation de hnRNPC, une suite de 5 uracyles, sur le pré-ARNm MALAT1 (Liu et al., 2015). Ce site de fixation peut former une tige-boucle avec une séquence complémentaire « RRACH » en aval. Si la tige-boucle se forme, hnRNPC ne peut se fixer sur son site de fixation. La méthylation de l'adénine présente dans la séquence complémentaire déstructure la tige-boucle, rétablissant la fixation de hnRNPC sur son site de fixation. Dans cette étude, il est évalué que la fixation de hnRNPC au niveau de ~2800 sites différents, principalement introniques, dépend de la méthylation d'adénines.

Dans une autre étude, plus de 2000 exons ont été montrés comme étant régulés à la fois par SRSF3 et SRSF10 (Xiao et al., 2016). Le premier permet l'inclusion de l'exon dans l'ARNm tandis

que le deuxième le réprime. La compétition entre les deux facteurs dépend de la méthylation d'une adénine (m6A) au sein des exons. En effet, lorsque l'adénine est méthylée, elle recrute YTHDC1 qui favorise le recrutement de SRSF3 permettant l'inclusion de l'exon (Fig.13, partie gauche). En absence de méthylation, SRSF10 se fixe à l'exon et en réprime l'inclusion (Fig.13, partie droite).

En conclusion, les mécanismes, au niveau de l'ARN, de régulation de l'épissage alternatif sont très divers. Ils impliquent la fixation de facteurs d'épissage, la formation de structures secondaires, ou

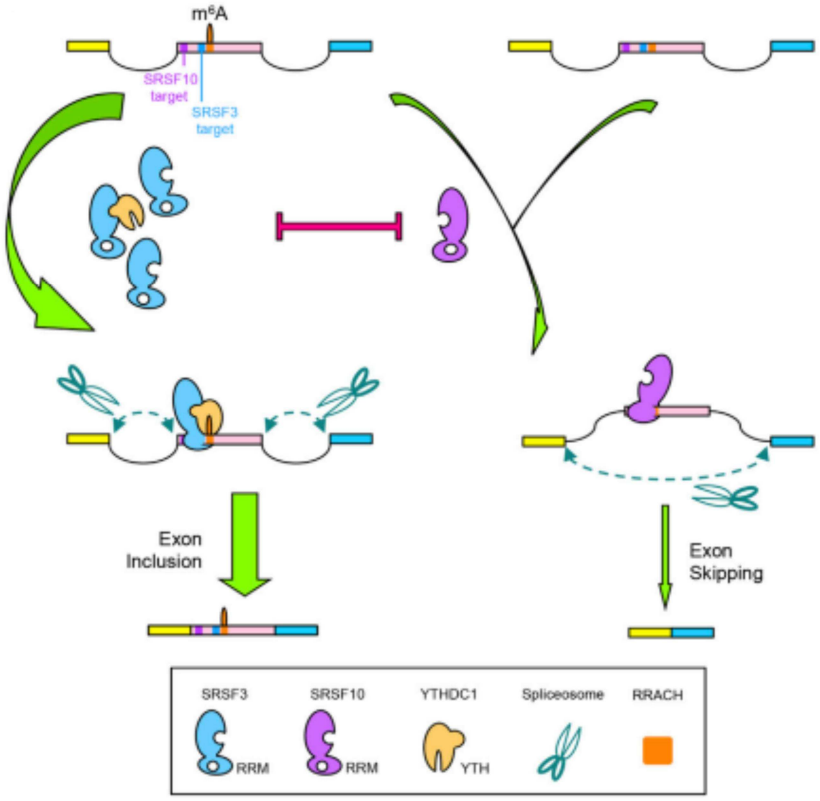


Figure 13 : La méthylation d'une adenine (m6A) dans un exon impacte la régulation d'épissage. La méthylation de l'adénine permet de recruter SRSF3 (particule bleu claire), via la protéine YTHDC1 (particule marron clair), ce qui favorise la détection de l'exon et son inclusion (partie gauche). L'absence de méthylation favorise l'interaction de SRSF10 (particule violette à droite) sur l'exon, ce qui réduit la détection de l'exon (Xiao et al., 2016).

des modifications chimiques des nucléotides. L'épissage alternatif peut aussi être régulé par des mécanismes associés à l'ADN et la chromatine.

4. Mécanismes de régulation de l'épissage alternatif : relation chromatine, transcription et épissage

4.1. Couplage transcription et épissage

Il est maintenant clairement établi que la majorité des introns sont épissés au cours de la transcription (Fig.14) (Carrillo Oesterreich et al., 2016; Herzel et al., 2017; Luco et al., 2011). Le fait que l'épissage se déroule pendant la transcription a des conséquences sur la reconnaissance des exons et des introns.

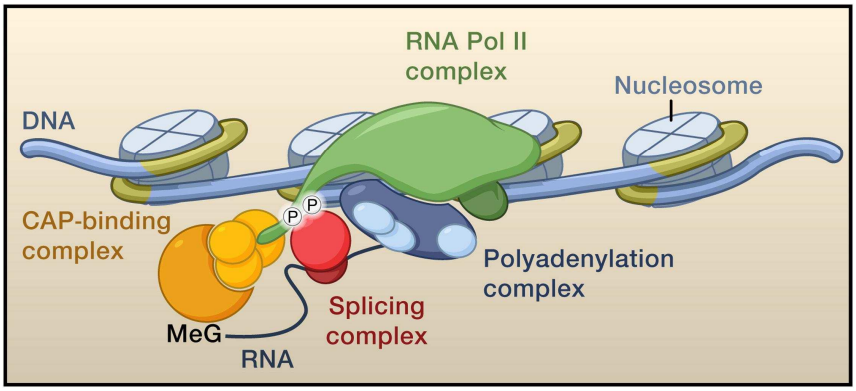


Figure 14 : Couplage entre la transcription et la maturation du pré-ARNm. La RNAPII (en vert) recrute des facteurs impliqués dans la modifications des ARNs naissants, comme le complexe ajoutant la coiffe (en jaune), le spliceosome (en rouge), et le complexe de poly-adénylation (en bleu) (Luco et al., 2011).

Tout d'abord, la machinerie transcriptionnelle et notamment la RNAPII peut permettre le recrutement, c'est-à-dire l'enrichissement local de facteurs d'épissage (Misteli and Spector, 1999; Morris and Greenleaf, 2000). Par exemple, il a été montré que la nature du promoteur placé en amont de gènes rapporteurs peut avoir un effet sur l'épissage (Cramer et al., 1999). Ainsi le promoteur MMTV contenant des éléments de réponse à la progestérone placé en amont des exons variables 4 et 5 du gène CD44 permet une variation d'épissage de ces exons après traitement à la progestérone alors qu'un promoteur contrôle (HSV) qui n'a pas d'éléments de réponse à la progesterone n'est pas associé à ces variations d'épissage (Auboeuf et al., 2002). Un autre exemple est la régulation de l'exon EDI par le promoteur de la fibronectine. Une mutation dans le promoteur favorise le recrutement du facteur d'épissage SF2 et l'exon EDI est alors mieux détecté donc mieux inclus. De même, il a été montré qu'un promoteur contenant un élément de réponse DR-1 permet la fixation du facteur de transcription PPARγ qui recrute le co-activateur transcriptionel PGC-1 qui permet alors le recrutement du facteur d'épissage SRp40 (Fig.15) (Kornblihtt, 2005; Monsalve et al., 2000).

Il est également intéressant de noter, que les facteurs de transcription ou les co-régulateurs transcriptionnels ne sont pas les seuls à permettre le recrutement de facteurs d'épissage. En effet, des facteurs d'épissage peuvent être recrutés directement via le domaine C-terminal (CTD) de la RNAPII. Ce domaine de la grande sous-unité de la RNAPII est constitué de 58 répétitions de 7 acides aminées dont la phosphorylation détermine la régulation de la transcription. Par exemple, l'exon E33 de la fibronectine a une fréquence accrue d'inclusion grâce à l'action de SRSF3. La troncature du CTD abolit l'action de SRSF3 (de la Mata and Kornblihtt, 2006). Plus généralement, le CTD de la RNAPII pourrait permettre le recrutement de nombreuses protéines SRs. Enfin, des facteurs d'épissage ou des éléments de la machinerie d'épissage peuvent être recrutés par des facteurs d'élongation de la

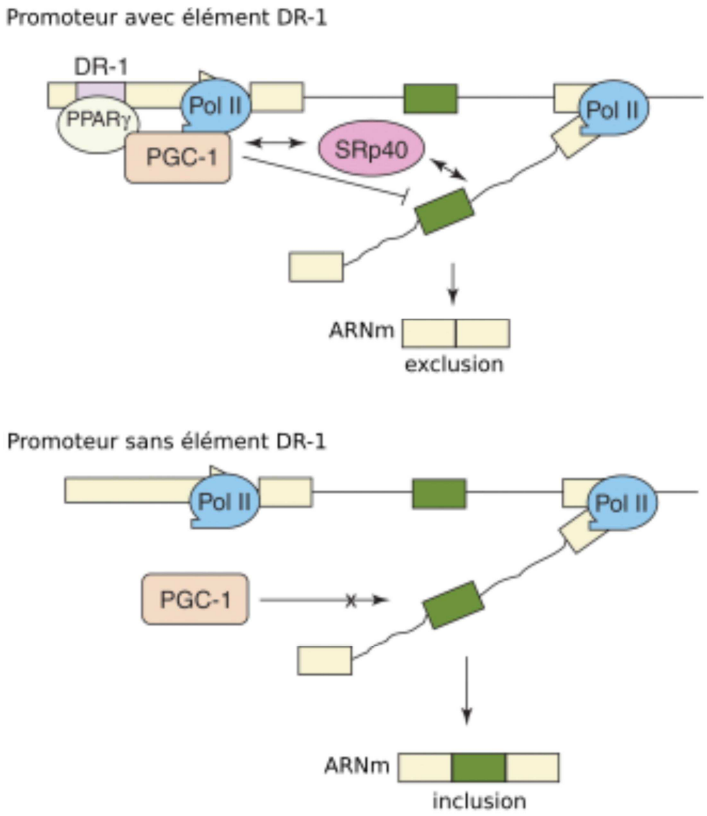


Figure 15 : Régulation de l'épissage par les facteurs de transcription recrutés au promoteur. La présence ou l'absence de l'élément DR-1 dans le promoteur du gène Fibronectine réprime ou active l'inclusion de l'exon IIIb de la fibronectine (en vert) via le recrutement de facteurs, comme PGC-1 qui permettent le recrutement de facteurs d'épissage comme SRp40 (Kornblihtt, 2005).

transcription comme TFIIS, ou TAT-SF1 qui est lui-même recruté par P-TEFb qui modifie le CTD de la RNAPII (Kadener et al., 2001; Nogués et al., 2003; Sánchez-Hernández et al., 2016).

Un autre mécanisme par lequel des facteurs de transcription peuvent avoir un effet sur l'épissage est lié à la vitesse d'élongation de la transcription (Fig.16) (Jonkers and Lis, 2015; Jonkers et al., 2014). Ainsi, une étude a démontré que l'antigène T du SV40, induisant un ralentissement de la RNAPII, augmentait la fréquence d'inclusion de l'exon EDI de la fibronectine placé en aval de ce

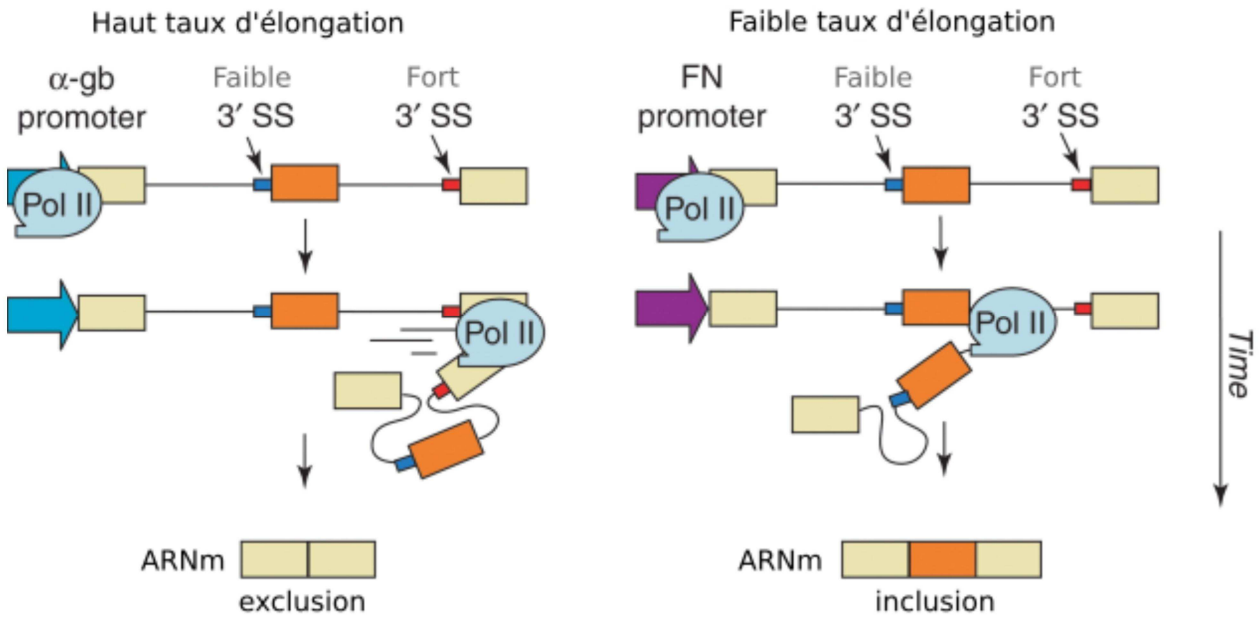


Figure 16 : Illustration de comment un promoteur peut affecter la régulation de l'épissage alternatif via le contrôle du taux d'élongation par la RNAPII. Le 3'SS (en bleu) de l'exon alternatif EDI (en orange) de la fibronectine est plus faible que le 3' SS (en rouge) de l'exon en aval. Un faible taux d'élongation sous le contrôle du promoteur de la fibronectine favorise l'inclusion (partie droite), tandis qu'un fort taux d'élongation sous le contrôle du promoteur de l'α-globine favorise l'exclusion de l'exon (partie gauche) (Kornblihtt, 2005).

promoteur. Si la vitesse de la RNAPII est augmentée avec l'activateur de transcription VP16, le taux d'inclusion de l'EDI est drastiquement diminué (Kadener et al., 2001).

4.2. Organisation de la chromatine et épissage alternatif

L'observation de l'effet de la dynamique de la transcription sur l'épissage, a conduit plusieurs équipes à analyser la relation entre épissage et organisation de la chromatine composée de l'ADN et de protéines associées.

En effet, l'ADN (environ 150 bp) s'enroule autour d'un octamère composé des histones H2A, H2B, H3 et H4, formant la structure élémentaire de la chromatine qu'est le nucléosome. Les

nucléosomes se retrouvent tout le long du génome et forment le premier ordre de compaction de la chromatine. Il est maintenant bien établi que les nucléosomes ne sont pas positionnés de façon aléatoire le long d'un génome (Bai and Morozov, 2010; Schwartz et al., 2009). L'octamère d'histone étant en interaction directe avec l'ADN, la composition en bases de la séquence ADN influe sur la probabilité que cette séquence s'enroule autour d'un nucléosome (Segal et al., 2006). Chez l'humain, les nucléosomes sont plus fréquemment positionnés sur des séquences enrichies en guanines et cytosines. A l'inverse, des séquences enrichies en adénines et thymines ont une probabilité plus faible d'être enroulées dans un nucléosome.

Les processus nucléaires impliquant la chromatine, comme la transcription, impactent aussi le positionnement des nucléosomes. Il n'y a généralement pas de nucléosome en amont du site où se forme le complexe d'initiation de transcription et le premier nucléosome d'un gène est situé juste après la position de début de transcription. (Fig.17A) (Bai and Morozov, 2010). Ce signal de positionnement est parmi les plus forts et se retrouve aussi chez d'autres Eucaryotes. Au sein même du gène, la répartition des nucléosomes est aussi inégale. Il a été montré une fréquence accrue de nucléosomes sur les exons en comparaison avec les introns (Fig.17B) (Andersson et al., 2009; Schwartz et al., 2009; Spies et al., 2009; Tilgner et al., 2009; Wilhelm et al., 2011). Ce positionnement au niveau des exons s'explique en partie par l'enrichissement en guanine et cytosine dans les séquences exoniques en comparaison des séquences introniques. La présence d'un nucléosome au niveau d'un exon est d'autant plus probable lorsque l'exon est isolé entre deux grands introns et a une taille proche de celle permettant un tour complet autour du nucléosome, c'est-à-dire environ 150 bp (Spies et al., 2009).

Il a été proposé que les nucléosomes forment un obstacle lors de l'élongation de la transcription (Gaykalova et al., 2015). Se faisant, les nucléosomes positionnés sur des exons induiraient un

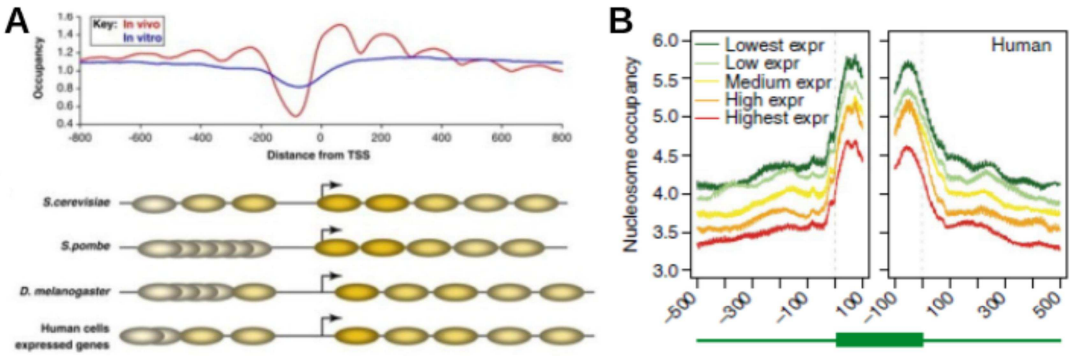


Figure 17 : A. Positionnement des nucléosomes à l'extrémité 5' d'un gène. En haut : signal de présence de nucléosome autour de la position de début de transcription mesuré par des expériences in vivo et in vitro. En bas : représentation du placement des nucléosome dans le même référentiel d'après les mesures d'occupation réalisées sur 4 espèces eucaryotes (Bai and Morozov, 2010). B. Mesure de la présence en moyenne de nucléosome autour des extrémités des exons, alignés sur le 3'SS (partie gauche) ou le 5'SS (partie droite), rassemblés en 5 groupes selon leur niveau d'expression (Schwartz et al., 2009).

ralentissement de la vitesse d'élongation de la RNAPII favorisant la reconnaissance des exons au sein du pré-ARNm naissant.

Les nucléosomes influenceraient également l'épissage au travers de leurs modifications. En effet, les histones peuvent être chimiquement modifiées par acétylation, méthylation, ubiquitinylation, sumoylation, et phosphorylation (Latham and Dent, 2007). Certaines marques d'histones sont associées à l'activité transcriptionnelle des gènes. Par exemple, la tri-méthylation de la lysine 4 de l'histone 3 (H3K4me3) est essentiellement détectée à l'extrémité 5' des gènes exprimés. La tri-méthylation de la lysine 36 de l'histone 3 (H3K36me3) est détectée dans le corps des gènes transcris. La tri-méthylation de la lysine 9 de l'histone 3 (H3K9me3) est détectée dans l'hétérochromatine.

Plusieurs études ont investi la relation entre les modifications d'histones et régulation de l'épissage. Ces études ont été réalisées soit sur des exons particuliers décrivant des mécanismes

spécifiques aux exons étudiés (Khan et al., 2016; Saint-André et al., 2011; Schor et al., 2009), soit à l'échelle du génome, permettant d'établir des corrélations (Andersson et al., 2009; Schwartz et al., 2009; Spies et al., 2009). Au travers de ces études, il a été montré que les marques H3K36me3, H3K27me2 et H3K4me3 sont enrichies sur les exons comparés aux introns (Spies et al., 2009) (Fig.18). Par ailleurs, les exons alternatifs sont moins enrichis en H3K36me3 que les exons constitutifs. Il a été montré que la marque H3K4me3 permet le recrutement de facteurs d'épissage et l'assemblage du spliceosome (Davie et al., 2015). Sa présence sur les exons 9 et 10 du gène c-MPL favorise leur inclusion. La marque H3K9me3 présente sur certains exons comme ceux de CD44 favoriserait leur inclusion (Saint-André et al., 2011).

Un autre type de modification d'histone, l'acétylation, interagit aussi avec la régulation de l'épissage. L'acétylation des histones est associée à une chromatine plus accessible pour les protéines.

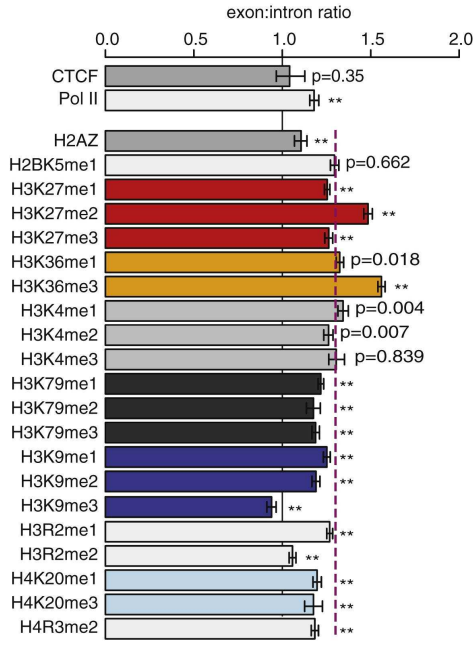


*Figure 18 : Comparaison d'enrichissement de marques d'histone entre l'exon et les régions introniques flaquantes. Ratio de référence à 1.0 pour CTCF et Pol II (RNAPII), à 1.3 pour les modifications d'histone. ** p < 0.01 après correction pour les tests multiples. Barre d'erreur : intervalles de confiance à 95 % (Spies et al., 2009).*

Il a été proposé que l' effet de l'acétylation des histones sur la compaction de la chromatine donc sur la vitesse d'élongation de la transcription a un effet sur la reconnaissance des exons au cours de l'épissage. Par exemple, il a été montré que la dépolarisation des cellules neuronales induit une augmentation de l'acétylation de H3K9 qui favorise une augmentation du taux d'élongation de la transcription impactant sur l'épissage de l'exon 18 du gène NCAM (Schor et al., 2009).

Un autre mécanisme par lequel les modifications d'histone ont un effet sur l'épissage est lié au recrutement de protéines par ces marques. En effet, les modifications d'histone peuvent être détectées par des protéines « lectrices » qui recrutent à leur tour des facteurs d'épissage et influencent ainsi la décision d'épissage (Fig.19) (Luco et al., 2011). Par exemple, il a été montré que H3K4me3 est reconnu par CHD1 qui recrute U2 snRNP et U2AF65 (Davie et al., 2015). La protéine MRG15

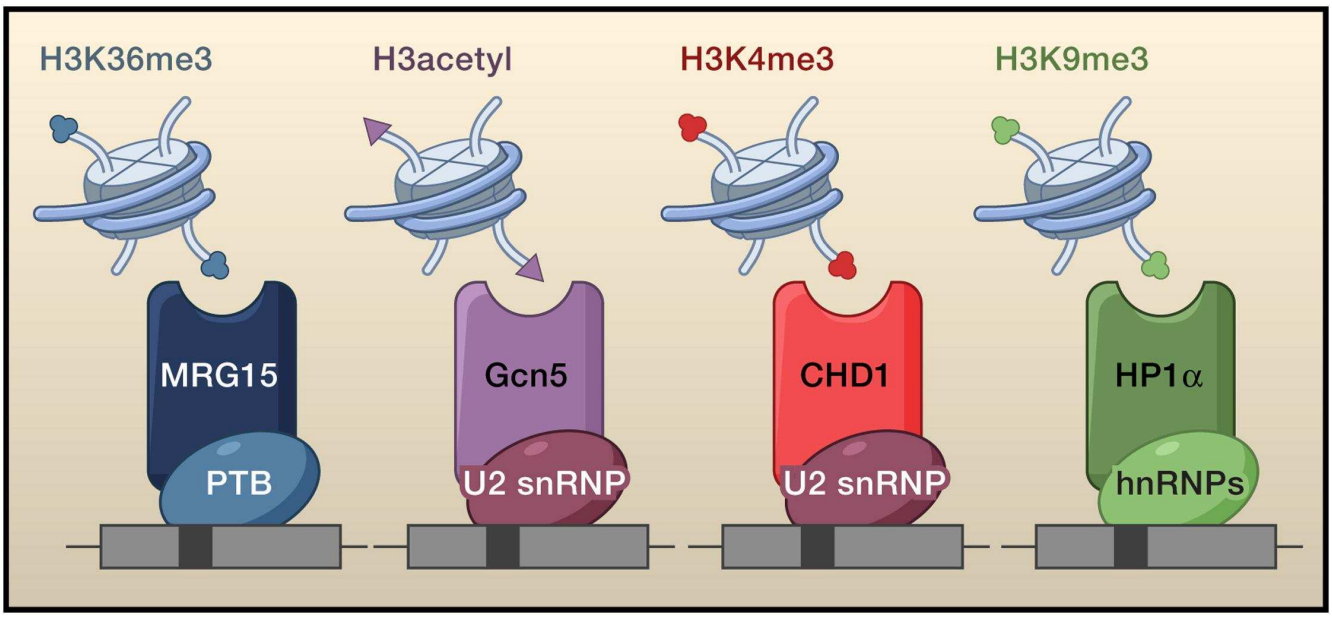


Figure 19 : Pontages entre des modifications d'histone et la régulation d'épissage par des paires d'interaction protéines « lectrices »/facteur d'épissage. Une protéine « lectrice », recrutée à la chromatine par une modification d'histone, recrute un facteur d'épissage qui impacte la régulation de épissage (Luco et al., 2011).

reconnaît la marque H3K36me3 et permet le recrutement du facteur d'épissage PTB (Luco et al., 2010).

En conclusion, les mécanismes de régulation de l'épissage alternatif, impliquant la transcription et l'organisation de la chromatine sont très divers et alimentent la complexité des mécanismes de régulation décrits au niveau de l'ARN. Cette diversité de mécanismes obscurcit notre compréhension de la régulation de l'épissage alternatif.

5. Organisation 1-D et 3-D de la chromatine

Le génome humain est composé à 41 % de guanines et cytosines. Ce pourcentage de GC n'est pas constant le long du génome. Le génome humain (comme celui d'autres espèces) peut être découpé en régions, nommées des isochores, qui ont une taille variant de quelques milliers à plusieurs centaines de milliers de bp dans lesquelles le pourcentage en GC est homogène (Fig.20A) (Bernardi, 2015). Cinq catégories d'isochores ont été définies : L1 (%GC:36), L2 (38.9), H1 (43.1), H2 (48.7) et H3 (54.5) (Costantini et al., 2009). Le taux de GC est souvent relié à l'organisation du génome en éléments fonctionnelles. Par exemple, les régions génomiques pauvres en GC contiennent de grands gènes, tandis que les régions génomiques riches en GC correspondent à de petits gènes contenant un îlot CpG dans leur promoteur (Beck et al., 2018; Liu et al., 2018).

Récemment, il a été montré que l'organisation du génome en isochores corrèle avec la structuration tri-dimensionnelle de la chromatine dans le noyau (Labena et al., 2018). La chromatine forme des boucles (ou TAD : « topologically associating domain ») sous l'action de protéines

insulatrices, comme CTCF et CP190, et la cohésine (Ramírez et al., 2018; Sequeira-Mendes and Gutierrez, 2016). Les TADs sont des domaines caractérisés, par exemple, par leur contenu en GC, les marques chromatiniennes, la condensation de la chromatine, ou encore l'activité transcriptionnelle. Ainsi, l'hétérochromatine est composée de TADs condensés, ayant des nucléosomes portant la marque H3K9me3, et correspondant à des régions génomiques ayant une faible activité transcriptionnelle (Fig.20C). Ces régions correspondent à des isochores pauvres en GC. L'euchromatine est composée de TADs décondensés, portant des marques chromatiniennes comme H3K4me3, H3K36me3 ou H3K27me3, et correspond à des régions génomiques ayant une forte activité transcriptionnelle. Ces régions correspondent à des isochores riche en GC (Sequeira-Mendes and Gutierrez, 2016).

A plus grande échelle, les TADs de l'hétérochromatine sont regroupés à la périphérie du noyau alors que les TADs de l'euchromatine sont rassemblés au centre du noyau (Bernardi, 2018; Sequeira-Mendes and Gutierrez, 2016) (Fig.20B). Il a été proposé que cette organisation de l'ADN dans l'espace nucléaire contribue à la régulation de l 'expression des gènes en favorisant par exemple la co-régulation des gènes qu'ils contiennent (Karathia et al., 2016; Nguyen and Bosco, 2015; Tsochatzidou et al., 2017).

Pour résumer, il existe un lien évident entre l'organisation 3D de la chromatine et la régulation transcriptionnelle et/ou les modifications des histones. Dans la mesure où, j'ai mentionné dans les parties ci-dessus qu'il existe une relation entre activité transcriptionnelle, modifications des histones et épissage, une question importante à adresser concerne la relation entre l'organisation 3D de la chromatine et épissage. D'autre part, comme l'organisation 3D de la chromatine est reliée aux biais de composition nucléotidiques (par exemple, le taux de GC), cela pose également la question de la relation entre biais de composition nucléotidique des gènes et épissage.

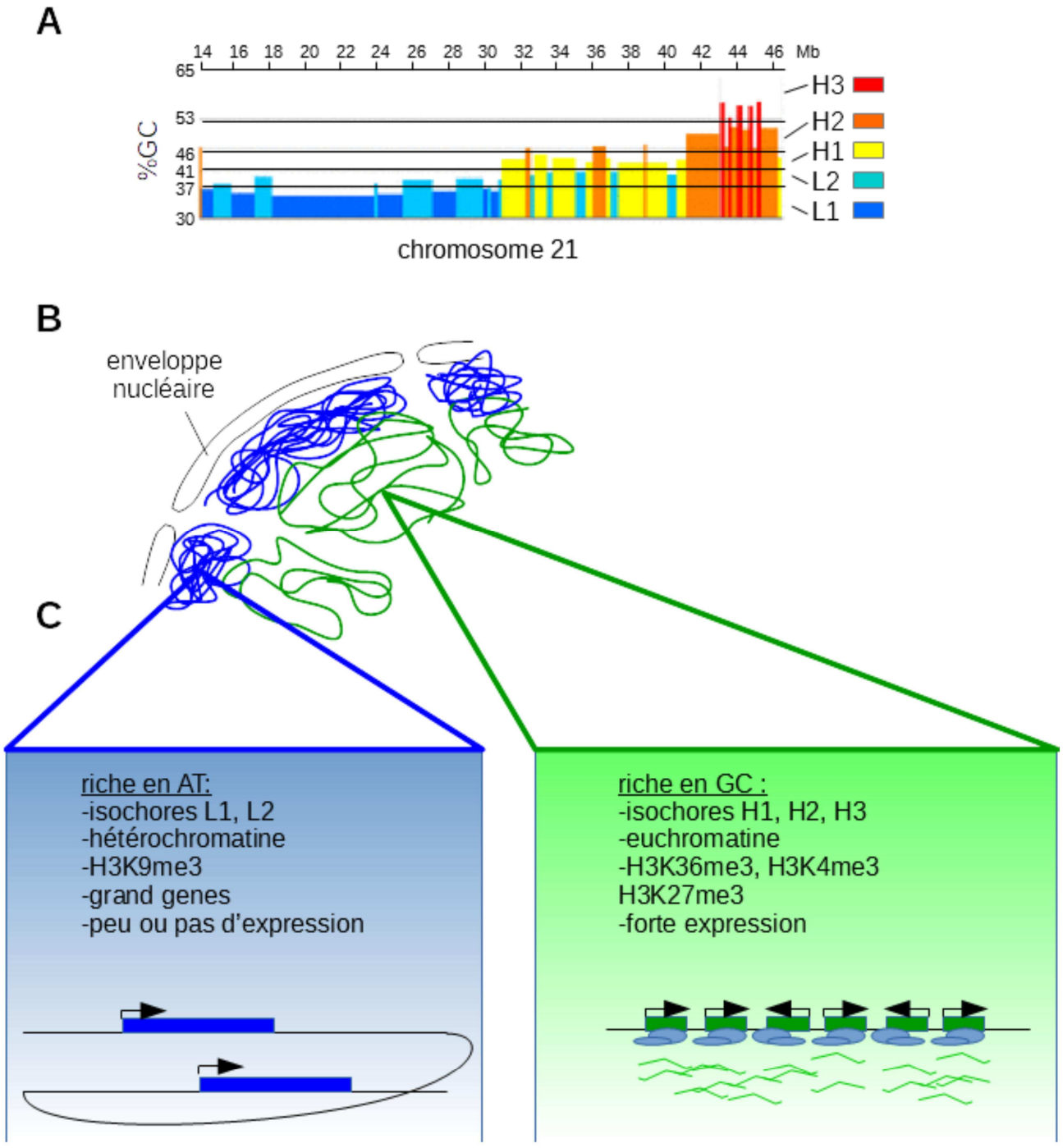


Figure 20 : A. Découpage du chromosome 21 en isochores de cinq catégories de taux de GC. (Cozzi et al., 2015) B. Organisation de la chromatine dans le noyau : hétérochromatine (bleu) à la périphérie et euchromatine (vert) au centre du noyau. C. Composition de la chromatine au sein du noyau en fonction du taux de GC dans la séquence nucléotidique.

6. Conclusion et Objectifs de thèse

Durant ces 20 dernières années, d'importants efforts et progrès ont été faits afin de caractériser les différents mécanismes contribuant à la régulation de l'épissage alternatif comme je les ai décrits ci-dessus. Ainsi, l'inclusion ou l'exclusion d'un exon au cours de l'épissage dépend de multiples signaux. Certains proviennent de l'organisation de la chromatine, d'autres de la transcription (Fig 20). Enfin, une multitude de signaux provenant des séquences de l'ARN ou de structures secondaires en interaction avec un grand ensemble de protéines sont intégrés aboutissant à l'inclusion ou l'exclusion d'exons (Fig.21).

Dans ce contexte, un enjeu majeur est maintenant de comprendre comment ces multiples signaux interagissent les uns avec les autres et sont intégrés pour aboutir à un phénomène biologique

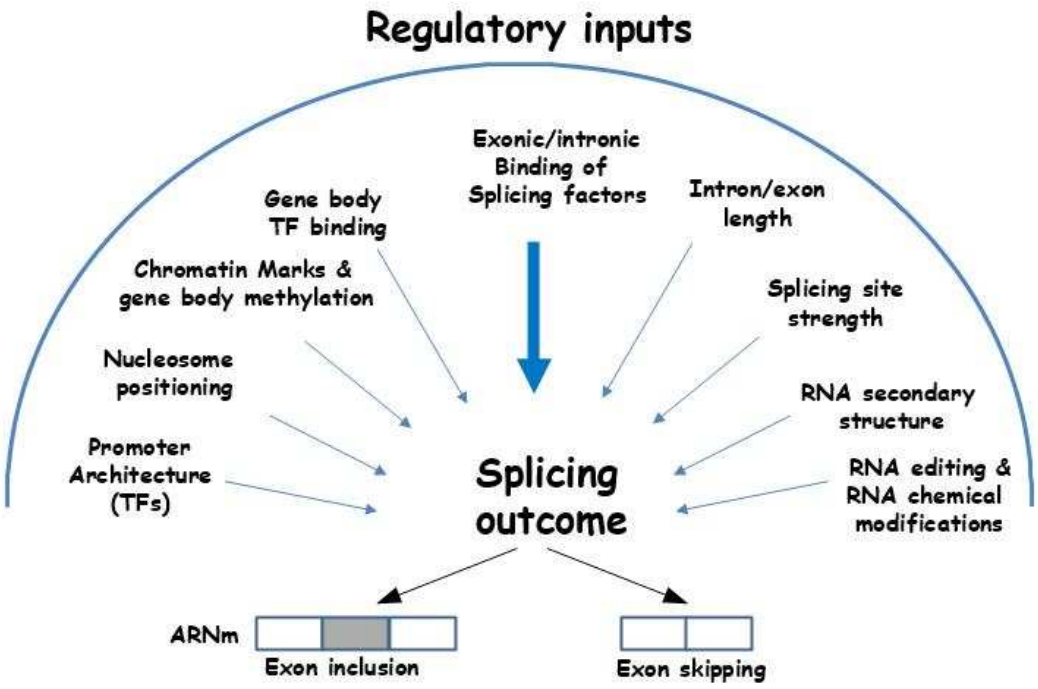


Figure 21 : Une multitude de paramètres influencent la régulation de l'épissage, notamment dans la détection des sites d'épissage. Le résultat de l'épissage est l'inclusion de l'exon, ou son exclusion de l'ARNm.

contrôlé et régulé. Face à la diversité des mécanismes de régulation de l'épissage, il est devenu essentiel d'identifier des règles générales permettant d'intégrer ces mécanismes dans un ensemble plus cohérent.

Afin d'adresser cette question, le laboratoire a analysé plusieurs dizaines de RNA-seq publiques déposés dans des bases de données. Les RNA-seq analysés ont été générés à partir de cellules transfectées avec des siRNAs ou des shRNAs ciblant spécifiquement certains facteurs d'épissage. Grâce à ces jeux de données, il a été possible d'établir les listes d'exons régulés par plus de 30 facteurs d'épissage. A l'aide de ces jeux de données, l'objectif est de rechercher les éléments communs à des exons co-régulés, c.-à-d. par exemple régulés par un même facteur d'épissage.

Pour cela, j'ai développé et utilisé une série d'analyses bio-informatiques. Dans la partie Résultats, je décrirai dans un premier temps, le travail réalisé montrant qu'un élément commun entre des exons co-régulés par un même facteur d'épissage concerne le biais en composition nucléotidique. Autrement dit, les exons co-régulés ont une composition nucléotidique similaire. Je décrirai les conséquences de cette observation en termes d'organisation de la chromatine et de reconnaissance des exons au cours de l'épissage.

Dans un deuxième temps, je décrirai comment les biais de composition nucléotidique dans des exons co-régulés pourraient contribuer à mieux comprendre les conséquences fonctionnelles de la régulation de l'épissage au niveau protéique. Ce travail est en révision dans Genome Research.

Bien que je n’en parle pas dans ce manuscrit de thèse, je souhaite également mentionner que mon travail a contribué à deux autres publications réalisées en collaboration avec des équipes du laboratoire:

“Nucleosome eviction in mitosis assists condensin loading and chromosome condensation.”

Toselli-Mollereau E, Robellet X, Fauque L, **Lemaire S**, Schiklenk C, Klein C, Hocquet C, Legros P, N'Guyen L, Mouillard L, Chautard E, Auboeuf D, Haering CH, Bernard P.

EMBO J. **2016** Jul 15;35(14):1565-81. PMID: 27266525

The proto-oncogenic protein TAL1 controls TGF-β1 signaling through interaction with SMAD3. Terme JM, **Lemaire S**, Auboeuf D, Mocquet V, Jalinot P.

Biochim Open. 2016 May 14;2:69-78. PMID: 29632840

Résultats

Article # 1

“Splicing factors and chromatin organization enhance exon recognition by alleviating constraints generated by gene nucleotide composition bias”

Introduction

Comme je l’ai décrit précédemment, la régulation de l’épissage dépend en partie de facteurs d’épissage qui sont des protéines de liaison à l’ARN se fixant aux pré-ARNm. Ces protéines reconnaissent des motifs et selon que ces motifs sont localisés dans les exons ou dans les introns, la fixation des facteurs d’épissage active ou réprime l’inclusion des exons. Néanmoins, il est attendu que les exons co-régulés par un même facteur d’épissage partagent différentes propriétés. Afin d’identifier ces propriétés, Hélène Polvèche, Ingénieure d’étude dans le laboratoire, a analysé plusieurs dizaines de RNA-seq déposés dans des bases de données publiques et qui ont été générés à partir de cellules dont l’expression de facteurs d’épissage spécifiques a été manipulée expérimentalement. Grâce à ces jeux de données, il a été possible d’établir les listes d’exons régulés par plus de 30 facteurs d’épissage.

Avec l’aide de Nicolas Fontrodona et de Jean-Baptiste Claude, bio-informaticiens au laboratoire, j’ai développé et utilisé une série d’analyses bio-informatiques afin de rechercher des éléments communs au sein des différents jeux d’exons co-régulés. Comme il est décrit dans l’article présenté ci-dessous, un point commun entre des exons co-régulés par un même facteur d’épissage repose sur le fait que des exons co-régulés ont une composition nucléotidique similaire. Nous montrons que ce biais de composition nucléotidique a des conséquences sur l’organisation de la chromatine et la régulation de l’épissage.

Splicing factors and chromatin organization enhance exon recognition by alleviating constraints generated by gene nucleotide composition bias

Short title: Gene nucleotide composition bias constrains the splicing regulatory process

Sébastien Lemaire[1,$], Nicolas Fontrodona[1,$], Jean-Baptiste Claude[1], Hélène Polvèche[2], Fabien Aubé[1], Laurent Modolo[3], Cyril F. Bourgeois[1], Franck Mortreux[1], Didier Auboeuf[1,*]

[1.] Univ Lyon, ENS de Lyon, Univ Claude Bernard, CNRS UMR 5239, INSERM U1210, Laboratory of Biology and Modelling of the Cell, 46 Allée d'Italie Site Jacques Monod, F-69007, Lyon, France

[2.] CECS, I-Stem, Corbeil-Essonnes, 91100, France.

[3.] LBMC biocomputing center, CNRS UMR 5239, INSERM U1210, 46 Allée d'Italie Site Jacques Monod, F-69007, Lyon, France

[$] Equal contribution

*Corresponding author: Didier Auboeuf, Laboratory of Biology and Modelling of the Cell, ENS de Lyon, 69007 Lyon, France. Didier.auboeuf@inserm.fr

Abstract

Increasing evidence suggests that the one-dimensional genome organization, defined by regional nucleotide composition bias determines the physical properties of the DNA polymer with consequences on the three-dimensional genome architecture. Both, the one- and three-dimensional genome organization then influence gene transcriptional activity. However, the interplay between genome organization and mRNA processing, in particular splicing is not known. In this report, we identified sets of exons activated by different splicing factors that are RNA binding proteins interacting with nucleotide compositionally-biased sequences. We uncovered a set of 12 splicing factors that preferentially activate GC-rich exons (labelled "GC-exons") flanked by small GC-rich introns and that are hosted by small GC-rich genes. Another set of 19 splicing factors preferentially activates AT-rich exons (labelled "AT-exons") flanked by large AT-rich introns and hosted by large AT-rich genes. We observed that the increase in the GC-load increases the stability of base-pairing interaction between complementary sequences at the 5' splicing site suggesting the formation of stable secondary structure. This feature is known to decrease the efficiency of 5' splicing site recognition by the U1 snRNA. Accordingly, GC-exons dependent on U1 snRNP-associated factors. Meanwhile, the increase in the AT-load weakens exon recognition by increasing the number of decoy signals at intron 3'-end, such as the number of branchpoints, and SF1- or U2AF2-binding sites. Accordingly, AT-exons dependent mostly on U2 snRNP-associated factors. We also observed that nucleotide composition bias influences local chromatin organization at the exon level since nucleosomes are well positioned on AT-exons, while both GC-exons and their flanking introns have a higher density of nucleosomes. In addition, genes hosting GC-exons contain a higher density of RNAPII and are associated with different histone modifications when compared to genes hosting AT-exons. Finally, genes hosting GC- or AT-exons are not distributed randomly across the genome but each set of exons clusters in different genomic regions, such as different isochores and topologically-associated domains. We propose a model in which regional- and gene-nucleotide composition bias associated with genome organization and/or gene transcriptional activity creates constraints at the exon level with consequences on exon recognition during the splicing process. In this model, local chromatin organization and splicing factors enhance exon recognition by counterbalancing nucleotide composition bias-associated constraints. Therefore, our work demonstrates that nucleotide composition bias establishes a straightforward link between genome organization and local molecular events, like splicing.

Introduction

Most eukaryotic genes are composed of exons and introns. Introns are defined at their 5'-end by the 5' splicing site (ss) recognized by the U1 snRNP that contains the U1 snRNA. The U1 snRNA interacts with the 5' ss [1]. At their 3'-end, introns are defined by the branch point (BP, recognized by SF1), the polypyrimidine tract (Py, recognized by U2AF65) and the 3' ss (recognized by U2AF35). SF1 and U2AF65 allow the recruitment of the U2 snRNP containing the U2 snRNA that interacts with the BP[1]. In addition to linear sequences (e.g., the Py-tract), RNA secondary structures play an important role in splicing. For example, secondary structures at the 5'ss can hinder the interaction between the 5' ss and the U1 snRNA[2,3]. Meanwhile, secondary structures at the 3'-end of short introns can replace the need of U2AF65[4,5]. Splicing-related signals are short and degenerated sequences and exons are much smaller than introns. Thus, a puzzling issue is how are exons precisely defined and how *bona fide* splicing-related signals are distinguished from pseudo- or decoy-signals.

Many (if not all) exons require a variety of splicing factors to be defined. Splicing factors that belong to different families of RNA binding proteins such as the SR and hnRNP families bind to short degenerated motifs present in the pre-mRNAs either within exons or introns[6,7]. Splicing factor binding sites are low complexity sequences made of the same nucleotide or dinucleotide[8-10]. When bound to pre-mRNAs, splicing factors modulate the recruitment of the spliceosome and/or modulate different steps of the spliceosome assembly[6,7,11].

The spliceosome assembly and the splicing process occurs mostly during transcription[6,7,12]. In this setting, the velocity of RNA polymerase II (RNAPII) can influence the exon recognition process[6,12]. However, the interplay between transcription and splicing is complex since, for example, speeding up transcription elongation can either enhance or repress exon inclusion[13]. RNAPII velocity is in turn influenced by the local chromatin organization, such as the presence of nucleosomes[11,12]. Nucleosomes are preferentially positioned on exons because exons have a higher GC-content than introns, which increases DNA bendability and the Py-tract (mostly made of Ts) upstream exons form a nucleosome energetic barrier[14-20]. Nucleosomes influence splicing by slowing down RNAPII in the vicinity of exons but also by modulating the local recruitment of splicing regulators[11,12]. Indeed, histone tails interact directly or indirectly with splicing factors depending on specific chemical modifications (e.g., methylation) of histone tail amino acids[11,12,21]. Therefore, exon recognition during the splicing process depends on a complex interplay between signals at the DNA level (e.g., nucleosome positioning) and signals at the RNA level (e.g., splicing factor binding sites).

Genes are not organized randomly across a genome and nucleotide composition biases over more or less large genomic regions plays an important role in genome organization. For example, large regions (several Kbs) of the human genomes have uniform GC-content that differs from adjacent regions. These regions or isochores are either GC-rich or GC-poor[22-26]. Interestingly, GC-rich

isochores have a high density of genes and these genes contain often small introns, while AT-rich genes contain often large introns[22,24]. It has been proposed that splicing of introns within GC- or AT-rich environment occurs through different operational modes[11,20,27]. For example, the splicing of short introns in a GC-rich context may occur through an intron definition model, while the splicing of large introns in an AT-rich context may occur through an exon definition model[11,20,27-30]. In the intron definition model, the U1 and U2 snRNP interact across introns when exons are flanked by short introns, while in the exon definition model, the U2 and U1 snRNP are stabilized across exons when exons are flanked by large introns[11,20,27-30]. In addition, since nucleosome positioning depend on the GC load (see above), nucleosomes are more precisely positioned on AT-rich exons than on GC-rich exons since AT-rich exons are flanked by long introns having a sharp different GC-load when compared to exons [11,20,27-30]. Collectively these observations support a model where the gene architecture (e.g., size of introns) and its nucleotide composition bias (e.g., GC load) can influence local processes (at the exon level) such as nucleosome positioning and intron removal. Since exon recognition depends also on the binding to the pre-mRNAs of splicing factors that interact with compositionally-biased sequences (see above), one interesting possibility is that the nature of splicing factor binding to pre-mRNAs depend at least in part on gene nucleotide composition bias. In this setting, we recently report that exons regulated by different splicing factors have different nucleotide composition bias[31].

In this report, we investigated the relationship between gene nucleotide composition bias, local chromatin organization and the splicing process. To this purpose, we identified sets of exons regulated by different splicing factors and we demonstrate that the analysis of nucleotide composition bias allows to better understand the interplay between chromatin organization and splicing-related features that collectively impact on exon recognition. We propose that nucleotide composition bias not only contributes to genome 3D organization in the nucleus space but has also consequences locally, at the exon level, during the splicing process.

Results

Splicing factor-dependent GC-rich and AT-rich exons.

Publicly available RNA-seq datasets generated after knocking down or over-expressing 33 different splicing factors across different cell lines were analyzed (Supplementary Table S1). Using our recently published FARLINE pipeline, that allows to quantify the inclusion rate of exons from RNA-seq datasets[32], we defined the sets of exons whose inclusion is activated by each of the 33 analyzed-splicing factors (Supplementary Table S2). We identified 14645 exons that were activated by at least one splicing factor among the 93680 exons whose inclusion rate can be quantified by FARLINE across all the datasets (Supplementary Table S2). We focused on splicing factor-activated exons in order to uncover the splicing-related features characterizing exons those recognition depends on splicing factors.

As expected, most of splicing factor-activated exons had weaker 5'- and 3'-ss scores when compared to the average scores of all human exons (Supplementary Fig S1A). The nucleotide composition of splicing factor-activated exon was determined (Supplementary Fig S2). Since our goal was to analyze sequence-dependent features at both the DNA and RNA level, we decided to systematically refer to thymidine (T) rather than Uridine (U) for clarity purpose event though Us replace Ts in RNAs. In addition, the values obtained for a set of splicing factor-activated-exons were normalized by the mean values measured for all human exons used as controls, in order to represent results in a consistent manner. As shown in Figure 1A (left panel), sets of exons activated by different splicing factors have differential GC-content when compared to the average nucleotide composition of all human exons (Figure 1A, Supplementary Fig Sup S3). Interestingly, the GC-content of splicing factor-activated exons positively correlated with the GC-content of their flanking intronic sequences (Figure 1B and Supplementary Fig Sup S3). Accordingly, splicing factor-activated GC-rich and AT-rich exons were flanked by GC-rich and AT-rich intronic sequences, respectively (Figure 1A, right panel). This observation is in agreement with reports indicating a positive correlation between the nucleotide composition of exons and those of their flanking introns[33].

The analysis of the size of introns flanking splicing factor-activated exons revealed that some sets of splicing factor-activated exons were flanked by small introns, while others were flanked by large introns when compared to the average size of all human introns (Figure 1C and Supplementary Fig Sup S1B). Interestingly, the GC-content of splicing-factor activated exons and the size of their flanking introns were negatively correlated (Figure 1D, upper panel). Likewise, the GC-content of introns and their size were also negatively correlated (Figure 1D, lower panel). To summarize, introns are generally shorter in a GC-rich context than in AT-rich context, as previously reported[34-36].

Based on these observations, we defined two groups of exons. The group of "GC-exons" corresponded to exons depending on splicing factors that activate GC-rich exons flanked by small introns (Figure 1E and Supplementary Table S2). The group of "AT-exons" corresponded to exons depending on splicing factors that activate AT-rich exons flanked by large introns (Figure 1E and Supplementary Table S2). To investigate the interplay between nucleotide composition bias and the splicing process, we excluded from our analyses, exons that were activated by splicing factors belonging to the two different groups and exons regulated by SRSF2, SRSF3 and hnRNPC since these splicing factors regulate GC-rich exons flanked by more or less large introns (Figure 1E, see Materials and Methods). We next analyzed different splicing-related features by comparing GC- and AT-exons representing two populations of exons that i) are activated by different splicing factors and ii) differ in terms of both GC-content and flanking intron size (Figures 1E and 1F).

Nucleotide composition bias and splicing-related features.

As already mentioned, exons and their flanking introns have similar nucleotide composition bias (e.g., Figure 1B). For example, intronic sequences flanking GC-exons have a higher density of G and/or C nucleotides when compared to intronic sequences flanking AT-exons (Figures 2A-2C). A high GC-load at the 5' ss was associated with a slightly weaker 5' ss score (Supplementary Figure S4), and was associated with a lower minimum free energy measured in a 50 nucleotide-long window centered at the 5' ss (Figure 2D, left panel). This suggests a higher stability of base-pairing interaction between complementary sequences and therefore formation of stable secondary structures at the 5' ss of GC-exons when compared to AT-exons. A similar feature was observed at the 3' ss of GC-exons when compared to AT-exons (Figure 2D, right panel).

GC-exons were impoverished in Ts but enriched in Cs just upstream their 3' ss and had therefore a similar Py-tract when compared to all human exons (Figures 2A and 2C). Meanwhile, AT-exons had a higher density of As upstream their 3' ss when compared to GC-exons and to all human exons (Figures 2A and 2C). We tested whether the high frequency of As was associated with a larger number of potential BP sites that are often made of an adenine[37,38]. As shown in Figure 2E (left panel), a higher proportion of AT-exons had more than two predicted BP sites in their upstream intronic sequence when compared to GC-exons and control exons. In addition, predicted BPs from the GC-exons were embedded in C-rich sequences when compared to BPs from the AT-exons (Figure 2F). It has been proposed that the interaction between the BP and the U2 snRNA is more stable when the BP is in a GC-rich context[37-39]. We therefore, computed the binding energy of BP sequences to the U2 snRNA sequence (AUGAUGUG, without the BP itself that is in a bulge in the structure, see Methods). A shown in Figure 2E (right panel), the U2 snRNA binding energy to the BP was higher for GC-exons when compared to exons of the AT-exons. This raises the possibility that the higher

number of potential BPs and their lower affinity for U2 snRNA may weaken the recognition of AT-exons, when compared to GC-exons (see below).

Since there was a higher density of As and Ts upstream AT-exons (Figures 2A and 2C), we next investigated whether this feature may interfere with the number of potential binding motifs for SF1 (that binds to UNA motifs) and U2AF65 (that binds to U-rich motifs) [1]. As shown in Figure 2G (left panel), there was a larger proportion of AT-exons containing more than two TNA motifs upstream their 3' ss when compared to the GC-exons. In addition, there was a larger proportion of AT-exons containing upstream their Py tract more than two low complexity sequences made of three Ts in 4 consecutive nucleotides (Figure 2G, right panel). Supporting the biological relevance of this observation, the analysis of U2AF65 ClIP-seq datasets revealed a higher U2AF65-related signal upstream AT-exons when compared to GC-exons (Figure 2H). In particular, the U2AF65-related signal level detected upstream AT-exons extended upstream the Py-tract (Figure 2H, green arrows) in agreement with the pattern of T density (Figure 2A).

To summarize, exons in a GC-rich context were predicted to have more stable secondary structures at their 5' and 3' ss when compared to all human exons and to exons embedded in AT-rich context (Figures 2A and 2D). Secondary structures at the 5'ss can hinder the interaction between the 5' ss and the U1 snRNA, while secondary structures at the 3' end of short introns replace the need of U2AF65[2-5]. Therefore, we anticipated that GC-exons were more sensitive to factors contributing to the recognition of 5' ss-related signals than to factors contributing the recognition of 3' ss-related signals. Meanwhile, our analysis suggested that AT-exons may be more dependent on factors contributing the recognition of 3' ss-related signals. Indeed, the increase of A- and T-density upstream AT-exons (Figure 2A) was associated with a lower binding energy between the U2 snRNA and BPs as well as an increase of potential decoy signals that may interfere with the recognition of the BP and/or the binding of SF1 and U2AF65. Indeed, AT-exons had a larger number of potential BPs, SF1- and U2AF65-binding sites (Figures 2E and 2G). In this setting, it is now well established that splicing-related signals can compete with each other's with consequences on splicing efficiency (see Discussion). Therefore, we hypothesized that the recognition of 3' ss-related signals may be weakened in an AT-rich context.

Nucleotide composition bias and dependency for specific spliceosome components.

In order to test the interplay between nucleotide composition bias and the dependency of exons to specific spliceosome-associated factors, we analyzed publicly available RNAseq datasets generated after knowing down a variety of spliceosome-associated factors (Supplementary Tables S1 and S2). By analyzing the nucleotide composition of exons whose inclusion depends on the presence of specific spliceosome components, we observed that exons that are skipped upon the depletion of

SNRPC or SNRNP70 (components of the U1 snRNP) were in a GC-rich environment when compared to the mean value of all human exons (Figure 3A). Likewise, exons that are skipped upon the depletion of the DDX5 and DXX17 RNA helicases, that enhance exon inclusion by favoring the binding of the U1 snRNP in highly structured 5' ss[40,41] were also in a GC-rich environment (Figure 3A). In addition, the 3' and 5' ss of SNRPC-, SNRNP70-, and DDX5/DDX17-dependent exons were predicted to be embedded in stable secondary structures when compared to all human exons (Figure 3B).

Exons skipped upon the depletion of SF1, U2AF2, SF3A3, and SF3B4 (but not U2AF1) that are involved in the definition of splicing-related sequences at intron 3'-ends were in an AT-rich environment when compared to the average value of all human exons (Figure 3A). In addition, a larger proportion of SF1- and U2AF2-dependent exons contain more than two predicted BPs and SF1 binding sites when compared to SNRPC- or SNRNP70-dependent exons (Figures 3C and 3D). Finally, a larger proportion of SF1- and U2AF2-dependent exons contained more than two T-rich motifs upstream the Py-tract when compared to SNRPC- or SNRNP70-dependent exons (Figure 3E). This is in agreement with the T-density pattern (Figure 3A) and with a broader signal observed when analyzing U2AF65 ClIP-seq datasets (Figure 3F).

To summarize, the nucleotide composition bias and splicing-related features of exons that depend on SNRPC, SNRNP70 and DDX5/17 were similar to those of GC-exons, while the nucleotide composition bias and splicing-related features of exons that depend on U2AF2 and SF1 were similar to those of the AT-exons. Accordingly, GC-exons had a higher probability to be dependent on SNRPC, SNRNP70, or DDX5/17 depletion than AT-exons that had in turn a higher probability to be sensitive to SF1, U2AF2, SF3A3 and SF3B4 depletion (Figure 3G). Of note, neither AT- nor GC-exons were selectively affected by U2AF1 whose dependent-exons did not show any specific nucleotide composition bias (Figures 3A and 3G). Collectively these results suggest that the dependency of exons to specific splicing-related factors can be explained, at least in part, by local nucleotide composition bias (see Discussion).

Nucleotide composition bias, gene features and chromatin organization

Since exons and their flanking intronic sequences have similar nucleotide composition bias (Figures 1B, 2A and 3A), we wondered whether the observed local nucleotide composition bias could be extended at the gene level. As shown in Figure 4A (left panel), there was a positive correlation between the GC content of splicing-factor activated exons and the GC content of their hosting gene. Accordingly, GC-exons and AT-exons belong to GC- and AT-rich genes, respectively when compared to the mean value of all human genes (Figure 4B, left panel). In addition, a negative correlation between the gene GC-content and the gene size was observed (Figure 4A, right panel), as already

reported[34-36]. Accordingly, GC-exons that are flanked by small introns (Figure 1) belong to genes containing small introns (Figure 4B, middle panels). Meanwhile, AT-exons that are flanked by large introns (Figure 1) belong to large genes containing large introns (Figure 4B, middle and right panels).

In this setting, several reports indicate that GC-rich genes are more expressed when compared to AT-rich genes[35,42-45]. Remarkably, a different pattern of RNAPII density was observed when comparing genes hosting either GC- or AT-exons (Figure 4C). While, there was a similar density of RNAPII on the promoter and first exon of genes hosting AT- or GC-exons, the RNAPII density was higher across exons and introns of genes hosting GC-exons (Figures 4C and 4D and Supplementary Figure S4C). A similar result was obtained when analyzing the pattern of RNAPII phosphorylated on ser2, suggesting that RNAPII loaded on genes hosting GC-exons was likely productive (Figures 4E and 4F).

In addition to be associated with gene transcriptional activity, the GC-load is associated with nucleosome positioning. Indeed, nucleosomes are more frequently detected on exons (see Introduction). The analysis of MNase-seq and H3-Chip-Seq datasets across different cell lines revealed a higher density signal related to nucleosomes on GC- than AT-exons (Figure 5A). Importantly, while similar signals were observed on the first exon of genes hosting either GC- or AT-exons, higher density signals related to nucleosomes were observed across all the exons and all along the genes hosting GC-exons (Figure 5B and 5C, and Supplementary Fig. S4D and S4E). In addition, a stronger signal was observed across introns of genes hosting GC-exons (Figure 5B and 5C), in particular within introns flanking the GC-exons themselves (Figure 5A). This could be because of the higher density of GCs in these introns and a lower density of Ts at their 3'-ends when compared to introns flanking AT-exons (Figure 2). In agreement with the fact that the density of Ts is also increased downstream AT-rich exons (Figure 2A), we observed a marked nucleosome-free regions both upstream and downstream AT-exons when compared to GC-exons when analyzing both MNase-seq and H3-Chip-seq datasets (Figure 5A, green arrows).

The different pattern of nucleosomes on GC- and AT-exons, prompted us to analyze the pattern of histone tail modifications that play a role in splicing regulation (see Introduction). For this purpose, we analyzed several publicly available ChIP-seq datasets that correspond to 10 histone marks and that were generated from the same cell line or from different cell lines (Supplementary Table S1). As shown in Figures 5D, a higher density signal corresponding to H3K4me1, H3K4me2, H3K4me3, H3K9ac, and H3K27ac was detected on GC-exons when compared to AT-exons across different samples. No significant differences were observed for H3K9me3, H3K27me3, H3K36me3, H3K79me3, and H4K20me1 (Figure 5D). The pattern of histone modifications did not seem to be specific to splicing factor-regulated exons. Indeed, while there was a similar density of the analyzed-histone marks on the promoter and first exons of genes hosting either AT- or GC-exons, the

H3K4me3 and H3K9ac density signals were higher across all the exons and introns of the genes hosting GC-exons (Figures 5E and 5F and Supplementary Figures S5 and S6). Collectively these observations suggest that GC- and AT-exons belong to genes that have different nucleotide composition bias (i.e., GC load) and architectures (i.e., intron size) and that are in different chromatin environments across different cell lines.

GC- and AT-exons are hosted by genes belonging to different isochores and chromatin domains.

It is now widely recognized that nucleotide composition bias plays an important role in genome organization. For example, large human genomic regions, named isochores have uniform GC content that differs from adjacent ones (see Introduction). In addition, there are increasing evidence that nucleotide composition bias plays a role in the genome 3D-organization in the nucleus space [22-26,46-52]. Since we observed a positive correlation between the GC-content of splicing factor activated-exons and their hosting genes, we challenged the possibility that splicing factor activated-exons are not dispersed randomly across the human genome.

As shown in Figure 6A (and see Supplementary Figure S7A), a large proportion (>80%) of AT-exons are within GC-poor isochores (<46% of GC corresponding to L1, L2 and H1). Meanwhile, a large proportion (~60%) of GC-exons are within GC-rich isochores (>46% of GC corresponding to H2 and H3). Furthermore, GC- and AT-exons cluster in different isochores (Figure 6B). Indeed, some isochores contain a larger number of GC- than AT-exons, while other isochores contain a larger number of AT-exons (Figure 6B).

It is now well established that chromosomes adopt specific structures in the nuclear space. Some DNA regions have been shown to be in close proximity to the nuclear envelop. These regions are called Lamina-associated Domains (LADs)[53]. Meanwhile, different DNA regions separated by several dozen of Kbs can be in close proximity in the nuclear space. These regions are named Topologically-Associated Domains (TADs)[54]. LADs and TADs have been annotated across different cell lines using different experimental approaches[53,54]. Interestingly, we observed that TADs contain a similar proportion of GC- and AT-exons, while LADs contain more frequently AT-exons (Figure 6C). This is in agreement with the fact that LADs have been reported to correspond to AT-rich genomic regions[53]. This result suggests that different splicing factor regulated-exons are not randomly distributed across the human genome and in the nuclear space. Further supporting this notion, we observed that some TADs contain a higher proportion of GC- than AT-exons while other TADs contain a higher proportion of AT-exons than GC-exons (Figure 6D). This result was confirmed using different annotations of TADs across different cell lines (Supplementary Figure S7B).

Collectively these observations support a model where the splicing process is linked to the genome architecture because nucleotide composition bias influence regulatory processes both globally (at the genome level) and locally (at the exon level).

Discussion

In this report, we uncovered a set of 12 splicing factors that preferentially activate exons that are embedded in GC-rich regions and that are flanked by small introns (labelled "GC-exons"), and a set of 18 splicing factors that preferentially activate exons embedded in AT-rich regions and flanked by large introns (labelled "AT-exons") (Figure 1E). Splicing factors, like hnRNPH, hnRNPF, PCBP1, RBFOX2, RBM22, RBM25, RBMX, SRSF1, SRSF5, SRSF6, and SRSF9 that activate GC-exons, bind to G-, C-, or GC-rich motifs [8-10] (Supplementary Figure S8). Meanwhile, splicing factors, like SFPQ, DAZAP1, KHSRP, TRA2, RBM15, hnRNPU, and QKI that activate AT-exons bind to A-, T-, or AT-rich motifs[8-10,55] (Supplementary Figure S8). An interesting possibility is that gene nucleotide composition bias (e.g., GC load within a gene, see below) could increase the probability of generating, locally (at the exon level), specific classes of splicing factor binding sites (e.g., GC-rich motifs). However, the interplay between gene nucleotide composition bias and splicing factor binding motifs needs to be further investigated. Indeed, splicing factors like FUS, hnRNPA2B1, and hnRNPK activating AT-exons bind to G-, C-, or GC-rich motifs and some splicing factors, like hnRNPL, PTBP1, hnRNPA1, and hnRNPM bind to motifs composed of CT, CA, or GT dinucleotides[8-10,55,56] (Supplementary Figure S8). In this setting, increasing evidence indicates that the recognition of specific binding sites by RNA binding proteins depends on the local nucleotide context that could, for example, impact on the formation of RNA secondary structures[8]. Some splicing factors binding to GC-rich motifs may preferentially interact when these motifs are exposed in unstructured regions (i.e., in an AT-rich context) or conversely some factors binding to AT-rich motifs may preferentially interact when these motifs are exposed in loops at the extremities of stems (i.e., in a GC-rich context)[8].

Along the same line, our analysis and reports from the literature support a model where the gene and local GC-load influences specific steps of the spliceosome assembly in part because of the formation of secondary structures at the RNA level[2-5]. Indeed, we observed that exons sensitive to the depletion of SNRPC and SNRNP70, two components of the U1 snRNP are embedded in a GC-rich environment that favors the formation of stable secondary structures (Figures 3A and 3B). This is in agreement with reports indicating that RNA structures can hinder the recognition of the 5' ss by occluding them, therefore limiting access of U1 snRNA to the 5' ss[2,3]. In other words, RNA intra-molecular interactions (i.e., RNA secondary structures) can compete for RNA inter-molecular interactions (i.e., U1 snRNA and 5' ss interaction). Supporting this hypothesis, DDX5 and DDX17

helicases, that activate exons embedded in a GC-rich environment (Figures 3A and 3B) enhance the recognition of 5' ss embedded in secondary structures through its RNA helicase activity[40,41]. Furthermore, most of the splicing factors that activate GC-exons, including hnRNPF, hnRNPH, RBM22, RBM25, RBFOX2, and several SRSF splicing factors have been shown to enhance the recruitment of the U1 snRNP at 5' ss[57-62]. Therefore, one interesting possibility is that high GC-load increases the probability of generating secondary structures at the 5' ss, which decreases exon recognition but simultaneously, high GC-load increases the probability of recruiting splicing factors binding to GC-rich motifs that enhance the recruitment of U1 snRNP. It is interesting to note that while secondary structures at the 5' ss negatively impact exon recognition, several studies performed in different species demonstrated that secondary structures at the 3' ss favor exon recognition and replace the need of U2AF2 for splicing[4,5]. In addition, a high GC-load, as well as G- and C-rich motifs upstream the BPs have been shown to enhance U2 snRNA binding and BP recognition [37,38,63]. Collectively, these observations are in agreement with our observation that exons embedded in GC-rich environment are more sensitive to U1 snRNP-associated factors than to U2 snRNP-associated factors (Figure 3G).

Our results also support a model in which a high load of AT nucleotides in large introns can negatively influence splicin. Indeed, high AT-content upstream exons is associated with a larger number of potential BPs and a decrease in the strength of the interaction between BPs and the U2 snRNA (Figure 2E). This observation is in agreement with a previous report showing that large introns have weaker and a larger number of potential BPs when compared to short introns[39]. Along the same line, a larger proportion of AT-exons and exons dependent on SF1 or U2AF2 have a larger number of SF1 and U2AF2-binding motifs when compared to GC-exons or to average of all human exons (Figure 2G, 3D and 3E). In this setting, it must be underlined that some splicing factors activating AT-exons, including hnRNPA1, hnRNPM, RBM15, RBM39, SFPQ, and TRA2 (Figure 1E) interact with and enhance the recruitment of SF1, U2AF65, U2AF35 and/or the U2 snRNP[64-69]. Of note, several splicing factors that activate AT-exons, including hnRNPK, hnRNPL, QKI, PTBP1, and hnRNPA1 are in competition with U2AF65 or SF1 for their binding to similar motifs[70-73]. While the competition for specific binding sites between splicing factors and U2AF65 or SF1 can result in exon skipping, an emerging theme is that this competition can actually increase exon inclusion[38,70,74-78]. Indeed, the binding of spliceosome-associated factors (e.g., SF1 or U2AF65) to pseudo- or decoy-signals can inhibit splicing by decreasing the efficiency of the spliceosome assembly[38,70,74-78]. The binding of splicing factors like hnRNPA1 and PTBP1 to decoy-splicing signals could "fill in" a surplus of splicing-related signals and consequently enhances the recognition of *bona fine* splicing sites[38,70,74-78]. Therefore, one interesting possibility is that high AT-load increases the probability of generating decoy splicing-related signals at intron 3'-ends, which decreases exon recognition but high AT-load simultaneously enhances the probability of recruiting splicing factors that cover decoy signals and therefore strengthen the recruitment of

spliceosome-related components (e.g., SF1 and U2AF65) to *bona fide* splicing sites and therefore enhance exon recognition. Supporting a model in which exons embedded in AT-rich regions are sensitive to spliceosome-associated factors defining intron 3'-ends, these exons had a higher probability to depend on SF1 and U2AF2 than GC-exons (Figure 3G). However, the interplay between local nucleotide composition bias affecting splicing-related features and the functions of splicing factors is expected to be more complex. Indeed, splicing factors activating GC-exons interact with splicing factors activating AT-exons and all these splicing factors interact with a wide array of spliceosome-associated components (Supplementary Figure S9).

It is now widely recognized that the synchronization between transcription and splicing plays an important role in splicing fidelity and efficiency as well as in alternative splicing regulation[6,7,11,12,79-84]. We propose that the coupling between transcription and splicing is operating through different mechanisms depending on the gene and local nucleotide composition bias impacting on chromatin organization and RNAPII dynamic. Indeed, at the chromatin level, nucleosomes are better positioned on exons in AT- than GC-rich context (Figure 5A and 5B), as already reported [14-20]. This could be the consequence of two interrelated phenomena. First, exons embedded in AT-rich context have a much higher GC-load than their flanking intronic sequences, in contrast to exons embedded in GC-rich context (Figure 2A). It has been proposed that the nature of GC-stacking interactions increases DNA structural polymorphisms that in turn increases DNA bendability and therefore favors DNA wrapping around nucleosomes[14-20]. Meanwhile, T- and A-rich sequences form more rigid structures that create nucleosome energetic barriers[14-20]. Therefore, increasing the intronic GC-load, which consequently decreases the density of Ts and As, will increase the probability of nucleosomes to slid from exons to introns. Meanwhile the decrease in intronic GC-load will favor nucleosome positioning on exons. As a consequence, transcription and splicing synchronization in an AT-rich environment could be local (at the exon level) and particularly dependent on nucleosomes well positioned on exons. Indeed, exonic nucleosome may increase exon recognition by locally slowing down RNAPII and favor the recruitment of splicing-related factors. In this setting, several components associated with the U2 snRNP interact with chromatin-associated factors[85-87]. Therefore, the increase in intronic AT-load would on the one hand creates decoy splicing-related signals (see above) while on the other hand enforce nucleosome-dependent splicing signals and slowing down locally RNAPII. Of note, large A- and T-stretches have also been shown to slow down RNAPII.

While local features (at the exon level) such as nucleosome positioning could mediate the synchronization of transcription and splicing when exons are within AT-rich environment, the coupling between transcription and splicing could be mediated by other mechanisms when exons are within a GC-rich environment. Indeed, both the higher density of nucleosomes across introns of GC-rich genes (Figure 5) as well as the higher stability of G:C base pairing, may create constraints

reducing RNAPII velocity across both exons and introns. Accordingly, the rate of elongation by RNAPII is negatively correlated with gene GC-content[88]. Therefore, an interesting possibility is that high gene GC-content may facilitate the synchronization between transcription and splicing by "smoothing" RNAPII dynamic all along the gene. In this setting, it must be pointed out that mRNA GC-content and RNA secondary structures have been proposed to enhance translation efficiency by "smoothing" translation elongation rate[89]. It is also interesting to note that extensive interactions between the U1 snRNP components and RNAPII-associated complexes have been reported[90,91]. RNAPII slowly moving through GC-rich gene could facilitate the recruitment of the U1 snRNP avoiding the formation of long and highly stable RNA secondary structures that would decrease the efficiency of exon recognition.

The interplay between gene GC-load, transcription and splicing is particularly interesting in regards to gene expression level. Indeed, it is now well established that transcription creates strong physical constrains on DNA and it has been proposed that a high density of GC nucleotides allows transcribed DNA regions to resist to transcriptional-mediated constraints. For example, GC dinucleotides form polymorphic structures in the DNA double helix that allow to "absorb" topological and torsional stresses (negative and positive supercoiling) generated during transcription, notably through the transition from B- to Z-DNA forms[92-97]. Therefore, a high GC-content may allow to increase the density of RNAPII on genes, in agreement with our observation (Figure 4). Accordingly, GC-rich genes are within gene-rich genomic domains and are highly expressed[22-26,34-36,42-45]. In this setting, RNA biogenesis and processing could further enhance transcription. Indeed, *in vitro* experiments demonstrated that GC-rich nascent RNA structures limit RNAPII pause by impeding backtracking along the template[98]. In addition, splicing can enhance transcription since for example, the U1 snRNP and the presence of 5' ss enhance transcription initiation and re-initiation[99]. Further supporting an important role of gene GC-load in the interplay between transcription, splicing and gene expression level, it has been shown that intron removal occurs more efficiently in highly-expressed genes[100] and GC-rich genomic regions preferentially associate with nuclear speckles rich in a variety of splicing factors[101-105]. In this setting, introns from the same genes have similar size and the same splicing rate[100] and co-transcriptionally co-regulated genes have similar nucleotide composition bias and similar intron size[106]. Collectively, these observations point out to a potential gene-specific and gene sequence-dependent nuclear compartmentalization of splicing.

Supporting this notion, we observed that AT- and GC-exons belong to different isochores and that a larger proportion of AT-exons are present in LADs (Figure 6). LAD represent specific AT-rich DNA regions attached to the nuclear lamina[53]. In addition, some TADs are enriched in GC-exons while others are enriched in AT-exons (Figure 6). TADs correspond to genomic regions that can be separated by several dozen and hundreds of Kb but that are in proximity in the nuclear space[54]. Genes belonging to the same TAD are associated with nucleosomes bearing the same histone

modifications, are transcriptionally co-regulated, and have similar nucleotide composition bias [22-26,46-52]. An emerging theme is that nucleotide composition bias determines the physical properties of a DNA region (e.g., its bendability) as well as the way nucleosomes are organized within this region. This next collectively determines in part the genome organization.

Based on these observations, we propose a model where gene transcription activity and/or genome organization create physical constraints on DNA driving nucleotide composition bias over dozen of Kbs (Figure 7). Nucleotide composition bias creates locally (at the exon level) constraints on the splicing process by impacting splicing-related features. However, nucleotide composition bias also creates "opportunities" for alleviating constraints and for regulatory processes. For example, exons embedded in AT-rich environment are weakened in terms of intron 3' end definition, which can be alleviated by an interplay between exonic nucleosomes, U2 snRNP-associated factors and splicing factors binding to AT-rich sequences (Figure 7). Meanwhile, exons embedded in GC-rich environment are weakened at their 5' ss because of the formation of RNA secondary structures which can be alleviated by an interplay between slow RNAPII, U1 snRNP-associated factors and splicing factors binding to GC-rich sequences. In this model, splicing factors would enhance the recognition of exons and therefore splicing by counteracting splicing-associated constraints resulting from nucleotide composition bias. This notion might be particularly relevant in light of recent discoveries indicating that the splicing process plays a major role in protecting DNA from transcriptional-associated physical constraints[107,108]. Indeed, the splicing process enhances the removal from the DNA template of the nascent transcript that would otherwise interact back to its template leading to DNA breaks[107,108]. Therefore, one of the main function of splicing factors could be to counterbalance constraints generated by gene nucleotide composition bias that otherwise decrease splicing efficiency, which would result in DNA damage. This model of course does not exclude the fact that splicing factors also contribute to increase the diversity of gene products.

Materials and Methods

RNA-seq dataset analyses

Publicly available RNA-seq datasets generated from different cell lines transfected with siRNAs or shRNAs targeting specific splicing factors or transfected with splicing factor expression vectors were recovered from GEO (Supplementary Table S1). These RNA-Seq datasets were analyzed using FARLINE, a computational program dedicated to analyze and quantify alternative splicing variations, as previously reported[32]. In this study, we focused on exons those inclusion depends on at least one splicing factor. All the genomic annotation (exons, introns, promoters) are from FASTERDB.

GC- and AT-rich exon groups

The percentage of nucleotide was measured for all human exons. The GC- and AT-rich exon groups were generated based on the results described on Figure 1E. Indeed, we generated sets of exons those inclusion is activated by splicing factors and defined two sets of splicing factors: those activating GC-rich exons flanked by small introns (i.e., SRSF9, PCBP1, RBMX, hnRNPF, RBFOX2, SRSF5, hnRNPH1, RBM22, RBM25, MBNL2, SRSF6, and SRSF1) and activating AT-rich exons flanked by large introns (i.e., TRA2A/B, RBM15, RBM39, hnRNPA2B1, KHSRP, hnRNPM, SRSF7, SFPQ, MBNL1, DAZAP1, PTBP1, hnRNPL, hnRNPK, FUS, QKI, hnRNPA1, PCBP2, hnRNPU). For further analyses, we eliminated exons (about ¼) that are found in the two sets leading to a list of 3182 GC-exons and a list of 4045 AT-exons.

Heatmaps

Each heatmap represents the relative median of a set of splicing-factor activated exons for a given feature. The formula (1) was used to compute the relative frequency of a feature D in a set of exons S = $\{x_1, \ldots, x_n\}$ such that x_i is an exon *i*.

$$RFreq(D) = \frac{Median(F_S) - Median(F_C)}{Median(F_C)} \times 100 \quad \textbf{(1)}$$

Where $F_S = \{v_1, \ldots, v_n\}$ such that v_i is the value of the exon *i* in S for the feature D and $F_C = \{ v_1, \ldots, v_m\}$ such that v_j is the value of the exon *j* in the control set of *m* exons noted C= $\{x_1, \ldots, x_m\}$. The control set of exon used corresponds to all human FasterDB exon.

Smallest flanking intron size

The size of every FasterDB introns was computed. Then, the size of every smallest flanking introns of a set of exons activated by any given splicing-factor was selected to build the figures.

Frequency Maps

The sequences corresponding to the intron-exon junctions (i.e the 100 last nucleotides of the upstream intron and the 50 first nucleotides of the downstream intron), have been recovered from a set of exons. Then, the frequency of a given nucleotide was computed at each positions of those

sequences. The same procedure was applied for the sequences corresponding to exon-intron junctions (I.e the 50 last nucleotides of the exon and the 100 first of the downstream intron).

Splicing sites score

The splice site scores were computed for each FasterDB exons with MaxEntScan[109]. MaxEntScan uses maximum entropy models to compute the likelihood of a sequence to be a real splicing site (the higher the score, the higher the probability that the associated sequence is a true splice-site).

Minimum free energy

The minimum free energy was computed on exonic junctions sequences using RNAFold from the VienneRNA package (v 2.4.1, 110). Analyzed sequences include 25 nucleotides within the intron and 25 nucleotides within the exon.

Branch point predictions

The number of putative branch points in a sequence corresponding to the 100 nucleotides preceding the 3'SS of a given exon was computed thanks to SVM-BP finder (111). Only the predicted branch point with a svm score > 0 were considered.

U2 binding energy

U2 binding energy corresponds to the number of hydrogen bounds between the nucleotides surrounding the branch point of an RNA sequence (without the branchpoint adenine) and the branch point binding sequence of U2 snRNA. The RNAduplex script in the ViennaRNA package (v2.4.1)[110] was used to determine the optimal hybridization structure between the branch point binding sequence of U2 snRNA (GUGUAGUA) and the RNA sequence. The RNA sequence is composed of 5 nucleotides before and 3 after the branch point. Then, the sum of hydrogen bound forming between the RNA and the U2 sequence are computed.

UNA motif counts

The number of UNA motifs (corresponding to UAA, UCA, UGA and UUA motifs) have been counted on the 50 last nucleotides of each intron upstream a set of interest exons.

Thymine-rich low complexity sequences

The number of thymine-rich low complexity sequences have been calculated using a sliding window with a size and a step of 4 and 1 nucleotide, respectively. We used this sliding window to read the intronic sequences from the 75^{th} to the 35^{th} nucleotides upstream the 3' splice site for every exon in a given set. We counted each window containing at least three thymines (I.e T-rich low complexity sequences). Then, we computed the proportion of exons in a given set with less or more than two T-rich low complexity sequences.

V-value : Exons regulation by U1 or U2 snrNP-associated factors

In figure 3G, we tested whether the proportions of GC- or AT-exons regulated by a given spliceosome associated factor were different. As we defined more AT-exons (4045) than GC-exons (3182), 10.000

random subsampling of AT-exons were made to have as many AT-exons in each sample than GC exons. Then, we compared the proportions of GC-exons and the sub-sampled AT-exons activated by a spliceosome-associated factor using a chi2 test for each AT-exons subsample. We then counter the number *x* of p-values lower than 0.05 among the 10.000 computed p-values for each spliceosome-associated factor. $n = \frac{x}{10.000}$ corresponds to an empirical p-value that summarize the 10.000 chi2 tests. Then we computed the value *v* for each spliceosome-associated factor using the formula :

$$v = log10\left(\frac{x}{10.000} + \varepsilon\right) \times s$$

Where s = 1 if the average proportion of the AT-exons samples activated by a given spliceosome associated factor is greater than the one of the GC-rich exons activated by the same factor. If the previous condition is not satisfied, then s = -1. $\varepsilon = 10^{-05}$ is present to allow the computation of log10(n) when n=0. In that case, the absolute maximum value of v is 4. *v* is a representation of the empirical p-value *n* that we call v-value. The doted line in the Figure 3G corresponds to log10(0.05). When the v-value is greater than log10(0.05) it means that the proportion of GC-exons activated by a spliceosome associated factor is significantly greater than the proportion of AT-exons activated by the same splicing factor. When the v-value is lower than -log10(0.05) it means that the proportion of GC-exons activated by a spliceosome associated factor is significantly lower than the proportion of AT-exons activated by the same splicing factor.

CLIPseq

Bed files from publicly available CLIPseq datasets generated using different antibodies and peak callers as described in Supplementary Table S2 were used to generate density maps. The bed files were first sorted and transformed into bedGraph files using the bedtools suite (v 2.25.0)[112] . The bedGraph files are then converted into bigWig files thank to bedGraphToBigWig (v.4)[113]. The 5'SS and 3'SS regions (from the splicing site, 200 nucleotides in intron and 50b in exon) are considered. The proportion of GC-, AT-exons or exons activated by SF1 and SNRPC having CLIP peak signals at each nucleotide of the 5'SS and 3'SS is computed at each nucleotide position.

Analysis of publicly available ChIPseq of RNAPII, H3, Mnase, and histone marks.

ChIP-seq or MNase-seq datasets were recovered from the Cistrome, ENCODE, and GEO databases. The coverage files (BigWig) were directly downloaded for datasets from Cistrome and for RNAPII ChIP-seq from ENCODE. Else, raw data were downloaded for analysis with homemade pipeline. The reads were trimmed and filter for minimum length of 25b using Cutadapt 1.16 (options: -m 25). Then, reads were trimmed at their 3' end for minimum quality of 20 (-q 20) and filter for minimum length of 25b (-m 25) in two consecutive steps. The processed reads were mapped to hg19 with Bowtie2 2.3.3 (options: --very-sensitive --fr -I 100 -X 300 --no-mixed) and filter for mapping quality over 10

with samtools view 1.6 (options: -b -q 10). If ChIP-seq experiments were generated using sonication only, the duplicates were removed with homemade tools, which checks for coordinates and CIGAR of the read and of the read 2 if paired-end sequencing. The fragments were reconstituted from the reads and fragment-coverage files were built with MACS2 2.1.1.20160309 (options: -g hs -B). The metaplots of ChIP/MNase-seq on genes were generated by recovering the fragment coverage (promoter: -1500b/+500 from the TSS; first exon, internal exons and introns: according to the coordinates of the annotation; regulated exons: -100 / +100 from the center or 500b in the intron and 50b in the exon from the splicing site [MNase and H3]; whole gene: according to the coordinates of the annotations and -200/+200 from the annotation). The annotations were lifted over from hg19 to hg38 if the coverage file came from Cistrome database. In the case of RNAPII coverage, only exons regulated in the corresponding cell line, or annotation from the gene hosting them, were considered. For internal exons, and introns, the coverages of the annotations from the same gene were concatenated respecting the order. Then the coverage recovered according to the coordinates of the annotations were split in 1000 bins. The 199 first bins of the "internal exons" or of the introns were removed to avoid display of signal influenced by the promoter. The metaplot were build by computing, at each position or bin, the average coverage across the annotations. Statistics were done by comparing average coverage in the annotations from two groups with a Wilcoxon's test. Average of the mean coverage on each regulated exons (-100/+100 from center) were computed. For each ChIP-seq experiment, the ratio of the averages : "GC" - "AT" / max("GC", "AT") was computed and used to build the boxplot in Figure 5C, and the statistics were made with a Wilcoxon's test of the averages per experiments of "GC" vs. "AT" exons.

Isochores, LADs and TADs

Three LADs and three TADs datasets were recovered from GEO (see Supplementary Table S1). Two Isochore datasets were retrieved from Costantini et al. (BMC Genomics. 2009 Apr 3;10:146) and Oliver et al. (Nucleic Acids Research 32: W287-W292). A third isochore dataset was computed with Isosegmenter (Cozzi P, et al. Bioinformatics 2015;11:253-261.). Isochore groups are labelled L1, L2, H1, H2 and H3 according to their GC% content (<37%, 37-41%, 41-46%, 46-53% and >53%, respectively). We used bedtools intersect command (Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 2014 Sep 8;47:11.12.1-34. PMID: 25199790) to determine for each exons and for each splicing factor-activated GC- or AT-exons at which isochore group they belonged to. This protocol was used for each isochore dataset, separately. The same bedtools intersect command was used to determine for each AT- and GC-exon if they were co-localized with LAD or TAD domains.

Figure legends

Figure 1

A. Right Panel: Heat map of the relative median frequency of GC and AT nucleotides in splicing factor-activated exons. Left Panel: Heatmap of the relative frequency of GC and AT nucleotides in the 100 intronic nucleotides upstream and downstream. splicing factor-activated exons. (*) Mann-Whitney test, FDR < 0,05.

B. Correlation between the relative GC content (when compared to all human exons) of splicing factor-activated exons and the relative GC content of their respective upstream intron (window of 100 nucleotides before the exons, upper panel, correlation coefficient = 0,77 and Pearson correlation P-value < 10^{-16}), or their respective downstream intron (window of 100 nucleotides after the exons, lower panel, correlation coefficient = 0,76 and Pearson correlation P-value < 10^{-16}).

C. Heat map of the relative median size of the smallest intron flanking splicing factor activated exons. The sets of splicing factor-activated exons are represented in the same order than in Figure 1A. (*) Mann-Whitney test, FDR < 0,05.

D. Correlation between the relative median size, in logarithmic scale, of the smallest introns flanking splicing factor-activated exons and their relative exonic GC-content (upper panel). or intronic GC content (lower panel).The Pearson correlation coefficients are -0.33 and -0,41 and the Pearson correlation test p-values are < 10^{-16} for the upper and lower panel respectively.

E. The x-axis represents the relative (%) median smallest size of the two introns flanking splicing-factor activated exons corresponding to each analyzed splicing factors when compared to all human introns. The y-axis represents the relative (%) median GC-content of splicing-factor activated exons corresponding to each analyzed splicing factors when compared to all human exons. Splicing factors in blue activate exons that are on average GC-rich and flanked by small introns when compared to all human exons. The splicing factors in green activate exons that are on average AT-rich and flanked by large introns when compared to all human exons. Pearson correlation coefficient = -0.83 and Pearson correlation test p-value = $3.06e^{-09.}$

F. Left panel: Violin plot representing the GC content of every exon activated by splicing factors that activate either GC- or AT-exons. Right panel: Violin plot representing the logarithmic nucleotide size of the smallest intron flanking each exon activated by splicing factors that activate either GC- or AT-exons in average. (***) Mann-Whitney test p-value < 0.001

Figure 2

A. Nucleotide density (%) maps in exons and their flanking sequences.

B. Heat map of the relative median frequency (%, when compared to all human exons) of A, T, G, or C nucleotides in a window of 25 nucleotides downstream GC-exons (left panel) or AT-exons (right panel).

C. Heat map of the relative frequency (%, when compared to all human exons) of A, T, G, or C nucleotides in a window of 25 nucleotides upstream GC-exons (left panel) or AT-exons (right panel).

D. Minimum free energy at the 5'ss (left panel) and at the 3'ss (right panel) of GC- and AT-exons. The minimum free energy is computed on intron-exon junction sequences of 50 nucleotides, with 25 nucleotides within the exons and 25 nucleotides within the intron. The red lines indicate the median values calculated for all human exons. (***) Mann-Whitney test p-values < 0.001

E. Proportion (%, left panel) of GC- or AT-exons having less or more than two predicted BPs in a window of 100 nucleotides in their upstream intron. The red lines indicate the proportion calculated for all human exons. (***) Chi-2 test p-value < 0.001. Number of H bounds measured between the U2 snRNA and the BP sequence found in the 25 last nucleotides of the upstream intron (right panel) of GC- and AT-exons. The red lines indicate the median values calculated for all human exons. (*) Mann-Whitney test p-values < 0.001

F. Weblogos generated using sequences flanking predicted BPs upstream GC- and AT-exons. The BP were predicted using the sequence corresponding to the 25 last nucleotides of the upstream intron of GC- and AT-exons.

G. Left panel: proportion (%) of GC- or AT-exons having at least two or more than two UNA sequences in a window corresponding to the last of 50 nucleotides in their upstream intron. (***) Chi-2 test p-value < 0.001. Right panel: proportion (%) of GC- or AT-exons having at least two or more than two T-rich low complexity sequences (at least three T in 4 consecutives nucleotides) in a window from positions -35 and -75 in regards to the 3'ss. The red lines indicate the average values calculated for all human exons. (***) Chi-2 test p-value < 0.001.

H. Density of reads obtained from publicly available U2AF2-ClIP datasets generated in HEK293T (left panel) or HeLa (right panel) cells and mapping upstream GC- and AT-exons. The green arrows indicate reads mapping upstream the Py tract.

Figure 3

A. Nucleotide frequency (%) maps in exons and their flanking sequences.

B. Minimum Free energy at the 5' ss exons activated by different spliceosome-associated factors The minimum free energy is computed on intron-exon junction sequences of 50 nucleotides with 25 nucleotides within the exons and 25 nucleotides within the intron The red lines indicate the median values calculated for all human exons. (***) Chi2-test FDR < 0.001; (**) chi2-test FDR < 0.01.

C. Proportion (%) of exons activated by spliceosome-associated factors having at least two or more than two predicted BPs in a window corresponding to the 100 last nucleotides in their upstream intron. The red lines indicate the proportion values calculated for all human exons. (***) Chi2 test FDR < 0.001

D. Proportion (%) of exons activated by spliceosome-associated factors having at least two or more than two UNA sequences in a window corresponding to the last 50 nucleotides in their upstream intron. The red lines indicate the proportion values calculated for all human exons (*) Chi2 test FDR < 0.05; (***) Chi2 test FDR < 0.001.

E. Proportion (%)of exons activated by spliceosome-associated factors having at least two or more than two T-rich low complexity sequences (at least three T in four consecutives nucleotides) in a window from positions -35 and -75 in regards to the 3' ss. The red lines indicate the proportion calculated for all human exons. Chi2 test FDR < 0.001

F. Density of reads from U2AF2-ClIP publicly available datasets generated in HEK293T (left panel) or HeLa (right panel) cell lines and mapping upstream SNRPC- or SF1-dependant exons. The green arrows indicate reads mapping upstream the Py tract.

G.V-value (see-method) of every spliceosome-associated factors. A v-value above the doted line is considered as significant.

Figure 4

A. Correlation between the relative GC content (when compared to all human exons) of splicing factor-activated exons and the relative GC content of their hosting genes (left panel). Pearson correlation coefficient = 0.72 and Pearson correlation test p-value < 10^{-16}. Correlation between the relative GC content and the relative median intron size of the genes hosting splicing factor-activated exons (right panel). Pearson correlation coefficient = -0,62 and Pearson correlation test p-value < 10^{-16}.

B. Violin plots representing the GC content (%, left panel), the intron size (middle panel), and the size (right panel) of genes hosting GC- and AT-exons. The red lines indicate the average values calculated for all human exons. (***) Mann-Whitney test p-value < 0.001.

C. Density of reads obtained after immunoprecipitation of RNAPII in K562 and HepG2 cell lines. Different parts of the genes hosting GC- and AT-exons are represented (see Methods).

D. Box plot of the mean coverage of GC- and AT-exons by RNAPII . (***) Wilcoxon's test p-value < 10^{-6}.

E. Density of reads obtained after immunoprecipitation of RNAPII phosphorylated on serine 2 (RNAPII-ser2) in K562 and HepG2 cell lines. Different parts of the genes hosting GC- and AT-exons are represented.

F. Box plot of the mean coverage of GC- and AT-exons by phosphorylated RNAPII. (***) Wilcoxon's test p-value < 10^{-6}.

Figure 5

A. Density of reads mapping to GC- and AT-exons and obtained after DNA treatment with MNase (left panels) or after immunoprecipitation of the histone H3 (right panels) in K562, HEK293T, and Hela cell lines.

B. Density of reads mapping across different parts of the genes hosting GC- and AT-exons and obtained after DNA treatment with MNase (left panels) or after immunoprecipitation of the histone H3 (right panels) in K562, HEK293T, and Hela cell lines.

C. Box plots of the mean coverage of GC- and AT-exons by H3. (***) Wilcoxon's test p-value < 10^{-6}.

D. Box plot representing the density ratio of reads mapping to GC- versus AT-exons and obtained after immunoprecipitation of DNA using antibodies against different histone modification, as indicated. Each box plot represent the values obtained from several publicly available datasets generated in different experiments and/or different cell lines. For each ChIP-seq dataset, the ratio of the averages : "GC" - "AT" / max("GC", "AT") was computed and used to build the boxplot. (***) Wilcoxon's test p-value < 10^{-6}.

E. Density of reads mapping across different parts of genes hosting GC- and AT-exons and obtained from the K562 cell line after immunoprecipitation of DNA using antibodies against different histone modifications, as indicated.

F. Box plot of the mean coverage of GC- and AT-exons by H3K4me3, H3Kças, H3K36me3, or H3K9me3. (***) Wilcoxon's test p-value < 10^{-6}.

Figure 6

A. Proportion of AT-, GC-, and all human exons distributed across different isochore families.

B. Number of AT- and GC-exons present in individual isochores defined by Isofinder (see Methods). Only isochores containing at least five GC-exons or five AT-exons are represented. The dashed line separates isochores containing preferentially GC-exons to isochores containing preferentially AT-exons.

C. Proportion of AT- and GC-exons in LADs annotated from three different datasets and in TADs annotated from three different datasets.

D. Number of AT- and GC-exons present in individual TADs annotated from the K562 cell line. Only TADs containing at least seven GC-exons or seven AT-exons are represented. The dashed line separates TADs containing preferentially GC-exons and TADs containing preferentially AT-exons.

Figure 7

Nucleotide composition bias determines the physical properties of the DNA polymer and therefore contribute to its i) folding and shape in the nuclear space and ii) resistance to transcriptional-associated physical constraints. Accordingly, GC-rich isochores and TADs contain a large number of genes ("Gene core") that are GC-rich and that contain small introns. Meanwhile, AT-rich isochores, TADs, and LADs contain a small number of genes ("Gene desert") that are AT-rich and that contain large introns. The regional nucleotide composition bias increases the probability of local (e.g., at the exon level) nucleotide composition bias. Local nucleotide composition bias influences local chromatin organization at the DNA level but also the splicing process at the RNA level. The high density of nucleosomes and GC nucleotides (upper panel) could generate a "smoothly" transcription across small genes favoring the synchronization (or coupling) between transcription and splicing. The high density of GC nucleotides increases the probability of secondary structures at the 5' ss, which consequences on splicing recognition during the splicing process. This constraint could be alleviated by the binding to GC-rich sequences of splicing factors enhancing the recruitment of the U1 snRNP. The high density of AT nucleotide (local panel) could favor a sharp difference between exon and intron in terms of nucleotide composition bias, which would favor nucleosome positioning on exons. A- and T-rich sequences located upstream AT-exons as well as exonic nucleosome could locally (at the exon level) slow down RNAPII favoring the synchronization (or coupling) between transcription and splicing. The high density of AT nucleotides increases the probability of generating decoy signals like pseudo BPs, SF1- or U2AF2 binding sites. This constraint could be alleviated by the binding to these decoy signals of splicing factors enhancing the recruitment of the U2 snRNP.

References

1 Wahl, M. C., Will, C. L. & Luhrmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**, 701-718, doi:10.1016/j.cell.2009.02.009 (2009).

2 Lin, C. L., Taggart, A. J. & Fairbrother, W. G. RNA structure in splicing: An evolutionary perspective. *RNA Biol* **13**, 766-771, doi:10.1080/15476286.2016.1208893 (2016).

3 Zhang, J., Kuo, C. C. & Chen, L. GC content around splice sites affects splicing through pre-mRNA secondary structures. *BMC Genomics* **12**, 90, doi:10.1186/1471-2164-12-90 (2011).

4 Gahura, O., Hammann, C., Valentova, A., Puta, F. & Folk, P. Secondary structure is required for 3' splice site recognition in yeast. *Nucleic Acids Res* **39**, 9759-9767, doi:10.1093/nar/gkr662 (2011).

5 Lin, C. L. *et al.* RNA structure replaces the need for U2AF2 in splicing. *Genome Res* **26**, 12-23, doi:10.1101/gr.181008.114 (2016).

6 Braunschweig, U., Gueroussov, S., Plocik, A. M., Graveley, B. R. & Blencowe, B. J. Dynamic integration of splicing within gene regulatory pathways. *Cell* **152**, 1252-1269, doi:10.1016/j.cell.2013.02.034 (2013).

7 Fu, X. D. & Ares, M., Jr. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat Rev Genet* **15**, 689-701, doi:10.1038/nrg3778 (2014).

8 Dominguez, D. *et al.* Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Mol Cell* **70**, 854-867 e859, doi:10.1016/j.molcel.2018.05.001 (2018).

9 Giudice, G., Sanchez-Cabo, F., Torroja, C. & Lara-Pezzi, E. ATtRACT-a database of RNA-binding proteins and associated motifs. *Database (Oxford)* **2016**, doi:10.1093/database/baw035 (2016).

10 Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172-177, doi:10.1038/nature12311 (2013).

11 Hollander, D., Naftelberg, S., Lev-Maor, G., Kornblihtt, A. R. & Ast, G. How Are Short Exons Flanked by Long Introns Defined and Committed to Splicing? *Trends Genet* **32**, 596-606, doi:10.1016/j.tig.2016.07.003 (2016).

12 Naftelberg, S., Schor, I. E., Ast, G. & Kornblihtt, A. R. Regulation of alternative splicing through coupling with transcription and chromatin structure. *Annu Rev Biochem* **84**, 165-198, doi:10.1146/annurev-biochem-060614-034242 (2015).

13 Dujardin, G. *et al.* How slow RNA polymerase II elongation favors alternative exon skipping. *Mol Cell* **54**, 683-690, doi:10.1016/j.molcel.2014.03.044 (2014).

14 Iyer, V. R. Nucleosome positioning: bringing order to the eukaryotic genome. *Trends Cell Biol* **22**, 250-256, doi:10.1016/j.tcb.2012.02.004 (2012).

15 Trifonov, E. N. Cracking the chromatin code: precise rule of nucleosome positioning. *Phys Life Rev* **8**, 39-50, doi:10.1016/j.plrev.2011.01.004 (2011).

16 Chen, W., Luo, L. & Zhang, L. The organization of nucleosomes around splice sites. *Nucleic Acids Res* **38**, 2788-2798, doi:10.1093/nar/gkq007 (2010).

17 Tilgner, H. *et al.* Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* **16**, 996-1001, doi:10.1038/nsmb.1658 (2009).

18 Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362-366, doi:10.1038/nature07667 (2009).

19 Segal, E. & Widom, J. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr Opin Struct Biol* **19**, 65-71, doi:10.1016/j.sbi.2009.01.004 (2009).

20 Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol* **16**, 990-995, doi:10.1038/nsmb.1659 (2009).

21 Luco, R. F. *et al.* Regulation of alternative splicing by histone modifications. *Science* **327**, 996-1000, doi:10.1126/science.1184208 (2010).

22 Costantini, M. & Musto, H. The Isochores as a Fundamental Level of Genome Structure and Organization: A General Overview. *J Mol Evol* **84**, 93-103, doi:10.1007/s00239-017-9785-9 (2017).

23 Jabbari, K. & Bernardi, G. An Isochore Framework Underlies Chromatin Architecture. *PLoS One* **12**, e0168023, doi:10.1371/journal.pone.0168023 (2017).

24 Arhondakis, S., Auletta, F. & Bernardi, G. Isochores and the regulation of gene expression in the human genome. *Genome Biol Evol* **3**, 1080-1089, doi:10.1093/gbe/evr017 (2011).

25 Schmidt, T. & Frishman, D. Assignment of isochores for all completely sequenced vertebrate genomes using a consensus. *Genome Biol* **9**, R104, doi:10.1186/gb-2008-9-6-r104 (2008).

26 Bernardi, G. Genome Organization and Chromosome Architecture. *Cold Spring Harb Symp Quant Biol* **80**, 83-91, doi:10.1101/sqb.2015.80.027318 (2015).

27 Amit, M. *et al.* Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep* **1**, 543-556, doi:10.1016/j.celrep.2012.03.013 (2012).

28 Zhang, X. H., Leslie, C. S. & Chasin, L. A. Dichotomous splicing signals in exon flanks. *Genome Res* **15**, 768-779, doi:10.1101/gr.3217705 (2005).

29 Fox-Walsh, K. L. *et al.* The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc Natl Acad Sci U S A* **102**, 16176-16181, doi:10.1073/pnas.0508489102 (2005).

30 Lim, L. P. & Burge, C. B. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A* **98**, 11193-11198, doi:10.1073/pnas.201407298 (2001).

31 Nicolas Fontrodona, F. A., Jean-Baptiste Claude, Helene Polveche, Sebastien Lemaire, Leon Charles Tranchevent, Laurent Modolo, Franck Mortreux, Cyril Bourgeois, Didier Auboeuf. Interplay between coding and exonic splicing regulatory sequences. *bioRxiv 334839; doi: https://doi.org/10.1101/334839* **In revision in Genome Research**.

32 Benoit-Pilven, C. *et al.* Complementarity of assembly-first and mapping-first approaches for alternative splicing annotation and differential analysis from RNAseq data. *Sci Rep* **8**, 4307, doi:10.1038/s41598-018-21770-7 (2018).

33 Vinogradov, A. E. Within-intron correlation with base composition of adjacent exons in different genomes. *Gene* **276**, 143-151 (2001).

34 Pozzoli, U. *et al.* Both selective and neutral processes drive GC content evolution in the human genome. *BMC Evol Biol* **8**, 99, doi:10.1186/1471-2148-8-99 (2008).

35 Versteeg, R. *et al.* The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res* **13**, 1998-2004, doi:10.1101/gr.1649303 (2003).

36 Zhu, L. *et al.* Patterns of exon-intron architecture variation of genes in eukaryotic genomes. *BMC Genomics* **10**, 47, doi:10.1186/1471-2164-10-47 (2009).

37 Mercer, T. R. *et al.* Genome-wide discovery of human splicing branchpoints. *Genome Res* **25**, 290-303, doi:10.1101/gr.182899.114 (2015).

38 Paggi, J. M. & Bejerano, G. A sequence-based, deep learning model accurately predicts RNA splicing branchpoints. *RNA* **24**, 1647-1658, doi:10.1261/rna.066290.118 (2018).

39 Corvelo, A., Hallegger, M., Smith, C. W. & Eyras, E. Genome-wide association between branch point properties and alternative splicing. *PLoS Comput Biol* **6**, e1001016, doi:10.1371/journal.pcbi.1001016 (2010).

40 Dardenne, E. *et al.* RNA helicases DDX5 and DDX17 dynamically orchestrate transcription, miRNA, and splicing programs in cell differentiation. *Cell Rep* **7**, 1900-1913, doi:10.1016/j.celrep.2014.05.010 (2014).

41 Kar, A. *et al.* RNA helicase p68 (DDX5) regulates tau exon 10 splicing by modulating a stem-loop structure at the 5' splice site. *Mol Cell Biol* **31**, 1812-1821, doi:10.1128/MCB.01149-10 (2011).

42 Gul, I. S. *et al.* GC Content of Early Metazoan Genes and Its Impact on Gene Expression Levels in Mammalian Cell Lines. *Genome Biol Evol* **10**, 909-917, doi:10.1093/gbe/evy040 (2018).

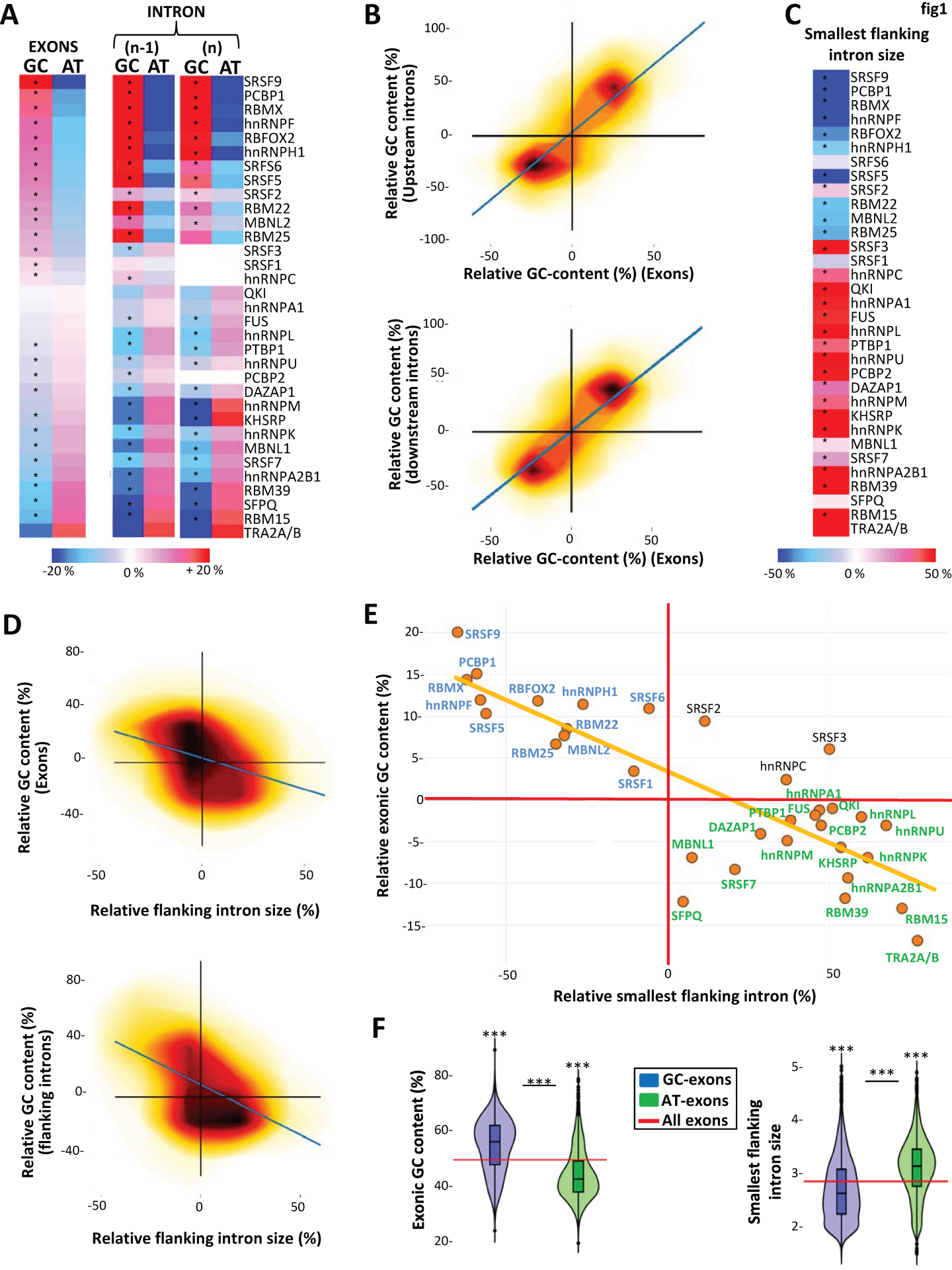
43 Frousios, K. *et al.* Transcriptome map of mouse isochores in embryonic and neonatal cortex. *Genomics* **101**, 120-124, doi:10.1016/j.ygeno.2012.11.006 (2013).

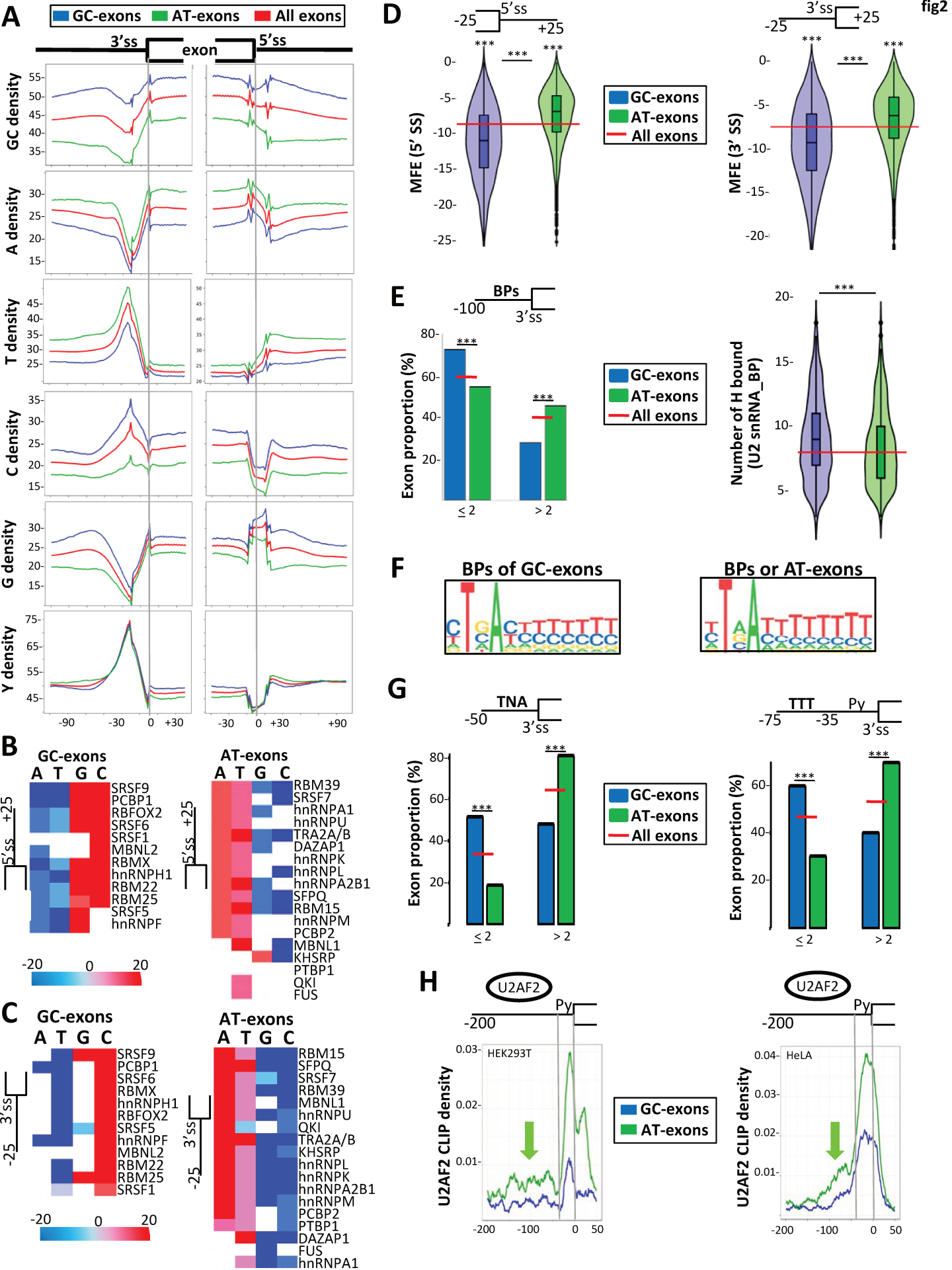
44 Kudla, G., Lipinski, L., Caffin, F., Helwak, A. & Zylicz, M. High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol* **4**, e180, doi:10.1371/journal.pbio.0040180 (2006).
45 Urrutia, A. O. & Hurst, L. D. The signature of selection mediated by expression on human genes. *Genome Res* **13**, 2260-2264, doi:10.1101/gr.641103 (2003).
46 Bessiere, C. *et al.* Probing instructions for expression regulation in gene nucleotide compositions. *PLoS Comput Biol* **14**, e1005921, doi:10.1371/journal.pcbi.1005921 (2018).
47 Todolli, S., Perez, P. J., Clauvelin, N. & Olson, W. K. Contributions of Sequence to the Higher-Order Structures of DNA. *Biophys J* **112**, 416-426, doi:10.1016/j.bpj.2016.11.017 (2017).
48 De Santis, P. & Scipioni, A. Sequence-dependent collective properties of DNAs and their role in biological systems. *Phys Life Rev* **10**, 41-67, doi:10.1016/j.plrev.2013.01.004 (2013).
49 Takasuka, T. E., Cioffi, A. & Stein, A. Sequence information encoded in DNA that may influence long-range chromatin structure correlates with human chromosome functions. *PLoS One* **3**, e2643, doi:10.1371/journal.pone.0002643 (2008).
50 Paz, A., Frenkel, S., Snir, S., Kirzhner, V. & Korol, A. B. Implications of human genome structural heterogeneity: functionally related genes tend to reside in organizationally similar genomic regions. *BMC Genomics* **15**, 252, doi:10.1186/1471-2164-15-252 (2014).
51 Kupper, K. *et al.* Radial chromatin positioning is shaped by local gene density, not by gene expression. *Chromosoma* **116**, 285-306, doi:10.1007/s00412-007-0098-4 (2007).
52 Liu, S. *et al.* From 1D sequence to 3D chromatin dynamics and cellular functions: a phase separation perspective. *Nucleic Acids Res* **46**, 9367-9383, doi:10.1093/nar/gky633 (2018).
53 van Steensel, B. & Belmont, A. S. Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell* **169**, 780-791, doi:10.1016/j.cell.2017.04.022 (2017).
54 Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat Rev Genet* **19**, 789-800, doi:10.1038/s41576-018-0060-8 (2018).
55 Huelga, S. C. *et al.* Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Rep* **1**, 167-178, doi:10.1016/j.celrep.2012.02.001 (2012).
56 Xu, Y. *et al.* Cell type-restricted activity of hnRNPM promotes breast cancer metastasis via regulating alternative splicing. *Genes Dev* **28**, 1191-1203, doi:10.1101/gad.241968.114 (2014).
57 Wang, E. & Cambi, F. Heterogeneous nuclear ribonucleoproteins H and F regulate the proteolipid protein/DM20 ratio by recruiting U1 small nuclear ribonucleoprotein through a complex array of G runs. *J Biol Chem* **284**, 11194-11204, doi:10.1074/jbc.M809373200 (2009).
58 Huang, S. C. *et al.* RBFOX2 promotes protein 4.1R exon 16 selection via U1 snRNP recruitment. *Mol Cell Biol* **32**, 513-526, doi:10.1128/MCB.06423-11 (2012).
59 Akker, S. A. *et al.* Pre-spliceosomal binding of U1 small nuclear ribonucleoprotein (RNP) and heterogenous nuclear RNP E1 is associated with suppression of a growth hormone receptor pseudoexon. *Mol Endocrinol* **21**, 2529-2540, doi:10.1210/me.2007-0038 (2007).
60 Rasche, N. *et al.* Cwc2 and its human homologue RBM22 promote an active conformation of the spliceosome catalytic centre. *EMBO J* **31**, 1591-1604, doi:10.1038/emboj.2011.502 (2012).
61 Cho, S. *et al.* Interaction between the RNA binding domains of Ser-Arg splicing factor 1 and U1-70K snRNP protein determines early spliceosome assembly. *Proc Natl Acad Sci U S A* **108**, 8233-8238, doi:10.1073/pnas.1017700108 (2011).
62 Kanno, T., Lin, W. D., Chang, C. L., Matzke, M. & Matzke, A. J. M. A Genetic Screen Identifies PRP18a, a Putative Second Step Splicing Factor Important for Alternative Splicing and a Normal Phenotype in Arabidopsis thaliana. *G3 (Bethesda)* **8**, 1367-1377, doi:10.1534/g3.118.200022 (2018).

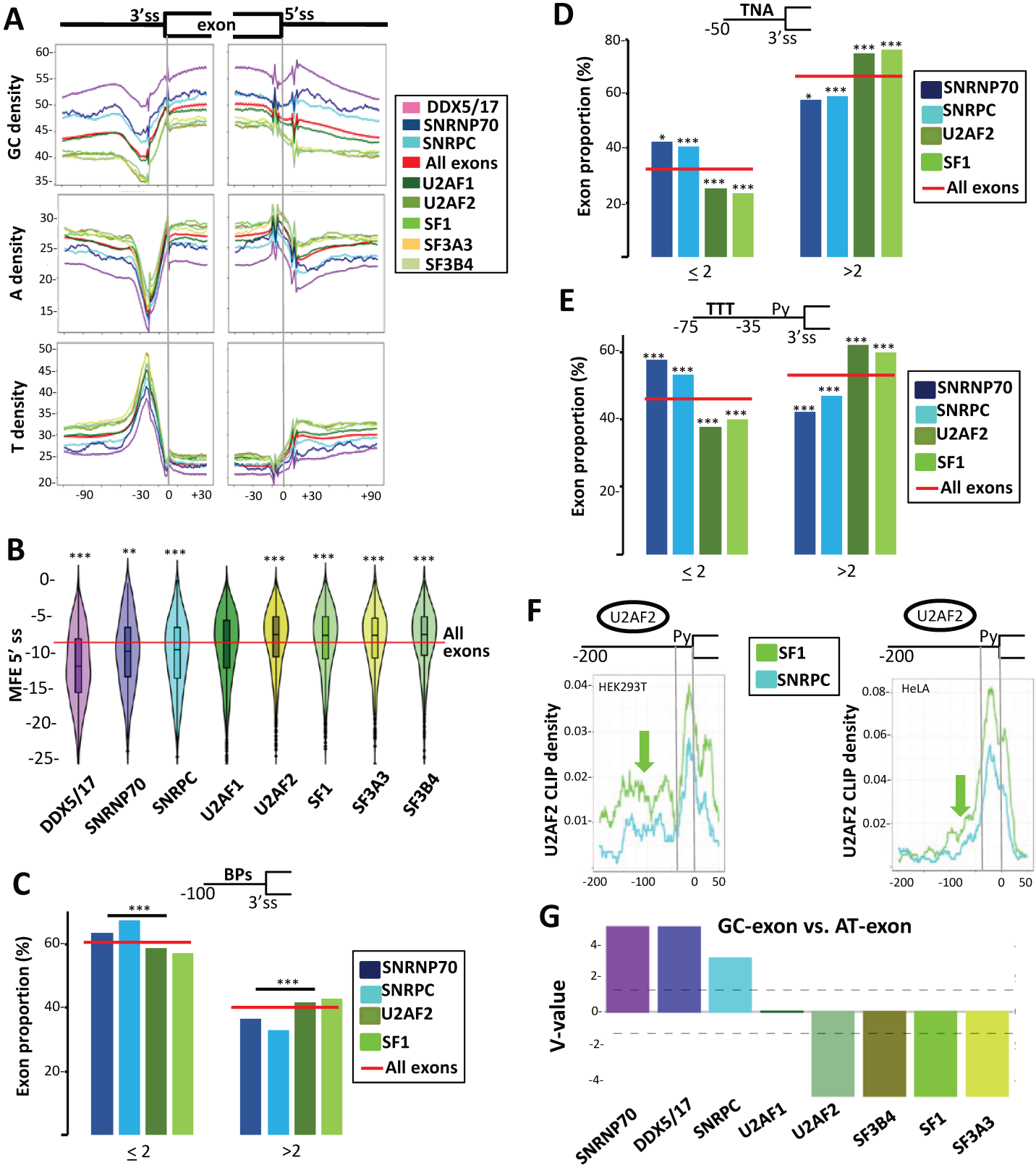
63 Murray, J. I., Voelker, R. B., Henscheid, K. L., Warf, M. B. & Berglund, J. A. Identification of motifs that function in the splicing of non-canonical introns. *Genome Biol* **9**, R97, doi:10.1186/gb-2008-9-6-r97 (2008).
64 Tavanez, J. P., Madl, T., Kooshapur, H., Sattler, M. & Valcarcel, J. hnRNP A1 proofreads 3' splice site recognition by U2AF. *Mol Cell* **45**, 314-329, doi:10.1016/j.molcel.2011.11.033 (2012).
65 Gozani, O., Patton, J. G. & Reed, R. A novel set of spliceosome-associated proteins and the essential splicing factor PSF bind stably to pre-mRNA prior to catalytic step II of the splicing reaction. *EMBO J* **13**, 3356-3367 (1994).
66 Zhang, L. *et al.* Cross-talk between PRMT1-mediated methylation and ubiquitylation on RBM15 controls RNA splicing. *Elife* **4**, doi:10.7554/eLife.07938 (2015).
67 Mai, S. *et al.* Global regulation of alternative RNA splicing by the SR-rich protein RBM39. *Biochim Biophys Acta* **1859**, 1014-1024, doi:10.1016/j.bbagrm.2016.06.007 (2016).
68 Wu, J. Y. & Maniatis, T. Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell* **75**, 1061-1070 (1993).
69 Cho, S. *et al.* hnRNP M facilitates exon 7 inclusion of SMN2 pre-mRNA in spinal muscular atrophy by targeting an enhancer on exon 7. *Biochim Biophys Acta* **1839**, 306-315, doi:10.1016/j.bbagrm.2014.02.006 (2014).
70 Howard, J. M. *et al.* HNRNPA1 promotes recognition of splice site decoys by U2AF2 in vivo. *Genome Res* **28**, 689-698, doi:10.1101/gr.229062.117 (2018).
71 Zong, F. Y. *et al.* The RNA-binding protein QKI suppresses cancer-associated aberrant splicing. *PLoS Genet* **10**, e1004289, doi:10.1371/journal.pgen.1004289 (2014).
72 Liu, G. *et al.* A conserved serine of heterogeneous nuclear ribonucleoprotein L (hnRNP L) mediates depolarization-regulated alternative splicing of potassium channels. *J Biol Chem* **287**, 22709-22716, doi:10.1074/jbc.M112.357343 (2012).
73 Cao, W., Razanau, A., Feng, D., Lobo, V. G. & Xie, J. Control of alternative splicing by forskolin through hnRNP K during neuronal differentiation. *Nucleic Acids Res* **40**, 8059-8071, doi:10.1093/nar/gks504 (2012).
74 Pineda, J. M. B. & Bradley, R. K. Most human introns are recognized via multiple and tissue-specific branchpoints. *Genes Dev* **32**, 577-591, doi:10.1101/gad.312058.118 (2018).
75 Sutandy, F. X. R. *et al.* In vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on regulation by cofactors. *Genome Res* **28**, 699-713, doi:10.1101/gr.229757.117 (2018).
76 Chen, L. *et al.* Stoichiometries of U2AF35, U2AF65 and U2 snRNP reveal new early spliceosome assembly pathways. *Nucleic Acids Res* **45**, 2051-2067, doi:10.1093/nar/gkw860 (2017).
77 Wu, T. & Fu, X. D. Genomic functions of U2AF in constitutive and regulated splicing. *RNA Biol* **12**, 479-485, doi:10.1080/15476286.2015.1020272 (2015).
78 Shao, C. *et al.* Mechanisms for U2AF to define 3' splice sites and regulate alternative splicing in the human genome. *Nat Struct Mol Biol* **21**, 997-1005, doi:10.1038/nsmb.2906 (2014).
79 Aslanzadeh, V., Huang, Y., Sanguinetti, G. & Beggs, J. D. Transcription rate strongly affects splicing fidelity and cotranscriptionality in budding yeast. *Genome Res* **28**, 203-213, doi:10.1101/gr.225615.117 (2018).
80 Wan, Y. & Larson, D. R. Splicing heterogeneity: separating signal from noise. *Genome Biol* **19**, 86, doi:10.1186/s13059-018-1467-4 (2018).
81 Davis-Turak, J., Johnson, T. L. & Hoffmann, A. Mathematical modeling identifies potential gene structure determinants of co-transcriptional control of alternative pre-mRNA splicing. *Nucleic Acids Res* **46**, 10598-10607, doi:10.1093/nar/gky870 (2018).
82 Chang, S. L., Wang, H. K., Tung, L. & Chang, T. H. Adaptive transcription-splicing resynchronization upon losing an essential splicing factor. *Nat Ecol Evol* **2**, 1818-1823, doi:10.1038/s41559-018-0684-2 (2018).

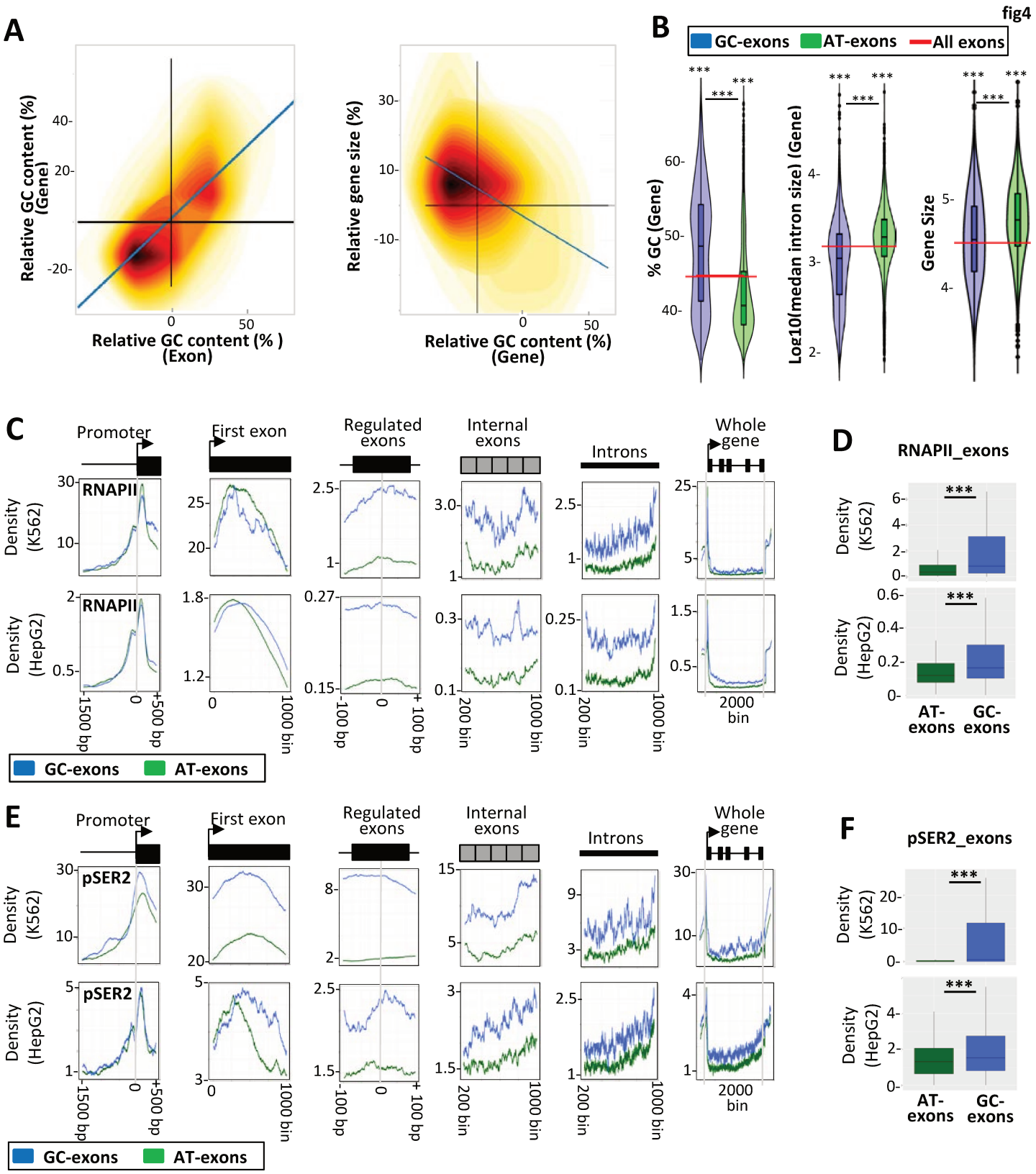
83 Brzyzek, G. & Swiezewski, S. Mutual interdependence of splicing and transcription elongation. *Transcription* **6**, 37-39, doi:10.1080/21541264.2015.1040146 (2015).
84 Fong, N. *et al.* Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev* **28**, 2663-2676, doi:10.1101/gad.252106.114 (2014).
85 Kfir, N. *et al.* SF3B1 association with chromatin determines splicing outcomes. *Cell Rep* **11**, 618-629, doi:10.1016/j.celrep.2015.03.048 (2015).
86 Allemand, E. *et al.* A Broad Set of Chromatin Factors Influences Splicing. *PLoS Genet* **12**, e1006318, doi:10.1371/journal.pgen.1006318 (2016).
87 Convertini, P. *et al.* Sudemycin E influences alternative splicing and changes chromatin modifications. *Nucleic Acids Res* **42**, 4947-4961, doi:10.1093/nar/gku151 (2014).
88 Veloso, A. *et al.* Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res* **24**, 896-905, doi:10.1101/gr.171405.113 (2014).
89 Gobet, C. & Naef, F. Ribosome profiling and dynamic regulation of translation in mammals. *Curr Opin Genet Dev* **43**, 120-127, doi:10.1016/j.gde.2017.03.005 (2017).
90 Chi, B. *et al.* Interactome analyses revealed that the U1 snRNP machinery overlaps extensively with the RNAP II machinery and contains multiple ALS/SMA-causative proteins. *Sci Rep* **8**, 8755, doi:10.1038/s41598-018-27136-3 (2018).
91 Harlen, K. M. *et al.* Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue. *Cell Rep* **15**, 2147-2158, doi:10.1016/j.celrep.2016.05.010 (2016).
92 Dans, P. D. *et al.* Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. *Nucleic Acids Res* **42**, 11304-11320, doi:10.1093/nar/gku809 (2014).
93 Naughton, C. *et al.* Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nat Struct Mol Biol* **20**, 387-395, doi:10.1038/nsmb.2509 (2013).
94 Reymer, A., Zakrzewska, K. & Lavery, R. Sequence-dependent response of DNA to torsional stress: a potential biological regulation mechanism. *Nucleic Acids Res* **46**, 1684-1694, doi:10.1093/nar/gkx1270 (2018).
95 Vinogradov, A. E. & Anatskaya, O. V. DNA helix: the importance of being AT-rich. *Mamm Genome* **28**, 455-464, doi:10.1007/s00335-017-9713-8 (2017).
96 Shin, S. I. *et al.* Z-DNA-forming sites identified by ChIP-Seq are associated with actively transcribed regions in the human genome. *DNA Res*, doi:10.1093/dnares/dsw031 (2016).
97 Vinogradov, A. E. DNA helix: the importance of being GC-rich. *Nucleic Acids Res* **31**, 1838-1844 (2003).
98 Zamft, B., Bintu, L., Ishibashi, T. & Bustamante, C. Nascent RNA structure modulates the transcriptional dynamics of RNA polymerases. *Proc Natl Acad Sci U S A* **109**, 8948-8953, doi:10.1073/pnas.1205063109 (2012).
99 Shaul, O. How introns enhance gene expression. *Int J Biochem Cell Biol* **91**, 145-155, doi:10.1016/j.biocel.2017.06.016 (2017).
100 Pai, A. A. *et al.* The kinetics of pre-mRNA splicing in the Drosophila genome and the influence of gene architecture. *Elife* **6**, doi:10.7554/eLife.32537 (2017).
101 Smith, K. P., Moen, P. T., Wydner, K. L., Coleman, J. R. & Lawrence, J. B. Processing of endogenous pre-mRNAs in association with SC-35 domains is gene specific. *J Cell Biol* **144**, 617-629 (1999).
102 Moen, P. T., Jr., Smith, K. P. & Lawrence, J. B. Compartmentalization of specific pre-mRNA metabolism: an emerging view. *Hum Mol Genet* **4 Spec No**, 1779-1789 (1995).
103 Shopland, L. S., Johnson, C. V., Byron, M., McNeil, J. & Lawrence, J. B. Clustering of multiple specific genes and gene-rich R-bands around SC-35 domains: evidence for local euchromatic neighborhoods. *J Cell Biol* **162**, 981-990, doi:10.1083/jcb.200303131 (2003).
104 Sanchez-Alvarez, M., Sanchez-Hernandez, N. & Sune, C. Spatial Organization and Dynamics of Transcription Elongation and Pre-mRNA Processing in Live Cells. *Genet Res Int* **2011**, 626081, doi:10.4061/2011/626081 (2011).

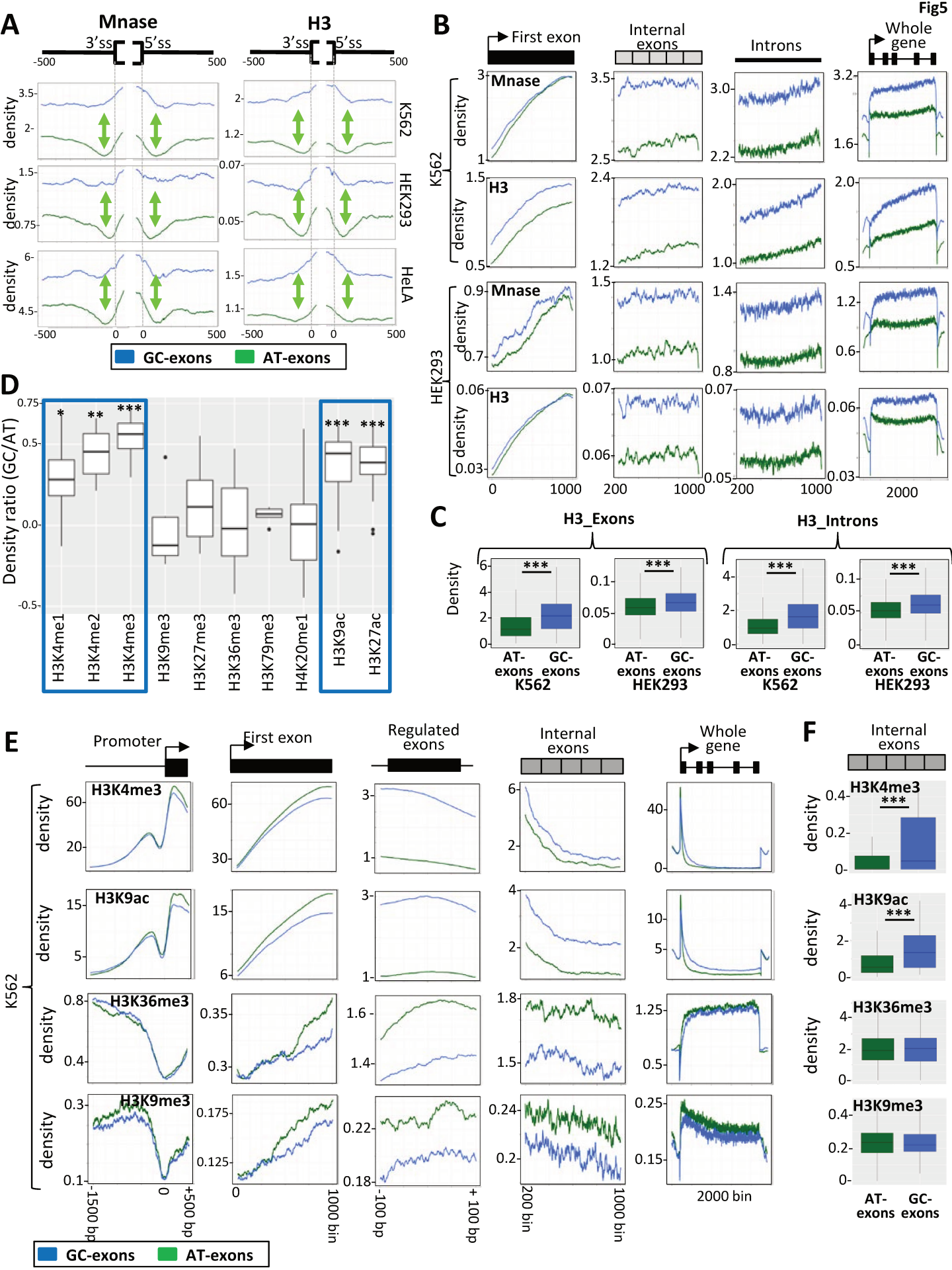
105 Zhu, J. *et al.* A novel role for minimal introns: routing mRNAs to the cytosol. *PLoS One* **5**, e10144, doi:10.1371/journal.pone.0010144 (2010).
106 Keane, P. A. & Seoighe, C. Intron Length Coevolution across Mammalian Genomes. *Mol Biol Evol* **33**, 2682-2691, doi:10.1093/molbev/msw151 (2016).
107 Bonnet, A. *et al.* Introns Protect Eukaryotic Genomes from Transcription-Associated Genetic Instability. *Mol Cell* **67**, 608-621 e606, doi:10.1016/j.molcel.2017.07.002 (2017).
108 Auboeuf, D. Alternative mRNA processing sites decrease genetic variability while increasing functional diversity. *Transcription* **9**, 75-87, doi:10.1080/21541264.2017.1373891 (2018).
109 Yeo, G. & Burge, C. B. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. Journal of Computational Biology 11, 377–394 (2004).
110 Lorenz, R. et al. ViennaRNA Package 2.0. Algorithms for Molecular Biology 6, (2011).
111 Corvelo, A., Hallegger, M., Smith, C. W. J. & Eyras, E. Genome-wide association between branch point properties and alternative splicing. PLoS Comput. Biol. 6, e1001016 (2010).
112 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
113 Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics 26, 2204–2207 (2010).

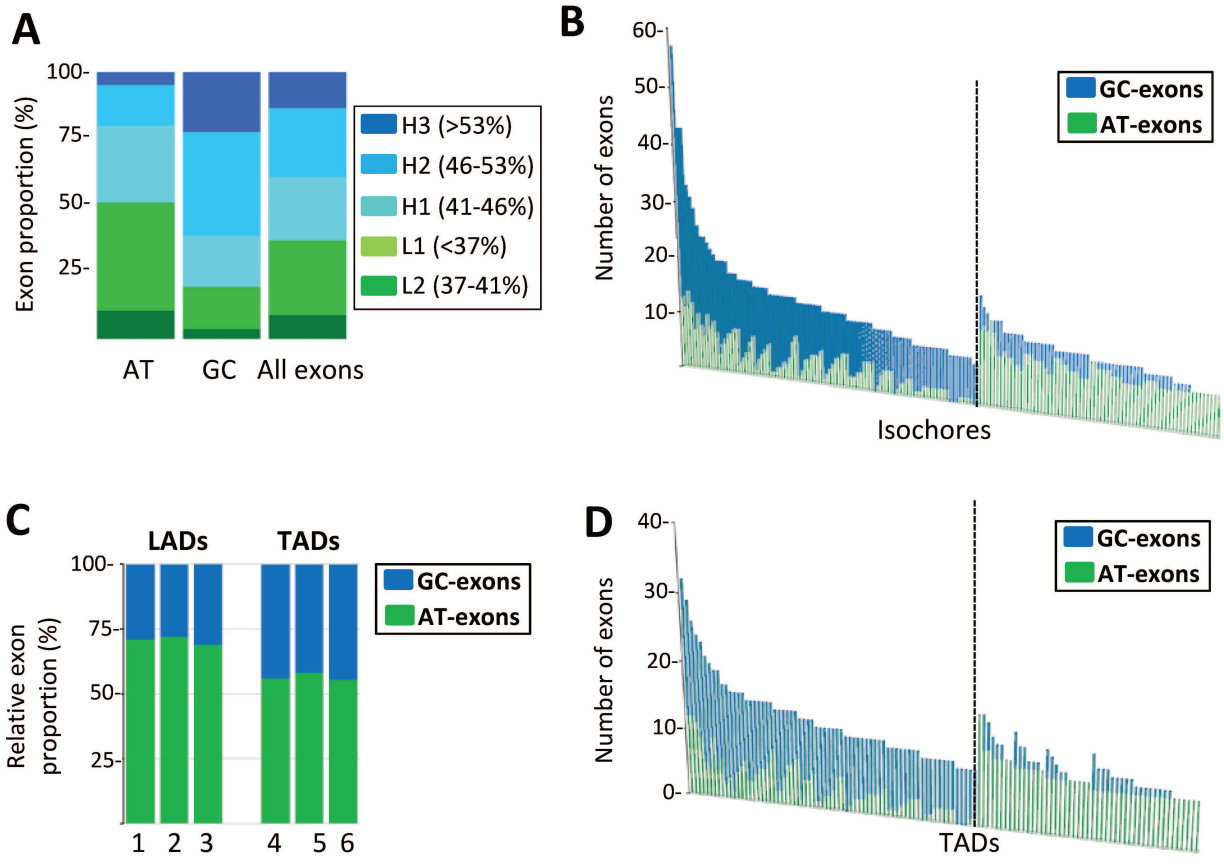

A
INTRON
EXONS
GC AT
(n-1)
(n)
GC AT GC AT
SRSF9
PCBP1
RBMX
hnRNPF
RBFOX2
hnRNPH1
SRFS6
SRSF5
SRSF2
RBM22
MBNL2
RBM25
SRSF3
SRSF1
hnRNPC
QKI
hnRNPA1
FUS
hnRNPL
PTBP1
hnRNPU
PCBP2
DAZAP1
hnRNPM
KHSRP
hnRNPK
MBNL1
SRSF7
hnRNPA2B1
RBM39
SFPQ
RBM15
TRA2A/B
-20 %
0 %
+ 20 %
B
Relative GC content (%) (Upstream introns)
Relative GC-content (%) (Exons)
Relative GC content (%) (downstream introns)
Relative GC-content (%) (Exons)
C
Smallest flanking intron size
-50 %
0 %
50 %
D
Relative GC content (%) (Exons)
Relative flanking intron size (%)
Relative GC content (%) (flanking introns)
Relative flanking intron size (%)
E
Relative exonic GC content (%)
Relative smallest flanking intron (%)
F
Exonic GC content (%)
GC-exons
AT-exons
All exons
Smallest flanking intron size

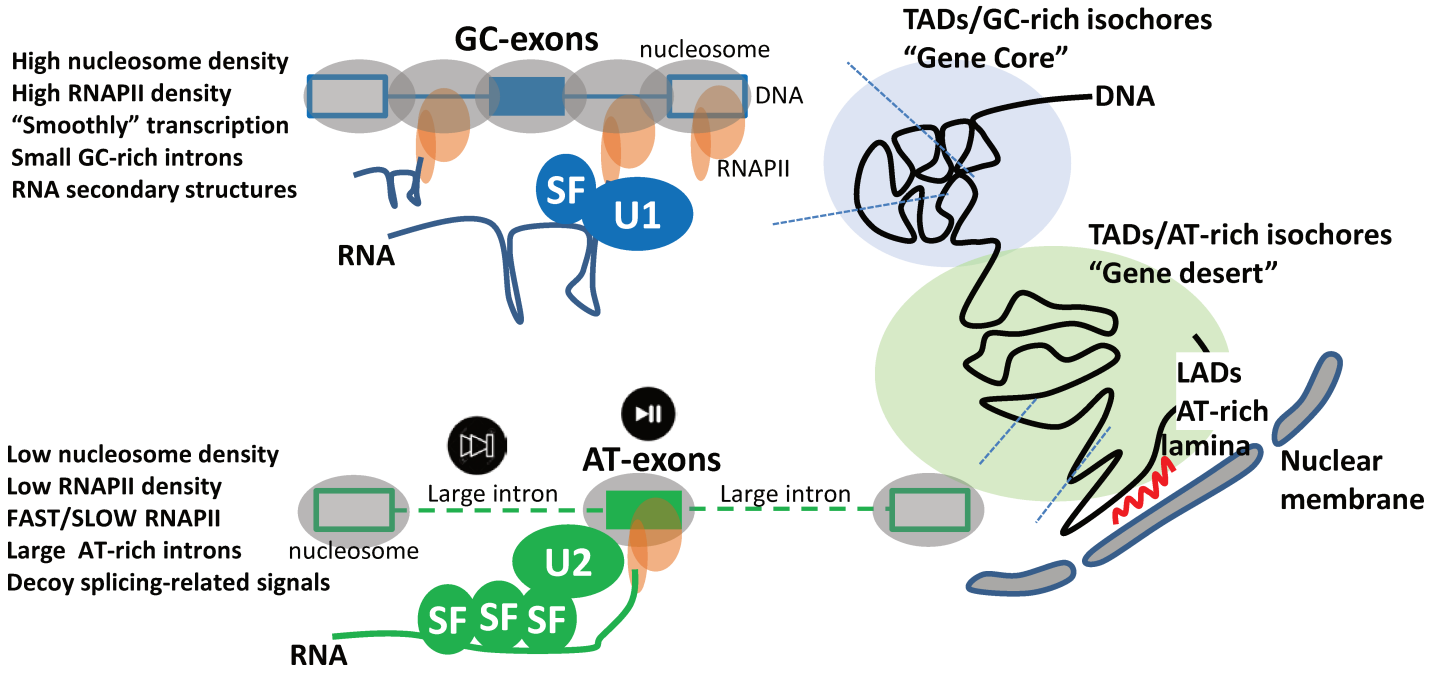

A
GC-exons
AT-exons
All exons
3'ss
exon
5'ss
GC density
A density
T density
C density
G density
Y density
-90
-30
0
+30
-30
0
+30
+90
B
GC-exons
AT-exons
A T G C
5'ss
+25
SRSF9
PCBP1
RBFOX2
SRSF6
SRSF1
MBNL2
RBMX
hnRNPH1
RBM22
RBM25
SRSF5
hnRNPF
RBM39
SRSF7
hnRNPA1
hnRNPU
TRA2A/B
DAZAP1
hnRNPK
hnRNPL
hnRNPA2B1
SFPQ
RBM15
hnRNPM
PCBP2
MBNL1
KHSRP
PTBP1
QKI
FUS
-20
0
20
C
GC-exons
AT-exons
A T G C
3'ss
-25
SRSF9
PCBP1
SRSF6
RBMX
hnRNPH1
RBFOX2
SRSF5
hnRNPF
MBNL2
RBM22
RBM25
SRSF1
RBM15
SFPQ
SRSF7
RBM39
MBNL1
hnRNPU
QKI
TRA2A/B
KHSRP
hnRNPL
hnRNPK
hnRNPA2B1
hnRNPM
PCBP2
PTBP1
DAZAP1
FUS
hnRNPA1
-20
0
20
D
5'ss
-25
+25
MFE (5' SS)
3'ss
MFE (3' SS)
***
GC-exons
AT-exons
All exons
E
BPs
-100
3'ss
Exon proportion (%)
≤ 2
> 2
Number of H bound (U2 snRNA_BP)
F
BPs of GC-exons
BPs or AT-exons
G
TNA
-50
3'ss
TTT
Py
-75
-35
H
U2AF2
Py
-200
HEK293T
HeLa
U2AF2 CLIP density
-200
-100
0
50


A
3'ss
exon
5'ss
GC density
A density
T density
DDX5/17
SNRNP70
SNRPC
All exons
U2AF1
U2AF2
SF1
SF3A3
SF3B4
B
MFE 5' ss
All exons
C
BPs
-100
3'ss
Exon proportion (%)
≤ 2
>2
D
TNA
-50
3'ss
Exon proportion (%)
E
TTT
Py
-75
-35
3'ss
F
U2AF2
Py
-200
HEK293T
HeLA
U2AF2 CLIP density
G
GC-exon vs. AT-exon
V-value


A
Relative GC content (%) (Gene)
Relative GC content (%) (Exon)
Relative gene size (%)
Relative GC content (%) (Gene)
B
GC-exons
AT-exons
All exons
% GC (Gene)
Log10(medan intron size) (Gene)
Gene Size
C
Promoter
First exon
Regulated exons
Internal exons
Introns
Whole gene
RNAPII
Density (K562)
Density (HepG2)
-1500 bp
+500 bp
1000 bin
-100 bp
+ 100 bp
200 bin
2000 bin
GC-exons
AT-exons
D
RNAPII_exons
AT-exons
GC-exons
E
pSER2
F
pSER2_exons


A
Mnase
H3
3'ss
5'ss
-500
500
density
K562
HEK293
HeLA
GC-exons
AT-exons
B
First exon
Internal exons
Introns
Whole gene
K562
HEK293
Mnase
H3
density
0
1000
200
2000
C
H3_Exons
H3_Introns
Density
***
AT-exons
GC-exons
K562
HEK293
D
Density ratio (GC/AT)
*
**
***
H3K4me1
H3K4me2
H3K4me3
H3K9me3
H3K27me3
H3K36me3
H3K79me3
H4K20me1
H3K9ac
H3K27ac
E
Promoter
First exon
Regulated exons
Internal exons
Whole gene
K562
H3K4me3
H3K9ac
H3K36me3
H3K9me3
density
-1500 bp
0
+500 bp
1000 bin
-100 bp
+100 bp
200 bin
2000 bin
GC-exons
AT-exons
F
Internal exons
H3K4me3
H3K9ac
H3K36me3
H3K9me3
density
***
AT-exons
GC-exons


A
100-
75-
50-
25-
Exon proportion (%)
H3 (>53%)
H2 (46-53%)
H1 (41-46%)
L1 (<37%)
L2 (37-41%)
AT
GC
All exons
B
60-
50-
40-
30-
20-
10-
Number of exons
GC-exons
AT-exons
Isochores
C
LADs
TADs
100-
75-
50-
25-
Relative exon proportion (%)
GC-exons
AT-exons
1 2 3 4 5 6
D
40-
30-
20-
10-
0-
Number of exons
GC-exons
AT-exons
TADs


High nucleosome density
High RNAPII density
“Smoothly” transcription
Small GC-rich introns
RNA secondary structures
GC-exons
nucleosome
DNA
RNAPII
SF
U1
RNA
TADs/GC-rich isochores
“Gene Core”
DNA
TADs/AT-rich isochores
“Gene desert”
LADs
AT-rich
lamina
Nuclear
membrane
Low nucleosome density
Low RNAPII density
FAST/SLOW RNAPII
Large AT-rich introns
Decoy splicing-related signals
AT-exons
Large intron
Large intron
nucleosome
U2
SF
SF
SF
RNA

Supplementary Table S1: GEO number of publicly available datasets analyzed in this work.

Supplementary Tables S2: List of exons activated by each analyzed splicing factors and of GC- and AT-exons

Figure S1

A. Violin plots representing the relative 3'ss score (upper panel) and the relative 5'ss score (lower panel) for each set of splicing-factor activated exons, when compared to all human exons. (*) Mann-Whitney test FDR < 0.05.

B. Violin plots representing the relative upstream intron size (upper panel) and the relative downstream intron size (lower panel) for each set of splicing-factor activated exons, when compared to all human exons. (*) Mann-Whitney test FDR < 0.05.

Figure S2

Violin plots representing the relative adenine, cytosine, guanine and thymine frequencies for each set of splicing-factor activated exons, when compared to all human exons. (*) Mann-Whitney test FDR < 0.05.

Figure S3

A. Violin plots representing the relative GC frequencies in each set of activated exons (upper panel) and their relative GC frequencies in the 100 last nucleotides of their upstream introns (middle panel) and in the first 100 nucleotides of their downstream introns(lower panel), when compared to all human exons .

Figure S4

A. Violin plots representing the 5'ss score (left panel), the 3'ss score (middle panel) and the pyrimidine frequencies in the 25 last nucleotides of the upstream intron (right panel) of GC- and AT-exons. (*) Mann-Whitney test P < 0.05.

B. Pyrimidine frequency (%) maps in exons (and their flanking sequences) activated by different splicing factors.

C. Density of reads obtained after immunoprecipitation of RNAPII in HelA, MCF-7 and HEK293 cell lines. Different parts of the genes hosting GC- and AT-exons are represented.

D. Density of reads mapping across different parts of the genes hosting GC- and AT-exons and obtained after DNA treatment with MNase or after immunoprecipitation of the histone H3 in Hela.

E. Density of reads mapping across promoters of the genes hosting GC- and AT-exons and obtained after DNA treatment with MNase (left panels) or after immunoprecipitation of the histone H3 (right panels) in K562, HEK293T, and Hela cell lines.

Supplementary Figure S5

Density of reads mapping across different parts of genes hosting GC- and AT-exons and obtained from the K562 cell line after immunoprecipitation of DNA using antibodies against different histone modifications, as indicated.

Supplementary Figure S6

A. Density of reads mapping across GC- and AT-exons and obtained from K562, HEK293, HeLA, HepG2 and MCF-7 cell lines after immunoprecipitation of DNA using antibodies against different histone modifications, as indicated.

B. Density of reads mapping across the genes hosting GC- and AT-exons and obtained from K562, HEK293, HeLA, HepG2 and MCF-7 cell lines after immunoprecipitation of DNA using antibodies against different histone modifications, as indicated.

Supplementary Figure S7

A. Proportion of AT-, GC-, and all human exons distributed across different isochores defined by Constantini et al (ref, left panel) or by IsoSegmenter (right panel).

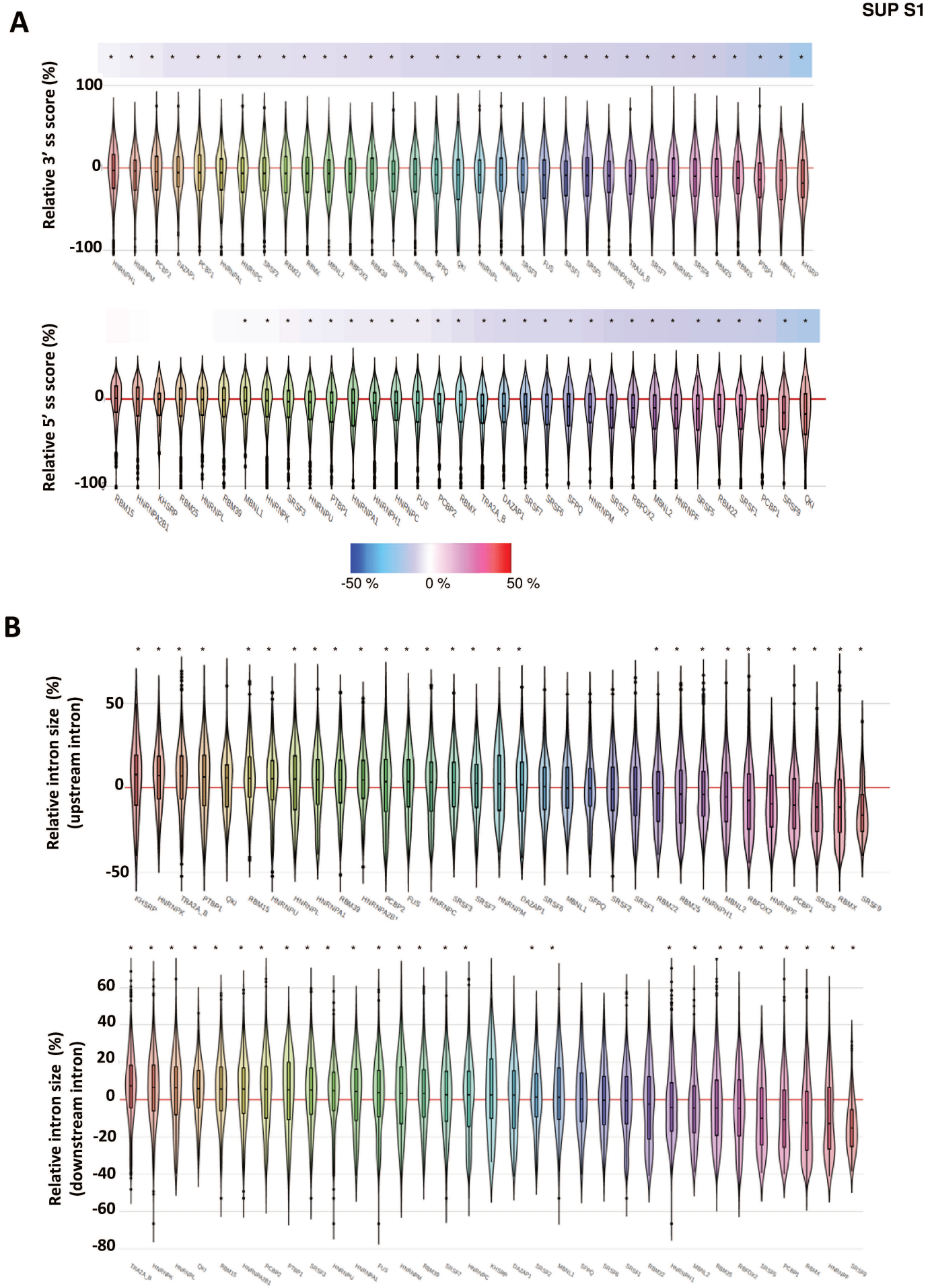
B. Number of AT- and GC-exons present in individual TADs defined in MCF-7 (left panel) or IMR90 (right panel) cell lines.

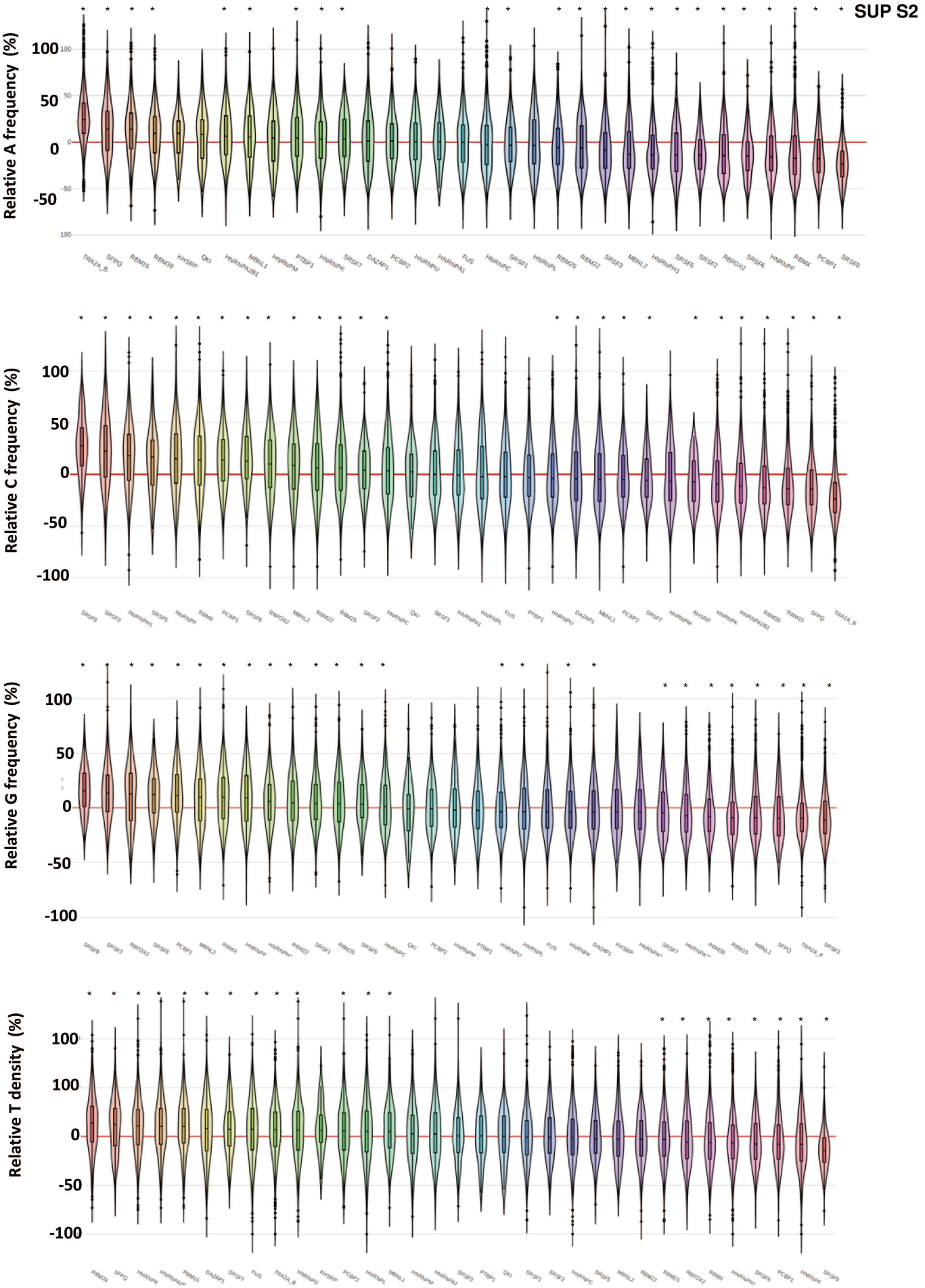
Supplementary Figure S8

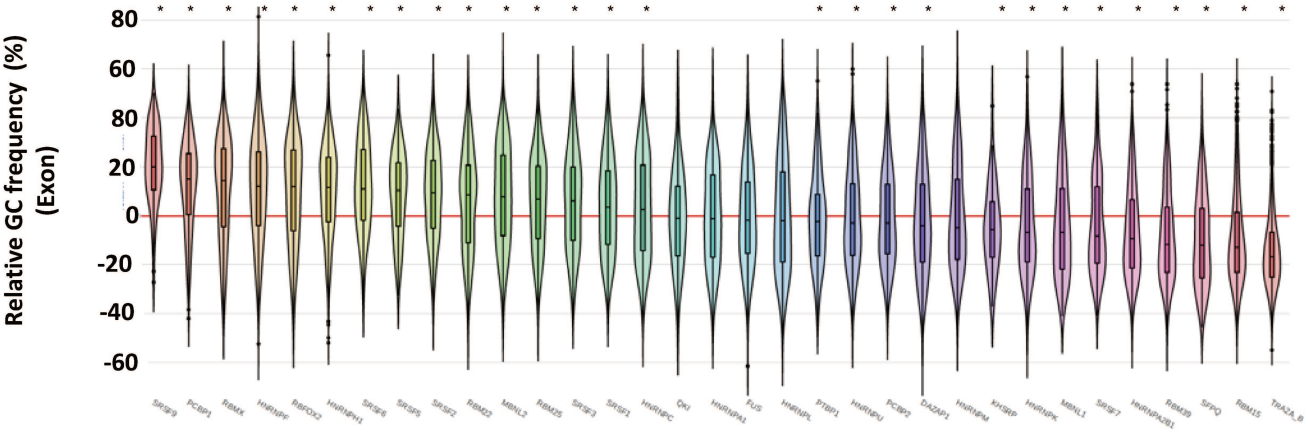
Splicing factor binding motifs retrieved from different resources.

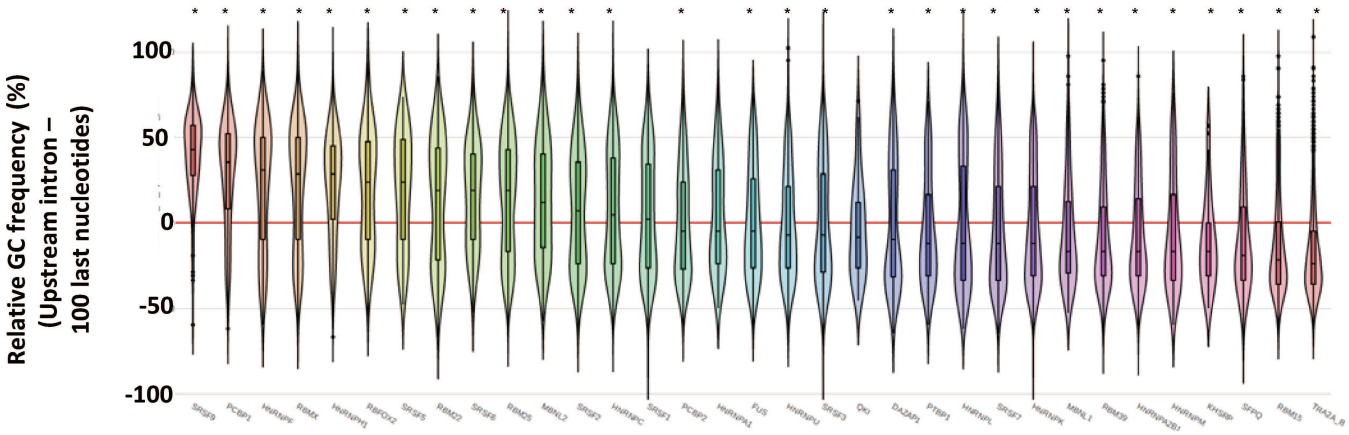
Supplementary Figure S9

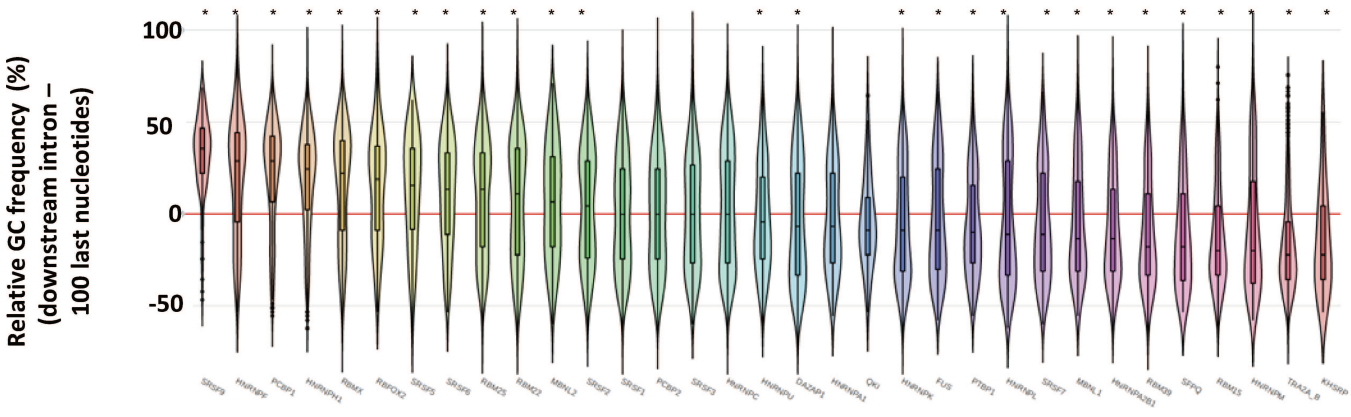
Interaction network between splicing factors and spliceosome-associated components generated. Splicing factors in blue activate GC-exons and the splicing factors in green activate AT-exons.

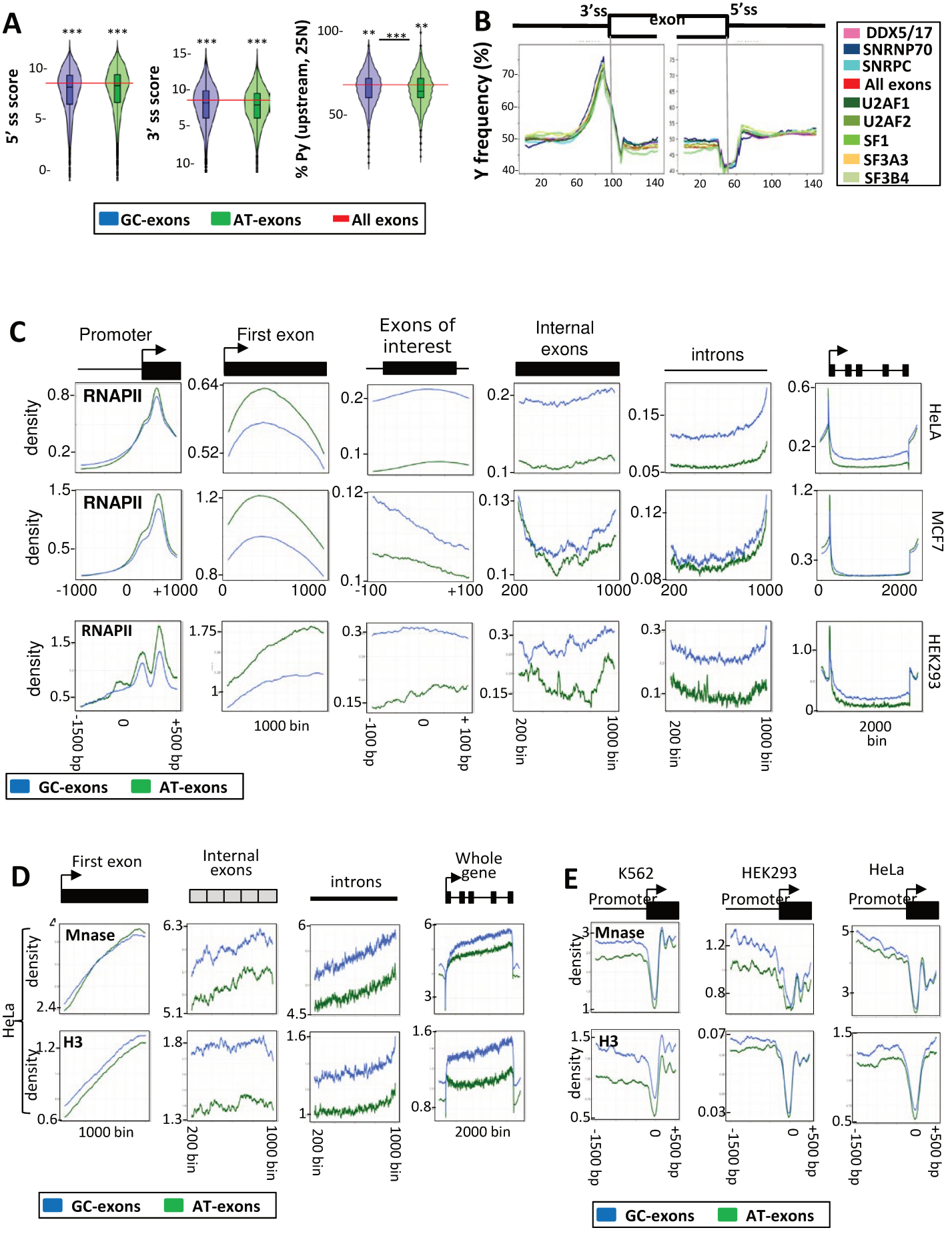

A
Relative 3' ss score (%)
100
0
-100
HNRNPH1 HNRNPM PCBP2 DAZAP1 PCBP1 HNRNPA1 HNRNPC SRSF2 RBM22 RBMX MBNL2 RBFOX2 RBM39 SRSF9 HNRNPK SFPQ QKI HNRNPL HNRNPU SRSF3 FUS SRSF1 SRSF5 HNRNPA2B1 TRA2A_B SRSF7 HNRNPF SRSF6 RBM25 RBM15 PTBP1 MBNL1 KHSRP
Relative 5' ss score (%)
0
-100
RBM15 HNRNPA2B1 KHSRP RBM25 HNRNPL RBM39 MBNL1 HNRNPK SRSF3 HNRNPU PTBP1 HNRNPA1 HNRNPH1 HNRNPC FUS PCBP2 RBMX TRA2A_B DAZAP1 SRSF7 SRSF6 SFPQ HNRNPM SRSF2 RBFOX2 MBNL2 HNRNPF SRSF5 RBM22 SRSF1 PCBP1 SRSF9 QKI
-50 %
0 %
50 %
B
Relative intron size (%) (upstream intron)
50
0
-50
KHSRP HNRNPK TRA2A_B PTBP1 QKI RBM15 HNRNPU HNRNPL HNRNPA1 RBM39 HNRNPA2B1 PCBP2 FUS HNRNPC SRSF3 SRSF7 HNRNPM DAZAP1 SRSF6 MBNL1 SFPQ SRSF2 SRSF1 RBM22 RBM25 HNRNPH1 MBNL2 RBFOX2 HNRNPF PCBP1 SRSF5 RBMX SRSF9
Relative intron size (%) (downstream intron)
60
40
20
0
-20
-40
-60
-80
TRA2A_B HNRNPK HNRNPL QKI RBM15 HNRNPA2B1 PCBP2 PTBP1 SRSF3 HNRNPU HNRNPA1 FUS HNRNPM RBM39 SRSF7 HNRNPC KHSRP DAZAP1 SRSF2 MBNL1 SFPQ SRSF6 SRSF1 RBM22 HNRNPH1 MBNL2 RBM25 RBFOX2 SRSF5 PCBP1 RBMX HNRNPF SRSF9

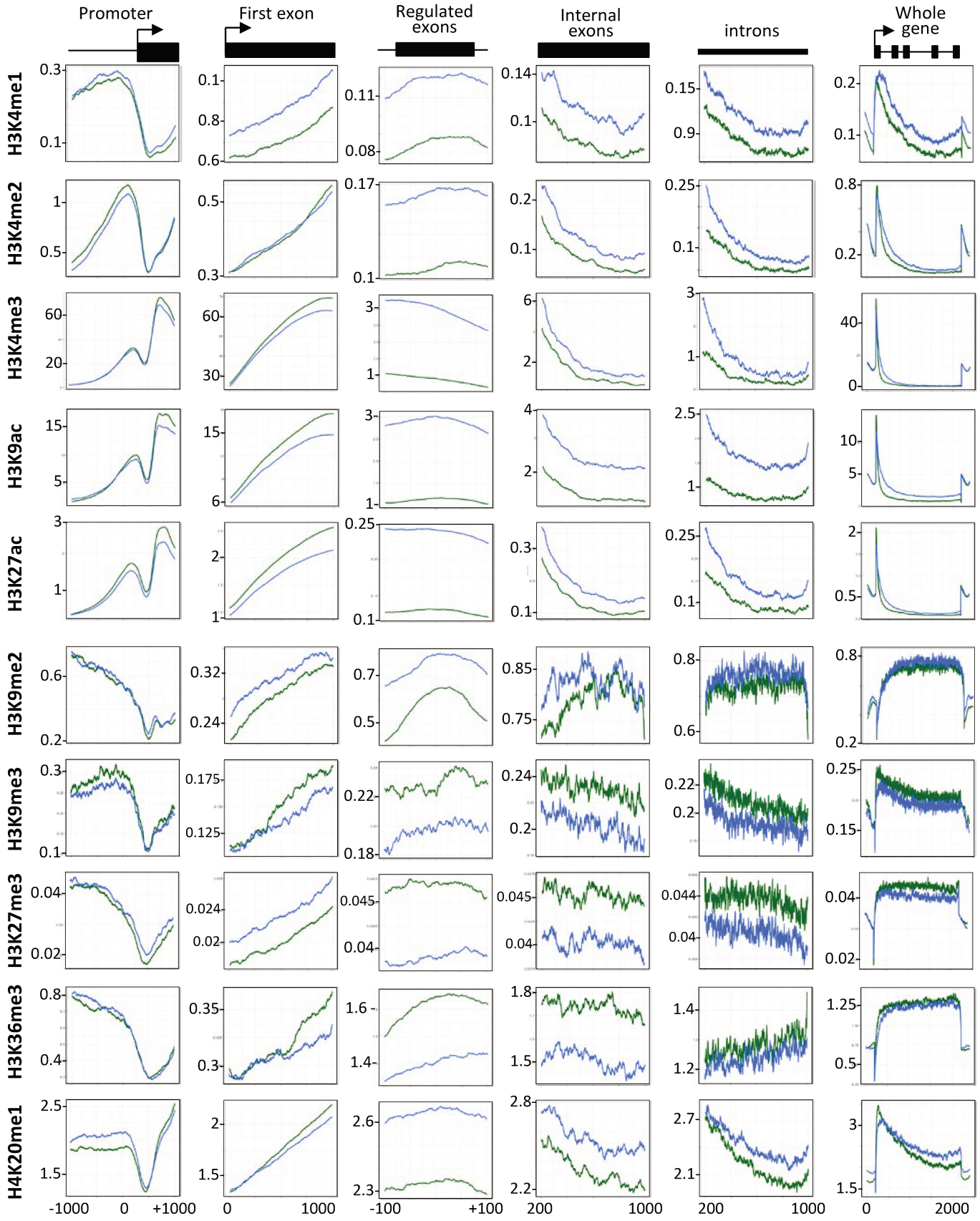

Relative A frequency (%)
Relative C frequency (%)
Relative G frequency (%)
Relative T density (%)

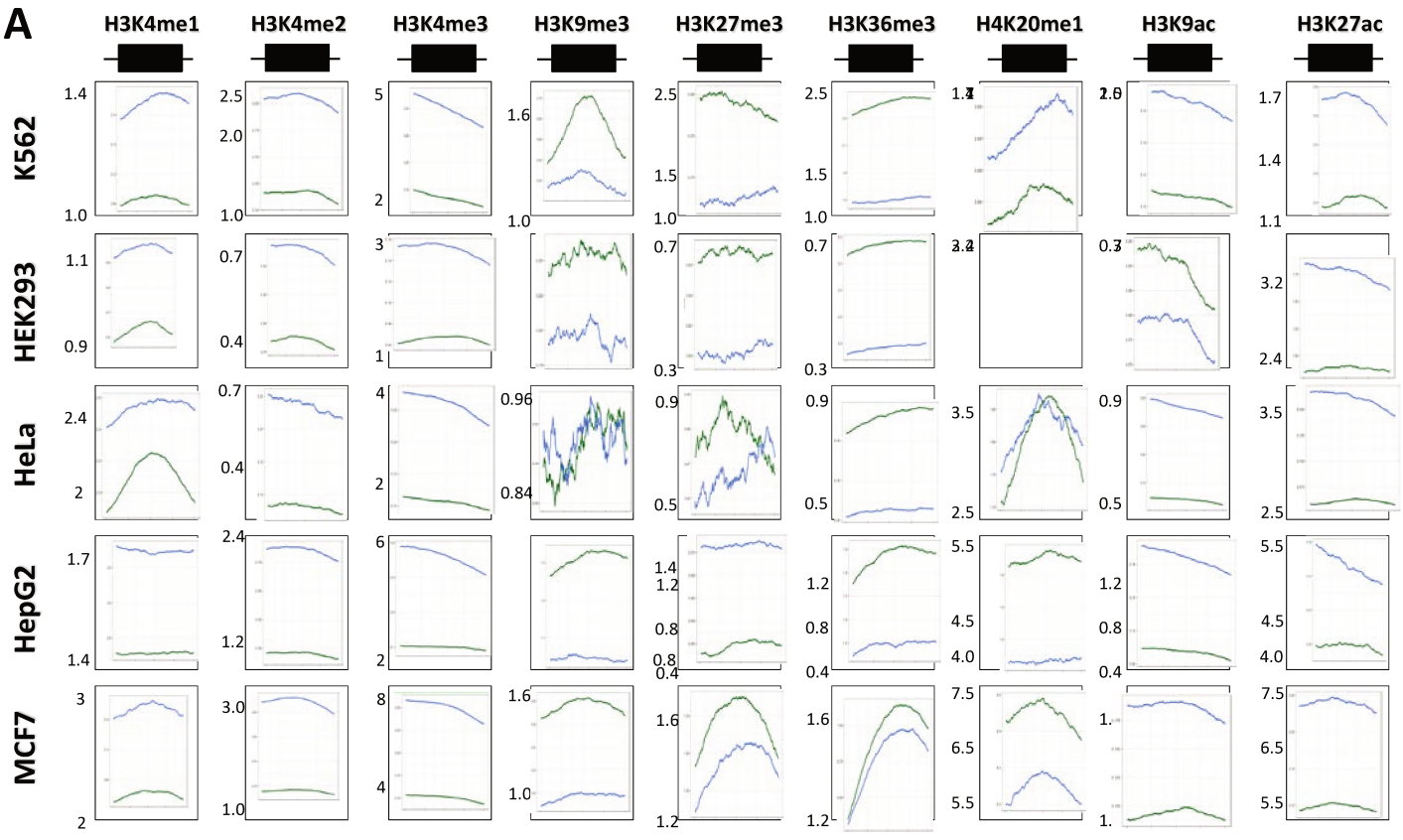

Relative GC frequency (%)
(Exon)
80
60
80
20
0
-20
-40
-60

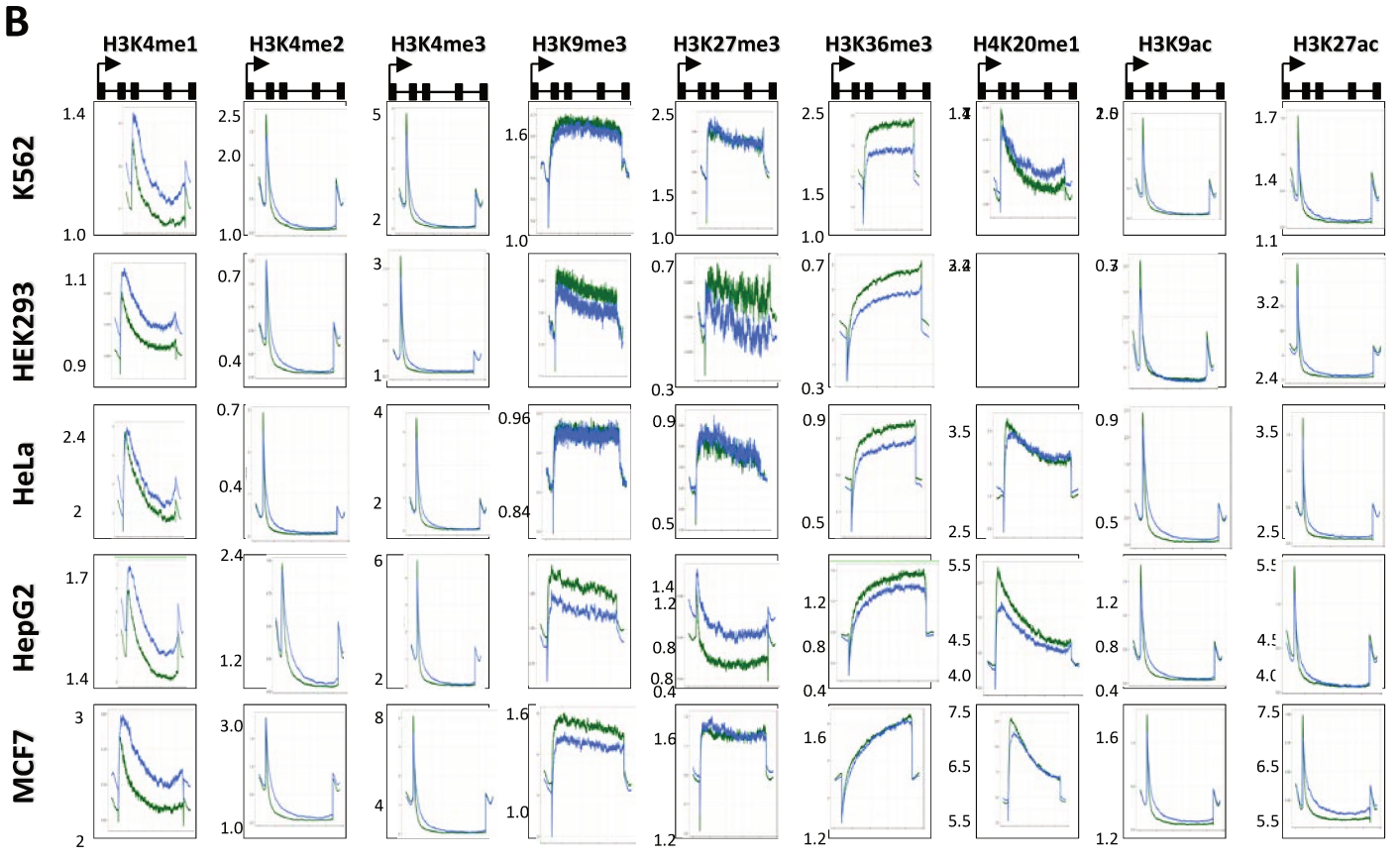

Relative GC frequency (%)
(Upstream intron –
100 last nucleotides)
100
50
0
-50
-100

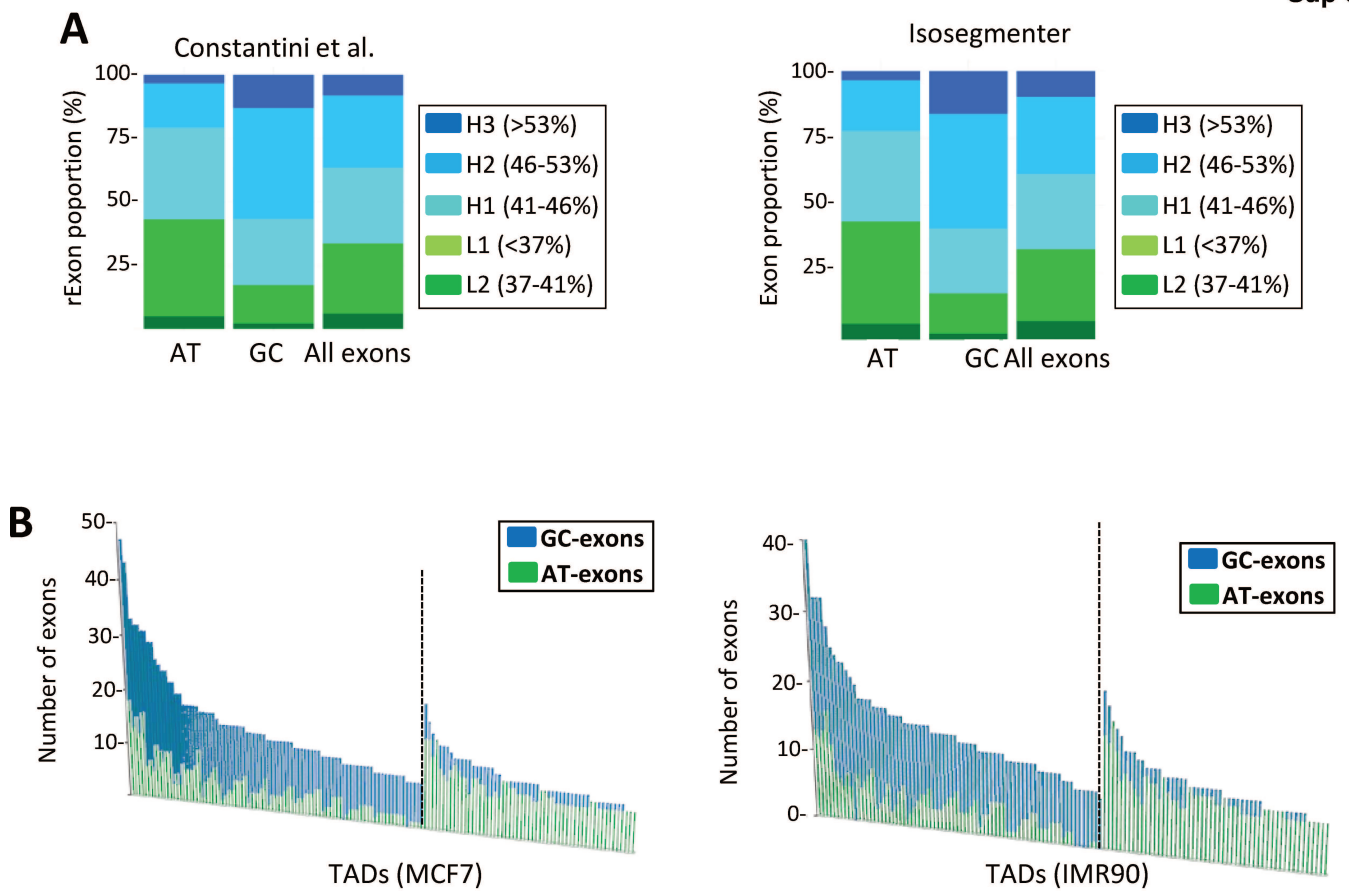

Relative GC frequency (%)
(downstream intron –
100 last nucleotides)
100
50
0
-50

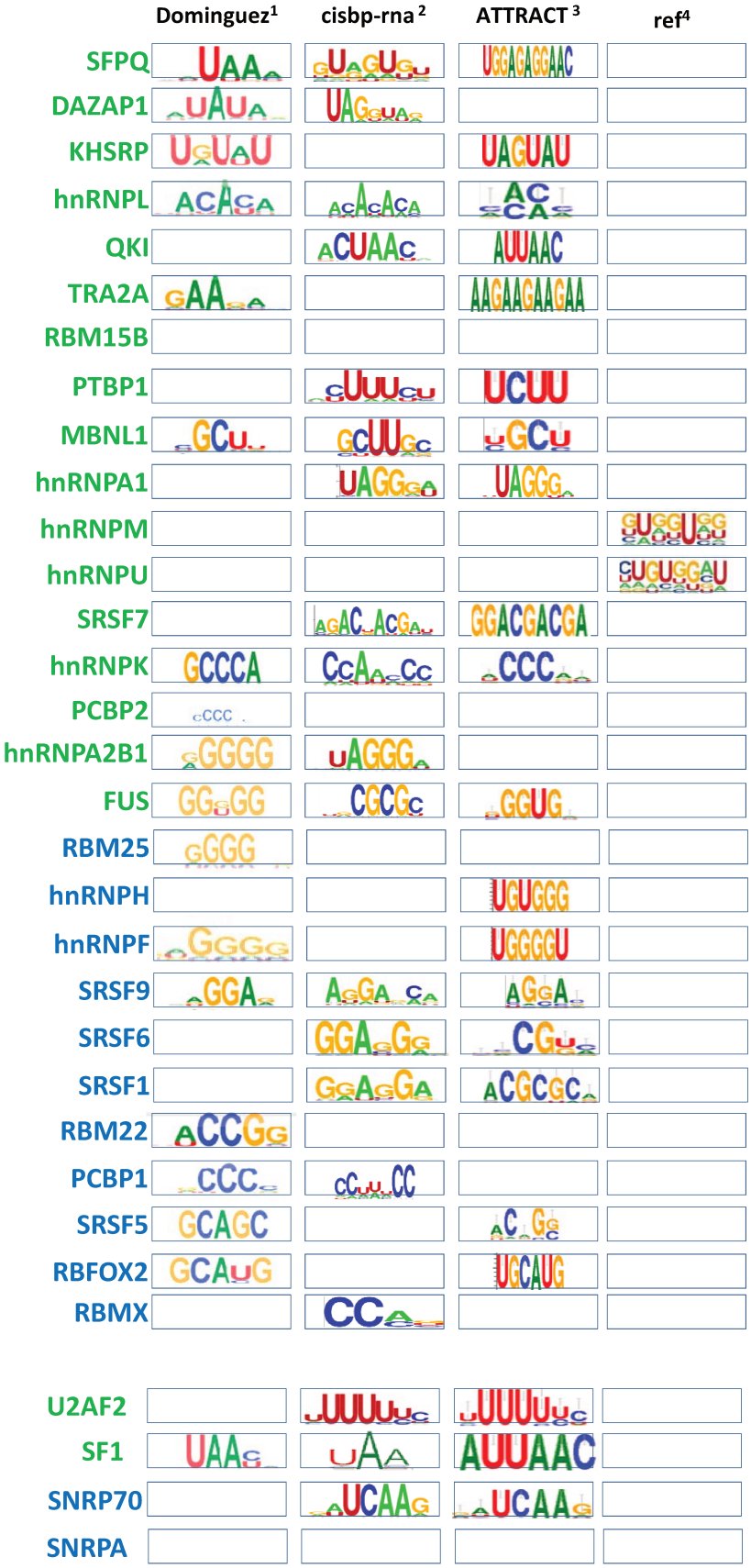

A
5' ss score
3' ss score
% Py (upstream, 25N)
GC-exons
AT-exons
All exons
B
3'ss
exon
5'ss
Y frequency (%)
DDX5/17
SNRNP70
SNRPC
All exons
U2AF1
U2AF2
SF1
SF3A3
SF3B4
C
Promoter
First exon
Exons of interest
Internal exons
introns
RNAPII
density
HeLA
MCF7
HEK293
-1500 bp
0
+500 bp
1000 bin
-100 bp
+100 bp
200 bin
2000 bin
GC-exons
AT-exons
D
First exon
Internal exons
introns
Whole gene
HeLa
Mnase
H3
density
1000 bin
200 bin
2000 bin
GC-exons
AT-exons
E
K562
HEK293
HeLa
Promoter
Mnase
H3
density
-1500 bp
0
+500 bp
GC-exons
AT-exons


Promoter
First exon
Regulated exons
Internal exons
introns
Whole gene
H3K4me1
H3K4me2
H3K4me3
H3K9ac
H3K27ac
H3K9me2
H3K9me3
H3K27me3
H3K36me3
H4K20me1
-1000
0
+1000
0
1000
-100
+100
200
1000
200
1000
0
2000

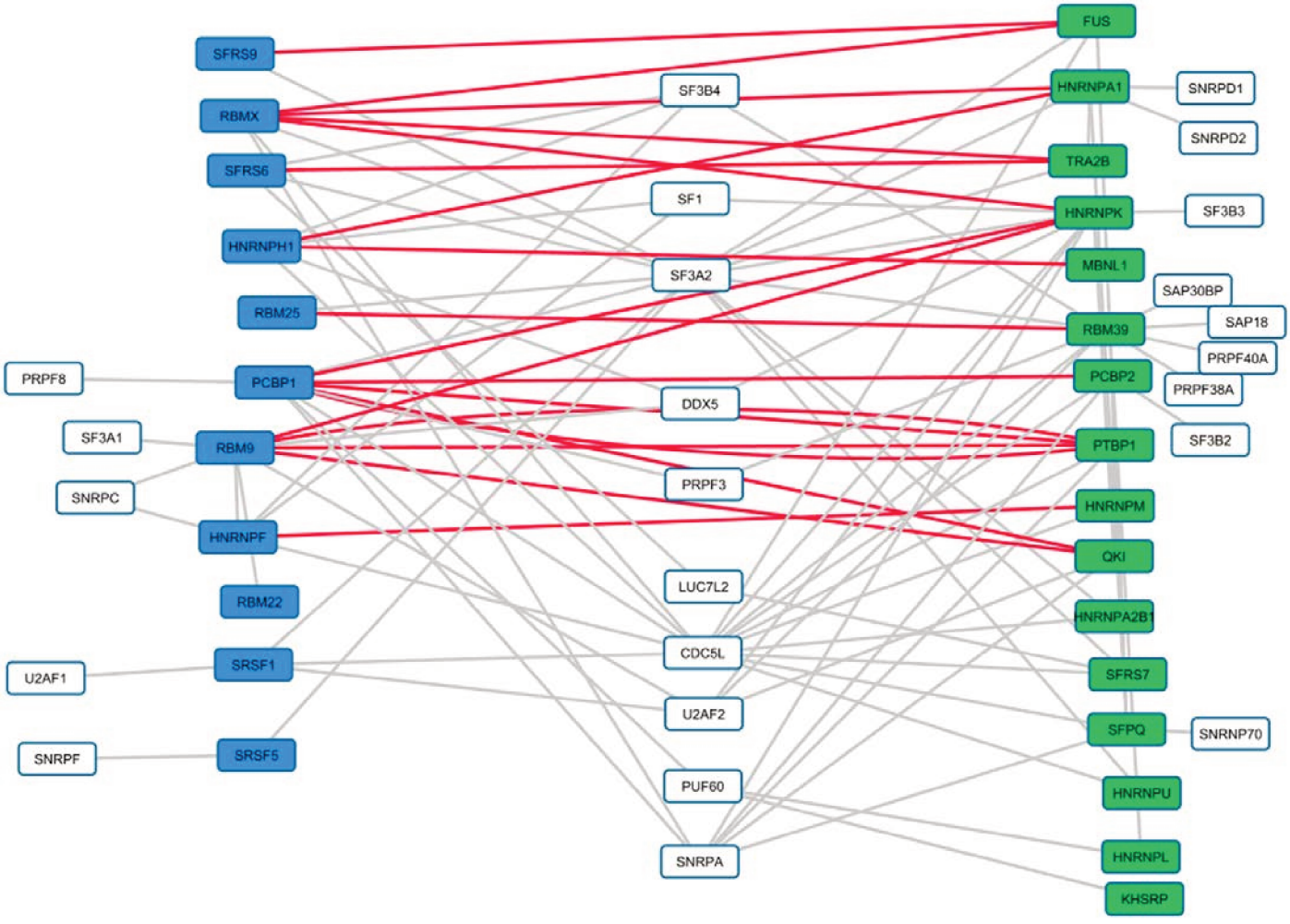

A
H3K4me1
H3K4me2
H3K4me3
H3K9me3
H3K27me3
H3K36me3
H4K20me1
H3K9ac
H3K27ac
K562
HEK293
HeLa
HepG2
MCF7


B
H3K4me1
H3K4me2
H3K4me3
H3K9me3
H3K27me3
H3K36me3
H4K20me1
H3K9ac
H3K27ac
K562
HEK293
HeLa
HepG2
MCF7


A
Constantini et al.
rExon poportion (%)
100-
75-
50-
25-
AT
GC
All exons
H3 (>53%)
H2 (46-53%)
H1 (41-46%)
L1 (<37%)
L2 (37-41%)
Isosegmenter
Exon proportion (%)
100-
75-
50-
25-
AT
GC
All exons
H3 (>53%)
H2 (46-53%)
H1 (41-46%)
L1 (<37%)
L2 (37-41%)
B
Number of exons
50-
40-
30-
20-
10-
GC-exons
AT-exons
TADs (MCF7)
Number of exons
40-
30-
20-
10-
0-
GC-exons
AT-exons
TADs (IMR90)



[1] Dominguez et al. PMID 29883606
[2] http://attract.cnic.es
[3] http://cisbp-rna.ccbr.utoronto.ca/index.php
4 ref


FUS
SFRS9
SF3B4
HNRNPA1
SNRPD1
RBMX
SNRPD2
SFRS6
TRA2B
SF1
HNRNPK
SF3B3
HNRNPH1
MBNL1
SF3A2
SAP30BP
RBM25
RBM39
SAP18
PRPF40A
PRPF8
PCBP1
PCBP2
PRPF38A
DDX5
SF3A1
RBM9
PTBP1
SF3B2
PRPF3
SNRPC
HNRNPM
HNRNPF
QKI
LUC7L2
RBM22
HNRNPA2B1
CDC5L
U2AF1
SRSF1
SFRS7
U2AF2
SFPQ
SNRNP70
SNRPF
SRSF5
PUF60
HNRNPU
SNRPA
HNRNPL
KHSRP

Conclusion de l’Article # 1

L’originalité du travail présenté ci-dessus repose sur plusieurs observations. Tout d’abord, nous montrons que des exons co-régulés par un même facteur d’épissage ont une composition nucléotidique similaire. Cela pourrait s’expliquer par le fait que chaque facteur d’épissage a une préférence de fixation pour des séquences ayant une composition nucléotidique particulière. Ensuite, nous montrons que des biais de composition nucléotidique génèrent des contraintes sur l’organisation de la chromatine (par exemple, le positionnement des nucléosomes) et sur l’épissage (par exemple, en favorisant la formation de structures secondaires). Enfin, l’originalité de ce travail repose sur l’observation que les exons ont les mêmes biais de composition nucléotidique que les gènes et isochores auxquels ils appartiennent. Or le biais de composition nucléotidique de portions d’ADN joue un rôle dans l’organisation 3D de la chromatine. Par conséquent, le biais de composition nucléotidique relie de façon directe, l’organisation de la chromatine sur de grandes régions d’ADN à des évènements moléculaires, comme l’épissage, se déroulant à l’échelle de l’exon. Je remettrai cette observation dans un contexte plus général dans la Discussion de cette thèse.

Article # 2

“Interplay between coding and exonic splicing regulatory sequences”

Introduction

L'épissage alternatif augmente la diversité du transcriptome et du protéome codé par un nombre limité de gènes. Deux questions connexes importantes concernant le processus d'épissage ont été peu abordées. La première question concerne l'information biologique codée par des exons co-régulés (par exemple des exons régulés par le même facteur d'épissage). Jusqu'à présent, il est principalement supposé que les exons régulés par un même facteur d'épissage sont divers et sans rapport en termes de caractéristiques des peptides codés par les exons. La deuxième question concerne la relation entre les séquences codantes et les séquences régulatrices d'épissage. Actuellement il est supposé que les séquences codantes peuvent s’accommoder de séquences régulatrices d'épissage (c.-à-d., sans affecter la nature des acides aminés codés) en raison de la redondance du code génétique. En effet, l’hypothèse couramment admise est que les sites de fixation des facteurs d’épissage au sein des exons influencent uniquement la troisième base des codons.

L’objectif du travail présenté ci-dessous a été de challenger ces hypothèses de travail en analysant la composition nucléotidique d’exons co-régulés et la composition en acides aminés des peptides codés par ces exons.

Interplay between coding and exonic splicing regulatory sequences

Nicolas Fontrodona[1,$], Fabien Aubé[1,$], Jean-Baptiste Claude[1], Hélène Polvèche[1,§], Sébastien Lemaire[1], Léon-Charles Tranchevent[2], Laurent Modolo[3], Franck Mortreux[1], Cyril F. Bourgeois[1], Didier Auboeuf[1,*]

[1.] Univ Lyon, ENS de Lyon, Univ Claude Bernard, CNRS UMR 5239, INSERM U1210, Laboratory of Biology and Modelling of the Cell, 46 Allée d'Italie Site Jacques Monod, F-69007, Lyon, France

[2.] Preoteome and Genome Research Unit, Department of Oncology, Luxembourg Institute of Health (LIH), Luxembourg, Luxembourg.

[3.] LBMC biocomputing center, CNRS UMR 5239, INSERM U1210, 46 Allée d'Italie Site Jacques Monod, F-69007, Lyon, France

[§.] Present address: INSERM UMR 861, I-STEM, 28, Rue Henri Desbruères, 91100 Corbeil-Essonnes – France

[$] Equal contribution

*Corresponding author: Didier Auboeuf, Laboratory of Biology and Modelling of the Cell, ENS de Lyon, 69007 Lyon, France. Didier.auboeuf@inserm.fr

Abstract

The inclusion of exons during the splicing process depends on the binding of splicing factors to short low-complexity regulatory sequences. The relationship between exonic splicing regulatory sequences and coding sequences is still poorly understood. We demonstrate that exons that are coregulated by any splicing factor share a similar nucleotide composition bias and preferentially code for amino acids with similar physicochemical properties because of the non-randomness properties of the genetic code. Indeed, amino acids sharing similar physicochemical properties correspond to codons that have the same nucleotide composition bias. In addition, coregulated exons encoding amino acids with similar physicochemical properties correspond to specific protein features. In conclusion, the splicing regulation of an exon by a splicing factor that relies on the affinity of this factor for specific nucleotide(s) is tightly interconnected with the encoded physicochemical properties because of the non-randomness of the genetic code. We therefore uncover an unanticipated bidirectional interplay between the splicing regulatory process and its biological functional outcome.

Introduction

Alternative splicing is a cellular process involved in the regulated inclusion or exclusion of exons during the processing of mRNA precursors. Alternative splicing is the rule in human since 95% of human genes produce several splicing variants (Pan et al. 2008; Wang et al. 2008). The exon selection process relies on RNA binding splicing factors that enhance or repress exon inclusion following two main principles. First, splicing factors bind to short intronic or exonic motifs (or splicing regulatory sequences) that are often low-complexity sequences composed of a repetition of the same nucleotide or dinucleotide (Fu and Ares 2014). The interaction of splicing factors with their cognate binding motifs often depends on the sequence context and on the presence of clusters of related-binding motifs (Zhang et al. 2013; Cereda et al. 2014; Fu and Ares 2014; Dominguez et al. 2018; Jobbins et al. 2018). The second principle states that the splicing decision (i.e., exon inclusion or skipping) depends on where splicing factors bind on pre-mRNAs with respect to the regulated exons. For example, hnRNP-like splicing factors often repress the inclusion of exons they bind to, but they enhance exon inclusion when they do not bind to exons but instead to their flanking introns (Erkelenz et al. 2013; Fu and Ares 2014; Geuens et al. 2016). Meanwhile, exonic binding of SR-like splicing factors (or SRSFs) usually enhances exon inclusion (Erkelenz et al. 2013; Fu and Ares 2014).

Since some splicing regulatory sequences lie within protein-coding sequences, a major challenge is to understand how coding sequences accommodate this overlapping layer of information (Itzkovitz et al. 2010; Lin et al. 2011; Savisaar and Hurst 2017b; Savisaar and Hurst 2017a). To date, a general assumption is that protein coding regions can accommodate overlapping information or “codes” (including the “splicing code”) as a direct consequence of the redundancy of the genetic code that allows the same amino acid to be encoded by several codons differing only on their third "wobble" nucleotide (Goren et al. 2006; Itzkovitz and Alon 2007; Itzkovitz et al. 2010; Lin et al. 2011; Shabalina et al. 2013; Savisaar and Hurst 2017b; Savisaar and Hurst 2017a). Therefore, coding and exonic splicing regulatory sequences could evolve independently because of the variation of the third nucleotide of codons. However, it has been shown that some amino acids are preferentially encoded near exon-intron junctions because of the presence of general splicing consensus sequences near splicing sites (Parmley et al. 2007; Warnecke et al. 2008; Smithers et al. 2015). In addition, recent evidence has suggested that exons that are coregulated in specific pathophysiological conditions may code for protein domains engaged in similar cellular processes (Irimia et al. 2014; Tranchevent et al. 2017). These observations raised the possibility that exons regulated by the same splicing regulatory process code for similar biological information. So far, the lack of large sets of coregulated exons limited studies addressing the interplay between the splicing regulatory process and peptide sequences encoded by splicing regulated exons. By focusing on exons

coregulated by different splicing factors, we uncover a bidirectional interplay between the physicochemical protein features encoded by exons and their regulation by splicing factors.

Results

Nucleotide composition bias of coregulated exons

To investigate the relationship between exonic splicing regulatory sequences and coding sequences, we analyzed RNA-seq datasets generated from different cell lines expressing or not Arg/Ser (RS) domain-containing splicing factors (SRSFs), focusing on SRSF1, SRSF2, SRSF3, and TRA2 (Supplementary Table S1). Each splicing factor regulated a common set of exons in several cell lines but many exons were regulated on a cell line-specific mode (Supplementary Figure S1). It has been shown that SRSF1, SRSF2, SRSF3, and TRA2 bind to GGA-rich motifs, SSNG motifs (where S=G or C), C-rich and G-poor motifs, and AGAA-like motifs, respectively (Grellscheid et al. 2011; Tsuda et al. 2011; Anko et al. 2012; Pandit et al. 2013; Ray et al. 2013; Best et al. 2014; Fu and Ares 2014; Anczukow et al. 2015; Hauer et al. 2015; Giudice et al. 2016; Luo et al. 2017). As expected, hexanucleotides enriched in SRSF1-, SRSF2-, SRSF3-, or TRA2-activated exons across different cell lines were enriched in purine-rich, S-rich, C-rich, or A-rich hexanucleotides, respectively, when compared to control exons and in contrast to exons repressed by the same factor (Figure 1A and Supplementary Table S2).

Each set of factor-specific exons had a specific nucleotide composition bias. Indeed, SRSF1-activated exons were enriched in G when compared to control exons and in contrast to SRSF1-repressed exons (Figure 1B, upper panel and Supplementary Table S2). SRSF2-activated exons were enriched in S (G or C) in contrast to SRSF2-repressed exons (Figure 1B, upper panel). SRSF3-activated exons were enriched in C and impoverished in G (Figure 1B, upper panel and Supplementary Figure S1). Finally, TRA2-activated exons were enriched in A in contrast to TRA2-repressed exons (Figure 1B, upper panel). Accordingly, a larger proportion of SRSF1-, SRSF2-, SRSF3-, or TRA2-activated exons had a high frequency of G, S, C, or A nucleotides, respectively, when compared to the corresponding repressed exons (Figure 1B, lower panels). While the enriched nucleotides within splicing factor-regulated exons could be randomly distributed across exons, we observed an increased frequency of specific dinucleotides and low-complexity sequences. For example, GG, SS, CC, or AA dinucleotides were more frequent in SRSF1-, SRSF2-, SRSF3-, or TRA2-activated exons, respectively, than in control exons or in the corresponding repressed exons (Figure 1C, upper panels and Supplementary Table S2). A larger proportion of SRSF1-, SRSF2-, SRSF3-, or TRA2-activated exons contained G-, S-, C-, or A-rich low-complexity sequences, respectively, when compared to the corresponding repressed exons (Figure 1C, lower panels).

We next analyzed the codon content of exons regulated by these splicing factors. In agreement with the nucleotide composition bias described above, SRSF1-, SRSF2-, SRSF3-, or TRA2-

activated exons were enriched in G-, S-, C-, or A-rich codons, respectively, compared to both sets of control exons or the corresponding repressed exons (Figure 1D, upper panels and Supplementary Figure S2 and Table S2). Importantly, the nucleotide composition bias was observed at the first and second codon positions (Figure 1D, lower panels and Supplementary Table S2), raising the possibility that different sets of SRSF-regulated exons may preferentially code for different amino acids.

SRSF-coregulated exons code for amino acids with similar physicochemical properties

As shown in Figure 2A (upper panels), amino acids more frequently encoded by SRSF1-, SRSF2-, SRSF3-, and TRA2-activated exons corresponded to G-, S-, C-, and A-rich codons, respectively (see also Supplementary Figure S3 and Supplementary Table S2). This was in sharp contrast to the corresponding repressed exons (Figure 2A, lower panels). For example, glycine (GGN codons) was more frequently encoded by SRSF1-activated exons than by control exons and SRSF1-repressed exons (Figure 2B). A counting of glycine encoded within SRSF1-activated versus SRSF1-repressed exons showed a mirrored distribution: a large proportion of activated exons (~60%) coded for more than 3 glycine, whereas nearly 70% of repressed exons coded for a maximum of 1 glycine (Figure 2C). Similarly, alanine (GCN codons), proline (CCN codons) and lysine (AAR codons) were more frequently encoded by SRSF2-, SRSF3- and TRA2-activated exons, respectively (Figures 2B and 2C).

The above observations revealed a nucleotide composition bias of splicing factor-regulated exons and a bias regarding the nature of the amino acids that are encoded by these exons. In this setting, it is well established that amino acids sharing similar physicochemical properties (e.g., size, hydropathy, charge) are encoded by similar codons (i.e., codons composed of the same nucleotides)(Woese 1965; Wolfenden et al. 1979; Taylor and Coates 1989; Biro et al. 2003; Prilusky and Bibi 2009). For example, small amino acids (Ala, Asn, Asp, Cys, Gly, Pro, Ser, Thr), and in particular very small amino acids (Ala, Gly, Ser, Cys) are encoded by S-rich codons while large amino acids (Arg, Ile, Leu, Lys, Met, Phe, Trp, Tyr) are encoded by S-poor codons (Figure 3A). SRSF2 binds to SSNG motifs and SRSF2-activated exons are S-rich (see Figure 1). Remarkably, the two sets of analyzed SRSF2-activated exons encoded more frequently for very small and small amino acids when compared to control exons in contrast to SRSF2-repressed exons (Figure 3B and Supplementary Table S2). Conversely, large amino acids were less frequently encoded by SRSF2-activated exons (Figure 3B). Accordingly, a larger proportion of SRSF2-activated exons corresponded to peptides containing very small amino acids, and a smaller proportion of these exons corresponded to peptides containing large amino acids, when compared to SRSF2-repressed exons (Figure 3C).

Amino acids can be classified in three families in regards to their hydropathy and each family is encoded by codons having different features. Hydrophilic amino acid (Arg, Asn, Asp, Gln, Glu, Lys) are encoded by A-rich codons, hydrophobic amino acids (Ala, Cys, Ile, Leu, Met, Phe, Val) are

encoded by U-rich and A-poor codons, and ambivalent or neutral amino acids (Gly, His, Pro, Ser, Thr, Tyr) are encoded by C-rich codons (Figure 3D) (Kyte and Doolittle 1982; Engelman et al. 1986; Chiusano et al. 2000; Biro et al. 2003; Pommie et al. 2004; Prilusky and Bibi 2009; Zhang and Yu 2011). TRA2 that binds to AGAA-like motifs activates the inclusion of A-rich exons (see Figure 1). Interestingly, TRA2-activated exons encoded hydrophilic amino acids more frequently and neutral or hydrophobic amino acids less frequently than TRA2-repressed or control exons (Figure 3E and Supplementary Table S2). This result was confirmed using different hydrophobicity propensity scales (Figure 3F). Accordingly, a larger proportion of TRA2-activated exons compared to TRA2-repressed exons encoded peptides containing hydrophilic amino acids (Figure 3G).

Polar uncharged amino acids (Asn, Gln, Ser, Thr, Tyr), including hydroxyl-containing amino acids (e.g., Ser and Thr) correspond to C-rich and G-poor codons, while polar charged amino acids (Asp, Glu, Lys, Arg) correspond to G-rich and C-poor codons (Figure 3H) (Biro et al. 2003; Zhang and Yu 2011). SRSF3 binds to C-rich motifs and activates C-rich and G-poor exons (Figure 1). Remarkably, two sets of SRSF3-activated exons encoded uncharged (including hydroxyl-containing) amino acids more frequently than control exons or SRSF3-repressed exons (Figure 3I and Supplementary Table S2). They also encoded more frequently hydropathically neutral amino acids (Figure 3I) that correspond to C-rich codons (Figure 3D). As shown in Figure 3J, a larger proportion of SRSF3-activated exons (compared to SRSF3-repressed exons) corresponded to peptides composed of polar uncharged than charged amino acids. In particular, a larger proportion of SRSF3-activated exons corresponded to peptides composed of hydroxyl-containing amino acids (that can be negatively charged after phosphorylation) than negatively charged amino acids (Figure 3J, right panel). These observations suggest a link between splicing-related nucleotide composition bias of splicing-regulated exons and physicochemical properties of the exon-encoded amino acids.

hnRNP-corepressed exons code for amino acids with similar physicochemical properties

SRSF-like splicing factors activate exons they bind to, in contrast to hnRNP-like splicing factors that repress exons they bind to. We analyzed several RNA-seq datasets generated from different cell lines expressing or not hnRNP-like RNA-binding proteins (Supplementary Table S1 and Supplementary Figure S1), focusing on hnRNPH1, hnRNPK, hnRNPL, and PTBP1 that bind to G-rich motifs, C-rich motifs, , CA-rich motifs, and CU-rich motifs, respectively (Klimek-Tomczak et al. 2004; Katz et al. 2010; Llorian et al. 2010; Ray et al. 2013; Rossbach et al. 2014; Hauer et al. 2015; Giudice et al. 2016). As shown in Figure 4A, hnRNPH1-repressed exons were enriched in Gs, while hnRNPK-repressed exons were enriched in Cs when compared to control exons. This nucleotide composition bias was observed at the first and second codon positions (Figure 4B). Accordingly, glycine (GGN codons) and proline (CCN codons) were more frequently encoded by hnRNPH1-repressed and by

hnRNPK-repressed exons, respectively (Figure 4C). As in the case of C-rich SRSF3-activated exons (Figure 3I), C-rich hnRNPK-repressed exons encode more frequently for uncharged and hydropathically neutral amino acids than control exons (Figure 4D).

As mentioned above, hnRNPL represses the inclusion of exons containing CA-rich motifs, while PTBP1 represses the inclusion of exons containing CU-rich motifs. CA and AC dinucleotides were enriched in hnRNPL-repressed exons, while CU and UC dinucleotides were enriched in PTBP1-repressed exons compared to control exons (Figure 4E, left panel). Interestingly, histidine (His, CAY codons), glutamine (Gln, CAR codons) and threonine (Thr, ACN codons) were more frequently encoded by hnRNPL-repressed exons than by control exons or PTBP1-repressed exons (Figure 4E, right panel). Meanwhile, serine (Ser, UCN codons) was more frequently encoded by PTBP1-repressed exons than by control exons or hnRNPL-repressed exons (Figure 4E, right panel). Both hnRNPL- and PTBP1-repressed exons encode more and less frequently for hydroxyl-containing and negatively charged amino acids, respectively than control exons (Figure 4F). This is consistent with the fact that hydroxyl-containing and charged amino acids are C-rich and C-poor, respectively (Figure 3H).

In conclusion, each set of SRSF- or hnRNP-coregulated exons has nucleotide composition bias and codes for amino acids with similar physicochemical properties.

Bidirectional interplay between the splicing regulatory process and its functional outcome

The physicochemical properties of amino acids are often more or less related, since for example, hydrophilic amino acids are often charged amino acids. In addition, we observed that exons regulated by a given splicing factor often coded for amino acids that have different physicochemical properties depending on whether the exons are activated or repressed by this factor (Figure 3). Consequently, each splicing factor induced a shift toward different combinations of protein physicochemical properties encoded by their regulated exons (Figures 5A and 5B).

Since local protein features depend on amino acid physicochemical properties, we measured an enrichment Z-score of annotated protein features to determine whether each factor-specific set of exons encode specific protein features using our recently developed Exon Ontology bioinformatics suite (Tranchevent et al. 2017). As shown in Figure 5C, all sets of SRSF-activated exons preferentially encode intrinsically unstructured protein regions (IUPR), in agreement with previous reports indicating that alternatively spliced exons often code for intrinsically disordered regions (Tranchevent et al. 2017). However, each set of SRSF-coregulated exons encodes specific sets of annotated protein features. For example, C-rich SRSF3-activated exons encode peptides that are enriched for experimentally validated phospho-serine and -threonine (“PTM”, Figure 5C and Figure 5D). These phosphorylation sites arise in serine- and proline-rich regions (Figure 5E). This observation is consistent with the fact that hydroxyl-containing amino acids and proline correspond to C-rich

codons (Figure 3H). Serine- and proline-rich regions have been shown to play a role in RNA-protein interactions that can be regulated by phosphorylation (Wang et al. 2006; Thapar 2015). In this setting, SRSF3-activated exons encode annotated “Nucleic Acid Binding” activity (Figure 5C).

Along the same line, A-rich TRA2-activated exons often encode for nuclear localization signal (“NLS”, Figure 5F). This is consistent with the fact that they encode hydrophilic amino acids, in particular lysine that correspond to A-rich codons (Figures 2B-2C and 3D-3G), a major amino acid of classical NLS (Marfori et al. 2011). In contrast, A-poor TRA2-repressed exons code for intramembrane protein parts (“Mne”, Figure 5F) which are intrinsically rich in hydrophobic amino acids. This is consistent with the fact that A-poor TRA2-repressed exons encode more frequently hydrophobic amino acids corresponding to A-poor codons (Figure 3D-3G). Collectively, these observations support a model where a splicing factor-related nucleotide composition bias of exons (Figures 1 and 2) impacts on the physicochemical properties of their encoded amino acids because of the non-randomness of the genetic code (Figure 3) with direct consequences on protein features encoded by splicing factor-regulated exons (Figures 5A-5F).

On one hand, splicing factors bind to sequences that have a biased nucleotide composition and on the other hand, amino acids with similar physicochemical properties are encoded by codons having the same nucleotide composition bias. Therefore, we anticipated that increasing the exonic density of specific nucleotides as measured in splicing factor-regulated exons would increase the density of encoded amino acids sharing the same physicochemical properties, as observed in those exons. To challenge this possibility, we generated random exonic coding sequences enriched in specific nucleotide(s) by following the human codon usage bias (labelled CUB sequences) or by randomly mutating human coding exons (labelled MUT sequences). For example, we generated 100 sets of 300 coding exons containing either 53% or 47% of S nucleotides, as measured in SRSF2-activated and SRSF2-repressed exons, respectively. Increasing by ~13% the density of S nucleotides in coding exons (S-CUB or S-MUT) increased (by ~15%) the frequency of encoded very small amino acids, while it decreased (by ~10%) the frequency of encoded large amino acids (Figure 5G, “S=53% vs 47%”), as observed when comparing SRSF2-activated and SRSF2-repressed exons (Figure 3B). The increase of the density of A nucleotides in coding exons (A-CUB and A-MUT) from 23% to 34%, as measured in TRA2-repressed and TRA2-activated exons, respectively, increased (by ~40%) the frequency of encoded hydrophilic amino acids and it decreased (by ~20%) the frequency of encoded hydrophobic amino acids (Figure 5G, “A= 34% vs 23%”), as observed when comparing TRA2-activated and TRA2-repressed exons (Figure 3E). Finally, an increase in the density of C nucleotides in coding exons (C-CUB and C-MUT) from 21% to 29%, as measured in SRSF3-repressed and SRSF3-activated exons, respectively, increased the frequency of encoded uncharged amino acids and neutral amino

acids while it decreased the frequency of encoded charged amino acids (Figure 5G, “C= 29% vs 21%”), as observed when comparing SRSF3-activated and SRSF3-repressed exons (Figure 3I).

We next generated exons coding for different proportions of amino acids sharing the same physicochemical features by mutating randomly-selected human coding exons. For example, we generated 100 sets of 300 mutated exons encoding different proportions of very small and large amino acids, using their respective frequency measured in SRSF2-activated or SRSF2-repressed exons (See Materials and Methods). Mutated exons coding more frequently very small rather than large amino acids had a higher frequency of S nucleotides and GC dinucleotides and contained more frequently S-rich low complexity sequences (Figure 5H, “SRSF2-like encoded properties”), as observed when comparing SRSF2-activated to SRSF2-repressed exons (Figure 1). Mutated exons encoding more frequently hydrophilic amino acids had a higher frequency of A nucleotides and AA dinucleotides and contained more frequently A-rich low complexity sequences (Figure 5H, ““TRA2-like encoded properties”), as observed when comparing TRA2-activated to TRA2-repressed exons (Figure 1). Finally, mutated exons encoding more frequently hydropathically neutral amino acids had a higher frequency of C nucleotides and CC dinucleotides and contained more frequently C-rich low complexity sequences (Figure 5H, “SRSF3-like encoded properties”), as observed when comparing SRSF3-activated to SRSF3-repressed exons (Figure 1).

Discussion

In this work, we uncover a direct link between the splicing regulatory process and its biological outcome relying on two straightforward principles: 1) splicing factors bind to exonic sequences that have a nucleotide composition bias with consequences on the nucleotide composition of codons from coregulated exons; and 2) codons having the same nucleotide composition bias encode amino acids with similar physicochemical properties with consequences on protein features encoded by the coregulated exons.

Exons coregulated by a given splicing factor are enriched for specific low-complexity sequences (often composed of a repeated (di)nucleotide) that correspond to the RNA binding sites of the cognate factor (Figures 1A and 1C). We showed that each set of exons of which inclusion (or exclusion) is enhanced by a given splicing factor is enriched for specific nucleotide(s) when compared to control exons or to exons repressed (or activated, respectively) by the same factor (Figures 1B, 2D, and 2H). To the best of our knowledge, this splicing-related exonic nucleotide composition bias has not been reported yet. However, it is in agreement with recent observations indicating that the interaction of a splicing factor with a binding motif depends on the sequence context and on the presence of clusters of related binding motifs (Zhang et al. 2013; Cereda et al. 2014; Fu and Ares

2014; Daniel Dominguez 2017; Jobbins et al. 2018). For example, increasing the exonic frequency of GGA-like motifs increases the probability of an exon to be regulated by the SRSF1 splicing factor that binds to GGAGGA-like motifs even though only one binding site is used (Jobbins et al. 2018).

Coding sequences overlap several kinds of regulatory sequences, including exonic splicing regulatory sequences. To date, it has been assumed that the redundancy of the genetic code permits protein-coding regions to carry this extra information (Goren et al. 2006; Itzkovitz and Alon 2007; Itzkovitz et al. 2010; Lin et al. 2011; Shabalina et al. 2013; Savisaar and Hurst 2017b; Savisaar and Hurst 2017a). This means that the sequence constraints imposed by splicing factor binding motifs would accommodate with coding sequences by impacting only the third codon position. However, we observed that nucleotide composition bias of splicing factor-regulated exons impacts the first and second codon positions (Figures 1D and 4B). Since amino acids having the same physicochemical properties correspond to codons with similar nucleotide composition bias, a direct consequence of the exonic nucleotide composition bias associated with the splicing regulatory process is that each set of splicing factor-regulated exons preferentially encodes amino acids having similar physicochemical properties (Figures 3 and 4). In addition, since specific local protein features depend on amino acid physicochemical properties, splicing factor-coregulated exons encodes specific sets of protein features (Figures 5A-5F).

Therefore, we propose that the interplay between coding and exonic splicing regulatory sequences that we report is based on straightforward principles related to both the non-randomness of the genetic code and the preferential binding of splicing factors to low-complexity sequences. Because of these properties, the high exonic density of a specific nucleotide related to splicing factor-binding features increases the probability that an exon encodes amino acids with similar physicochemical properties (Figure 5G). Conversely, the high density of amino acids corresponding to specific physicochemical properties increases the probability of generating exonic nucleotide composition bias and nucleotide low complexity sequences (Figure 5H).

In this setting, the function of splicing factors would not only be to regulate the production of individual specialized protein isoforms, but it would also be to control more globally the intracellular content of specific protein regions having specific physicochemical properties. Each splicing factor would control a specific combination of exon-encoded protein physicochemical properties accordingly to its affinity for specific nucleotides. Our work unravels how a complex phenomenon (e.g., splicing regulatory process and its biological consequences) can rely on straightforward principles.

Materials and Methods

RNA-seq dataset analyses

Publicly available RNA-seq datasets generated from different cell lines expressing or not various splicing factors as described in Supplementary Table S1 were analyzed using FARLINE. FARLINE is a computational program dedicated to analyze and quantify alternative splicing variations, as previously reported(Benoit-Pilven et al. 2018).

Frequency of hexanucleotides, dinucleotides, nucleotides, codons, amino acids, and amino acid physicochemical features in exon sets.

The formula (1) was used to compute the frequencies of words (D_n) of size n within a set of exons $S_n = \{y_1, \ldots, y_N\}$ such that y_i is an exon i having a number L_i of nucleotides:

$$\begin{cases} if\ n \in \{2,1\} \Rightarrow \dfrac{\sum_{i=1}^{N}\left(\frac{x_i}{L_i-(n-1)}\right)}{N} \\ else\ if\ n = 6 \Rightarrow \dfrac{\sum_{i=1}^{N}\left(\frac{x_i}{L_i-(n-1)} \times min\left(\left(\frac{L_i}{P}\right),1\right)\right)}{\sum_{i=1}^{N} min\left(\frac{L_i}{P},1\right)} \\ else\ if\ n = 3 \Rightarrow \dfrac{\sum_{i=1}^{N}\left(\frac{x_i}{L_i/3} \times min\left(\left(\frac{L_i}{3P}\right),1\right)\right)}{\sum_{i=1}^{N} min\left(\frac{L_i}{3P},1\right)} \end{cases} (1)$$

where x_i is the number of occurrences of D_n in exon i, n is set to 6, 2, and 1 for hexanucleotides, dinucleotides and nucleotides, respectively. For codons, amino acids and amino acid physicochemical properties, n is set to 3. P=50 is a penalty size used to decrease the border effects seen in small exons and N *is* the number of exons in the set S_n. For hexanucleotides and dinucleotides, the occurrences x_i of D_n are overlapping whereas they are contiguous for the others. In the particular case of amino acids and amino physiochemical properties D_n represents a group of codons encoding the same amino acid or the same physiochemical properties respectively. When coding phase is mandatory, incomplete codons at exon borders were deleted.

Very small (Ala, Gly,Ser, Cys), small (Ala, Asn, Asp, Cys, Gly, Pro, Ser, Thr), large (Arg, Ile, Leu, Lys, Met, Phe, Trp, Tyr), polar uncharged (Asn, Gln, Ser, Thr, Tyr), charged (Asp, Glu, Lys, Arg), hydroxyl-containing (Ser, Thr, Tyr), hydrophilic (Arg, Asn, Asp, Gln, Glu, Lys), hydro-neutral (Gly, His, Pro, Ser, Thr, Tyr), and hydrophobic (Ala, Cys, Ile, Leu, Met, Phe, Val) amino acids were classified as previously reported (*43-45*).

The hydrophobicity scale was calculated as defined by Kyte et al.(*44*) and Engelman et al(*45*). TRA2-activated or -repressed exons larger or equal than 30 amino acids were selected to calculate the average of hydrophobicity using a sliding window of 5 amino acids with a step of 1 amino acid for

the 30 first and 30 last amino acids. The mean and the standard deviation of the hydrophobicity values corresponding to each exon set were then calculated for each position of the window.

Generation of sets of control exons and statistical analyses.

To test whether a feature was enriched in a set S_N of N exons, a randomization test was made by sampling, from FasterDB(Mallinjoud et al. 2014), 10000 sets of control exons, **C** = {C_1, … , C_{10000}}, with C_l = { $y_{l,1}$, … , $y_{l,i}$} such that $y_{l,i}$ is the exon i having a number of $L_{l,i}$ nucleotides following the constraints:

$$L_{l,i} = \begin{cases} if L_i < 50 \Rightarrow L_{l,i} \in \left[\frac{L_i}{3}, max(L_i \times 3{,}50)\right] \\ else\ if\ 50 \leq L_i \leq 300 \Rightarrow L_{l,i} \in \left[\frac{L_i}{2}, L_i \times 2\right] \\ else\ L_i > 300 \Rightarrow L_{l,i} \in [300, +\infty] \end{cases}$$

where 50 and 300 nucleotides correspond to the 5th and 95th quantile of the distribution of the length of all the exons defined in FasterDB.

The relative frequency of a feature D_n in S_n compared to the sets of control exons **C** was calculated by the formula:

$$RFreq(D_n) = \frac{Freq_{obs}(D_n) - \frac{1}{10000} \times \left(\sum_{l=1}^{10000} Freq_{control,l}(D_n)\right)}{\frac{1}{10000} \times \left(\sum_{l=1}^{10000} Freq_{control,l}(D_n)\right)}$$

Where $Freq_{obs}(D_n)$ is the frequency (as in equation (1)) of a word D_n of size n in S_n and $\frac{1}{10000}\sum_{l=1}^{10000} Freq_{control,l}(D_n)$ is the average frequency (as in equation (1)) of D_n in **C**.

In order to calculate an empirical p-value, the number of control frequencies upper or lower than the frequency in the set of interest is determined. Then the smallest number between those two is kept and divided by the number of control sets (*i.e.,* 10000).

All p-values obtained for each set of features have been corrected using Benjamini-Hochberg correction.

The nucleotide composition of enriched codons or codons corresponding to enriched amino acids was calculated after recovering codons or amino acids whose frequency was 10% higher in the set of exons of interest than their average frequency in sets of control exons.

Low complexity sequences and random sequences.

Low complexity sequences were defined as sequences of n (n=5 to 10) nucleotides containing at least (n-1) occurrences of the same nucleotide.

Random exonic sequences (from 50- to 300-long nucleotides) with specific nucleotide composition bias were generated using two strategies. First, random codons sequences respecting the human codon usage bias (CUB exons) were generated. These sequences were then mutated

randomly, one nucleotide at a time, to increase or decrease the frequency of a specific nucleotide. Only mutation increasing or decreasing the frequency toward $E \sim N\left(Freq_{obs}(D_1), \frac{1}{30}\right)$(where $Freq_{obs}(D_1)$ is the nucleotide frequency observed in a specific set of activated or repressed exons by a given splicing factor) were kept. The mutation procedure was stopped when E was reached. Second, exonic sequences (MUT exons), selected by sampling human coding exons, were mutated using the same principle used for CUB sequences. In each case, 100 sets of 300 exonic sequences with specific features were generated. A t-test was performed to compare the frequency of amino acid physicochemical properties between the generated sets.

Exonic sequences encoding for specific amino acid physicochemical properties were generated from sampled human coding exons (with the same criteria as CUBs). These sequences were modified by codon substitution to increase the frequency of amino acids encoding for a given physicochemical property P1 and to decrease the frequency of another given physicochemical property P2. Codons that encode P2 were substituted toward codons encoding P1 following the human codon usage bias. SRSF2-like encoded properties exons were generated using the frequency of very small (0.27) and large (0.34) amino acids measured in SRSF2-activated exons or the frequency of very small (0.21) and large (0.38) amino acids measured in SRSF2-repressed exons. TRA2-like encoded properties exons were generated using the frequency of hydrophilic (0.4) and hydrophobic (0.33) amino acids measured in TRA2-activated exons or the frequency of hydrophilic (0.26) and hydrophobic (0.39) amino acids measured in TRA2-repressed exons. SRSF3-like encoded properties exons were generated using the frequency of hydro-neutral (0.38) and charged (0.17) amino acids measured in SRSF3-activated exons or the frequency of hydro-neutral (0.31) and charged (0.22) amino acids measured in SRSF3-repressed exons. The same procedure as for CUBs was used to compare the frequencies of nucleotides or dinucleotides with a t-test.

Density charts

For each exon, the frequency of each nucleotide or each amino acid physicochemical property was calculated using the formula (1). The exonic sequences were parsed using a sliding window (of size 1 and step 1). Truncated codons (at 3' or 5' exon extremities) or codons downstream of stop codons were ignored. Frequency histograms were then computed with R software.

Acknowledgments

We gratefully acknowledge support from the PSMN (Pôle Scientifique de Modélisation Numérique) of the ENS de Lyon for the computing resources. We thank the members of the LBMC biocomputing center for their involvement in this project. This work was funded by Fondation ARC (PGA120140200853), INCa (2014-154), ANR (CHROTOPAS), AFM-Téléthon, and LNCC. J.B.C was supported by Fondation de France. None of the authors have any competing interests.

Figure legends

Figure 1

A. Color-code of the relative frequency (%) of hexanucleotides in SRSF-activated and -repressed exons. After recovering the 10 most enriched hexanucleotides in SRSF-activated exons, their relative frequency was calculated by comparing it to the average frequency calculated in 10 000 sets of control exons. The relative frequency of these hexanucleotides was also calculated in SRSF-repressed exons. Red and green colors indicate when the frequency is higher and lower, respectively, in the set of regulated exons compared to the sets of control exons. The sets of SRSF-regulated exons originated from publicly available RNAseq datasets generated from the K562 (1), HepG2 (2), GM19238 (3), HeLa (4), K562 (5), Huh7 (6), HepG2 (7), GM19238 (8), and MDA-MB-231 (9) cell lines. Purine, SSNG motifs, Cs, and As are underlined in the enriched hexanucleotides identified in SRSF1-, SRSF2-, SRSF3-, and TRA2-activated exons, respectively. (*) Randomization test FDR-adjusted P-value<0.05 (for the 10 most enriched hexanucleotides).

B. The upper panels represent the relative frequency (%) when compared to sets of control exons of G, S, C, and A nucleotides in SRSF1-, SRSF2-, SRSF3-, and TRA2-regulated exons, respectively, identified in different cell lines as described in Figure 1A. (*) Randomization test FDR-adjusted P-value <0.05. The lower panels represent the density chart of G, S, C, and A nucleotides in SRSF1-, SRSF2-, SRSF3-, and TRA2-regulated exons, respectively. (**) KS-test $<10^{-13}$.

C. The upper panels represent the color-code of the relative frequency (%) of dinucleotides, compared to sets of control exons, in SRSF-activated and SRSF-repressed exons across different cell lines as depicted in Figure 1A. The frequency of each dinucleotide was calculated in SRSF-activated and SRSF-repressed exons and expressed as the % of the average frequency calculated in sets of control exons. Red and green colors indicate when the dinucleotide frequency is higher and lower, respectively, in the sets of regulated exons when compared to sets of control exons. Only the two most enriched dinucleotides in SRSF-activated compared to SRSF-repressed exons are represented. (*) Randomization test FDR adjusted P-value<0.05. The lower panels represent the proportion of exons containing at least one low complexity (LC) sequence of 6, 7, 9, or 10 nucleotides. In a sliding window of n nucleotides, the number of the same nucleotide (G, S, C, or A) must be equal to or greater than (n-1). The average of 4 datasets is represented for SRSF1. A logistic regression analysis was performed to test if activated or repressed exons for a given splicing factor have a different content in low complexity sequences while accounting for cell-line variations. (*) P-value<0.05.

D. The upper panels represent the color-code of the relative frequency (%), compared to sets of control exons, of some codons in SRSF-activated and SRSF-repressed exons across different cell lines as depicted in Figure 1A. The frequency of each codon was calculated in SRSF-activated and SRSF-

repressed exons and expressed as the % of the average frequency calculated in sets of control exons. Red and green colors indicate when the codon frequency is higher and lower, respectively, in the sets of regulated exons when compared to sets of control exons. Only some enriched codons identified in SRSF-activated exons are represented. (*) Randomization test FDR-adjusted P-value <0.05. The lower panels represent the relative frequency (%) when compared to sets of control exons, of G (G1-2), S (S1-2), C (C1-2), or A (A1-2) nucleotides at the first and second codon positions in SRSF-activated and -repressed exons identified across different cell lines as depicted in Figure 1A. (*) Randomization test FDR adjusted P-value <0.05.

Figure 2

A. Nucleotide composition of codons corresponding to amino acids more frequently encoded, when compared to sets of control exons, by SRSF1-, SRSF2-, SRSF3-, or TRA2-activated (upper panels) and -repressed exons (lower panels).

B. Relative frequency (%) when compared to sets of control exons, of glycine (Gly corresponding to GGN codons), alanine (Ala corresponding to GCN codons), proline (Pro corresponding to CCN codons), and lysine (Lys corresponding to AAR codons) encoded by SRSF1-, SRSF2-, SRSF3-, or TRA2-activated and -repressed exons identified across different cells lines as depicted in Figure 1A. (**) Randomization test FDR adjusted P-value < 0.05.

C. Proportion (%) of exons from SRSF1-, SRSF2-, SRSF3, or TRA2-regulated exons encoding for 0, 1, 2, and more than three Gly, Ala, Pro, or Lys amino acids, respectively. The average of 4 datasets is represented for SRSF1. A logistic regression analysis was performed to test if activated or repressed exons for a given splicing factor have a different content in codons encoding a particular amino acid while accounting for cell-line variations. (*) P-value<0.05.

Figure 3

A. Nucleotide composition of codons encoding small, very small and large amino acids. S=G or C.

B. Relative frequency (%) when compared to sets of control exons, of very small, small and large amino acids encoded by two sets of SRSF2-activated and -repressed exons identified in the K562 (1) and Huh7 (2) cell lines. (**) Randomization test FDR-adjusted P-value <0.05.

C. Density chart of SRSF2-activated and -repressed exons identified in K562 cells coding for very small and large amino acids. (**) KS-test $<10^{-5}$.

D. Nucleotide composition of codons encoding hydrophobic, neutral and hydrophilic amino acids.

E. Relative frequency (%) when compared to sets of control exons, of hydrophilic, neutral and hydrophobic amino acids encoded by TRA2-activated and TRA2-repressed exons. (**) Randomization test FDR-adjusted P-value <0.05.

F. Hydrophobic scales of TRA2-activated and TRA2-repressed exons. The green bottom line indicates the Mann-Whitney test P-value <0.05 at each amino acid position.

G. Density chart of TRA2-activated or –repressed exons coding for hydrophilic amino acids. (**) KS-test $<10^{-13}$.

H. Nucleotide composition of codons encoding polar uncharged, hydroxyl-containing and charged amino acids.

I. Relative frequency (%) when compared to sets of control exons, of polar uncharged, hydroxyl-containing, charged or neutral (in terms of hydropathy) amino acids encoded by two sets of SRSF3-activated and -repressed exons identified from HepG2 (1) and GM19238 (2) cell lines. (**) Randomization test FDR-adjusted P-value <0.05.

J. Density chart of SRSF3-activated and -repressed exons coding for polar uncharged, hydroxyl-containing, or charged amino acids. (**) KS-test $<10^{-4}$.

Figure 4

A. Relative frequency (%), when compared to sets of control exons, of G and C nucleotides in hnRNPH1- and hnRNPK-repressed exons identified in 293T (1), K562 (2), GM19238 (3), and HepG2 (4) cell lines. (**) Randomization test FDR-adjusted P-value < 0.05.

B. Relative frequency (%) when compared to sets of control exons, of G (G1-2) or C (C1-2) nucleotides at the first and second codon position from hnRNPH1- or hnRNPK-repressed exons identified in 293T (1), K562 (2), GM19238 (3), and HepG2 (4). (**) Randomization test FDR-adjusted P-value <0.05.

C. Relative frequency (%) when compared to sets of control exons, of glycine (Gly corresponding to GGN codons) and proline (Pro corresponding to CCN codons) encoded by hnRNPH1- and hnRNPK-repressed exons identified in 293T (1), K562 (2), GM19238 (3), and HepG2 (4) cell lines. (**) Randomization test FDR-adjusted P-value <0.05.

D. Relative frequency (%) when compared to sets of control exons, of polar uncharged, charged or neutral (in terms of hydropathy) amino acids encoded by three sets of hnRNPK-repressed exons identified from K562 (1), GM19238 (2), and HepG2 (3) cell lines. (**) Randomization test FDR-adjusted P-value <0.05.

E. The left panel represents the average of the relative frequency (%) when compared to sets of control exons of CA, CT, AC, and TC dinucleotides calculated from four sets of hnRNPL- or PTBP1-repressed exons. The right panel represents the average of the relative frequency (%) when

compared to sets of control exons of histidine (His corresponding to CAY codons), glutamine (Gln corresponding to CAR codons), leucine (Leu corresponding to CTN codons), threonine (Thr corresponding to ACN codons), and serine (Ser corresponding to TCN codons) encoded by four sets of hnRNPL- and hPTBP1repressed exons. (*) Mann-Whitney test P-value<0.05.

F. Relative frequency (%) when compared to sets of control exons, of hydroxyl-containing and negatively charged amino acids encoded by four sets of hnRNPL-repressed exons (K562 (1), HepG2 (2), LNCaP (3), GM19238 (4)) and four sets of PTBP1-repressed exons (HepG2 (5), 293T (6), HeLA (7), K562 (8) cells). (**) Randomization test FDR-adjusted P-value <0.05.

Figure 5

A. Color-code corresponding to the relative frequency (%) of amino acid physicochemical properties as indicated when comparing all the SRSF1-, SRSF2-, SRSF3-, and TRA2-activated exons to all the SRSF1-, SRSF2-, SRSF3-, and TRA2-repressed exons, respectively, or when comparing all the hnRNPH1, hnRNPK, hnRNPL, and PTBP1-repressed exons to all the hnRNPH1, hnRNPK, hnRNPL, and PTBP1-activated exons, respectively.

B. Relative frequency (%) of very small, large, hydrophilic, neutral, hydrophobic, charged, uncharged, negatively charged (charged -), and positively charged (charged +) amino acids when comparing all the SRSF1-, SRSF2-, SRSF3-, or TRA2-activated exons to all the SRSF1-, SRSF2-, SRSF3-, or TRA2-repressed exons, respectively. (*) Mann-Whitney FDR-adjusted p-value <0.05.

C. Color-code corresponding to the Z-score of annotated protein features encoded SRSF1-, SRSF2-, SRSF3-, and TRA2-activated exons compared to all human coding exons. IUPR, intrinsically unstructured regions; CBR, compositionally biased protein region; PTM, post-translational modifications. FDR-adjusted p-value <0.05.

D. Z-score of experimentally validated phosphorylated serine (pS) and threonine (pT) encoded by SRSF3-activated and -repressed exons compared to all human coding exons. FDR-adjusted p-value <0.05.

E. Sequence logo generated from the PhosphoSite website using sequences surrounding experimentally validated phosphorylated residues coded by SRSF3-activated exons.

F. Z-score of nuclear localization signal (NLS) and intramembrane peptides (Mne) terms encoded by TRA2-activated and -repressed exons compared to all human coding exons. FDR-adjusted p-value <0.05.

G. The left panel represents the relative frequency (%) of very small and large amino acids encoded by 100 sets of 300 generated-exonic sequences containing a high frequency (53%) of the S nucleotide (S-CUB and S-MUT) compared to 100 sets of 300 generated-exonic sequences containing a low S-

nucleotide frequency (47%). The middle panel represents the relative frequency (%) of hydrophilic and hydrophobic amino acids encoded by exonic sequences containing a high frequency (34%) of A nucleotide (A-CUB and A-MUT) compared to exonic sequences containing a low A-nucleotide frequency (23%). The right panel represents the relative frequency (%) of polar uncharged, charged and neutral (in terms of hydropathy) amino acids encoded by exonic sequences containing a high frequency (29%) of Cs(C-CUB and C-MUT) compared to exonic sequences containing a low C-nucleotide frequency (21%). (**) t-test p-value $<10^{-30}$.

H. The left panels represent the relative frequency (%) of the S nucleotide and GC dinucleotide as well as the relative proportion (%) of exons with S-rich low-complexity (LC) sequences of 100 sets of 300 mutated exons encoding for the same physicochemical properties than SRSF2-activated exons compared to 100 sets of 300 mutated exons encoding for the same physicochemical properties than SRSF2-repressed exons. The middle panels represent the relative frequency (%) of the A nucleotide and AA dinucleotide as well as the relative proportion (%) of exons with A-rich low-complexity sequences of mutated exons encoding for the same physicochemical properties than TRA2-activated exons compared to mutated exons encoding for the same physicochemical properties than TRA2-repressed exons. The right panels represent the relative frequency (%) of the C nucleotide and CC dinucleotide as well as the relative proportion (%) of exons with C-rich low-complexity (LC) sequences of mutated exons encoding for the same physicochemical properties than SRSF3-activated exons compared to mutated exons encoding for the same physicochemical properties than SRSF3-repressed exons. (**) t-test p-value $<10^{-30}$.

References

Anczukow O, Akerman M, Clery A, Wu J, Shen C, Shirole NH, Raimer A, Sun S, Jensen MA, Hua Y et al. 2015. SRSF1-Regulated Alternative Splicing in Breast Cancer. *Molecular cell* **60**(1): 105-117.

Anko ML, Muller-McNicoll M, Brandl H, Curk T, Gorup C, Henry I, Ule J, Neugebauer KM. 2012. The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. *Genome Biol* **13**(3): R17.

Benoit-Pilven C, Marchet C, Chautard E, Lima L, Lambert MP, Sacomoto G, Rey A, Cologne A, Terrone S, Dulaurier L et al. 2018. Complementarity of assembly-first and mapping-first approaches for alternative splicing annotation and differential analysis from RNAseq data. *Sci Rep* **8**(1): 4307.

Best A, Dalgliesh C, Kheirollahi-Kouhestani M, Danilenko M, Ehrmann I, Tyson-Capper A, Elliott DJ. 2014. Tra2 protein biology and mechanisms of splicing control. *Biochem Soc Trans* **42**(4): 1152-1158.

Biro JC, Benyo B, Sansom C, Szlavecz A, Fordos G, Micsik T, Benyo Z. 2003. A common periodic table of codons and amino acids. *Biochem Biophys Res Commun* **306**(2): 408-415.

Cereda M, Pozzoli U, Rot G, Juvan P, Schweitzer A, Clark T, Ule J. 2014. RNAmotifs: prediction of multivalent RNA motifs that control alternative splicing. *Genome Biol* **15**(1): R20.

Chiusano ML, Alvarez-Valin F, Di Giulio M, D'Onofrio G, Ammirato G, Colonna G, Bernardi G. 2000. Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code. *Gene* **261**(1): 63-69.

Daniel Dominguez PF, Maria S. Alexis, Amanda Su, Myles Hochman, Tsultrim Palden, Cassandra Bazile, Nicole J. Lambert, Eric L. Van Nostrand, Gabriel A. Pratt, Gene W. Yeo, Brenton Graveley, Christopher B. Burge. 2017. Sequence, Structure and Context Preferences of Human RNA Binding Proteins. *bioRxiv 201996*.

Dominguez D, Freese P, Alexis MS, Su A, Hochman M, Palden T, Bazile C, Lambert NJ, Van Nostrand EL, Pratt GA et al. 2018. Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Molecular cell* **70**(5): 854-867 e859.

Engelman DM, Steitz TA, Goldman A. 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem* **15**: 321-353.

Erkelenz S, Mueller WF, Evans MS, Busch A, Schoneweis K, Hertel KJ, Schaal H. 2013. Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *RNA* **19**(1): 96-102.

Fu XD, Ares M, Jr. 2014. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat Rev Genet* **15**(10): 689-701.

Geuens T, Bouhy D, Timmerman V. 2016. The hnRNP family: insights into their role in health and disease. *Hum Genet* **135**(8): 851-867.

Giudice G, Sanchez-Cabo F, Torroja C, Lara-Pezzi E. 2016. ATtRACT-a database of RNA-binding proteins and associated motifs. *Database (Oxford)* **2016**.

Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G. 2006. Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers. *Molecular cell* **22**(6): 769-781.

Grellscheid SN, Dalgliesh C, Rozanska A, Grellscheid D, Bourgeois CF, Stevenin J, Elliott DJ. 2011. Molecular design of a splicing switch responsive to the RNA binding protein Tra2beta. *Nucleic acids research* **39**(18): 8092-8104.

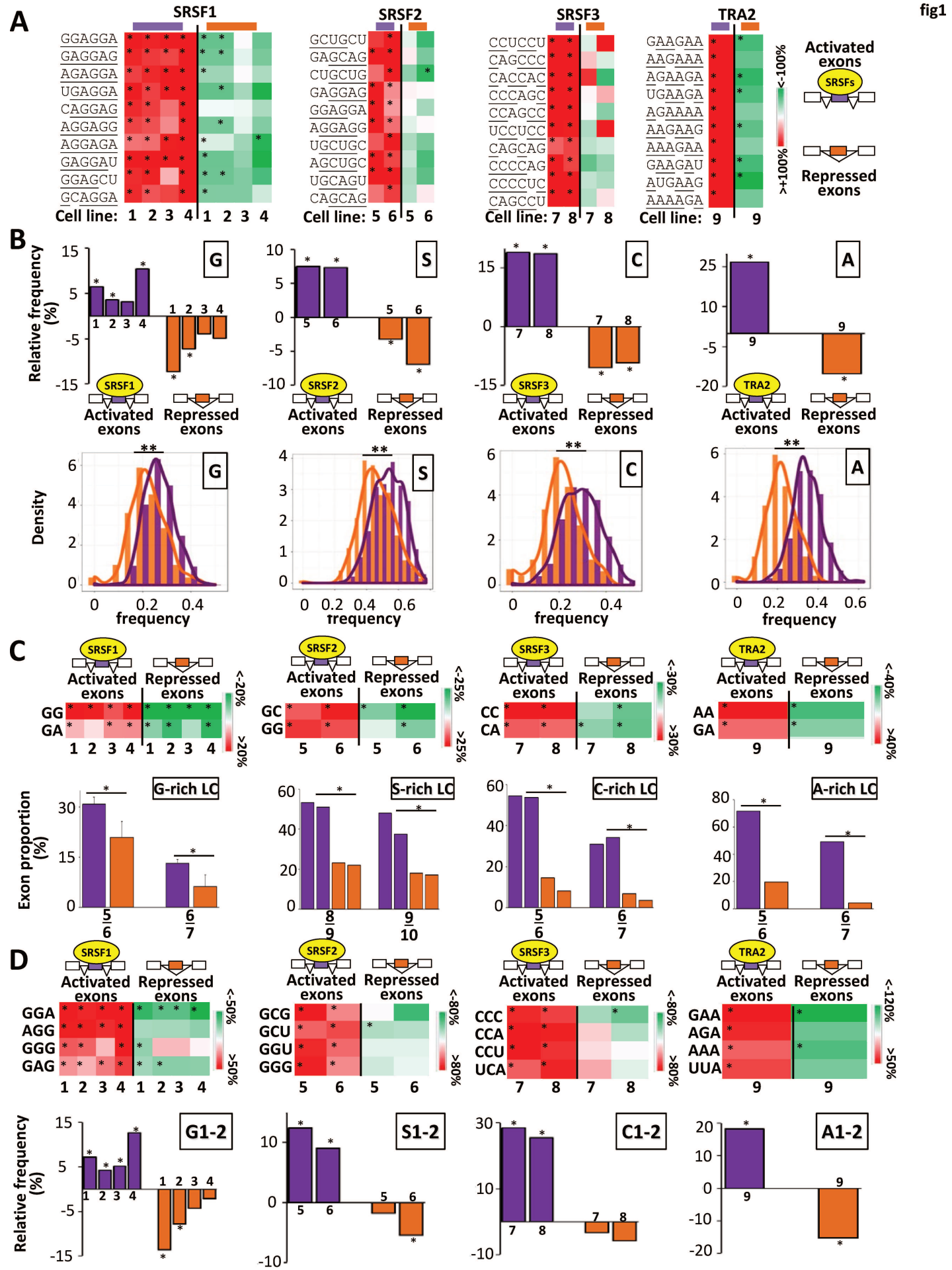
Hauer C, Curk T, Anders S, Schwarzl T, Alleaume AM, Sieber J, Hollerer I, Bhuvanagiri M, Huber W, Hentze MW et al. 2015. Improved binding site assignment by high-resolution mapping of RNA-protein interactions using iCLIP. *Nature communications* **6**: 7921.

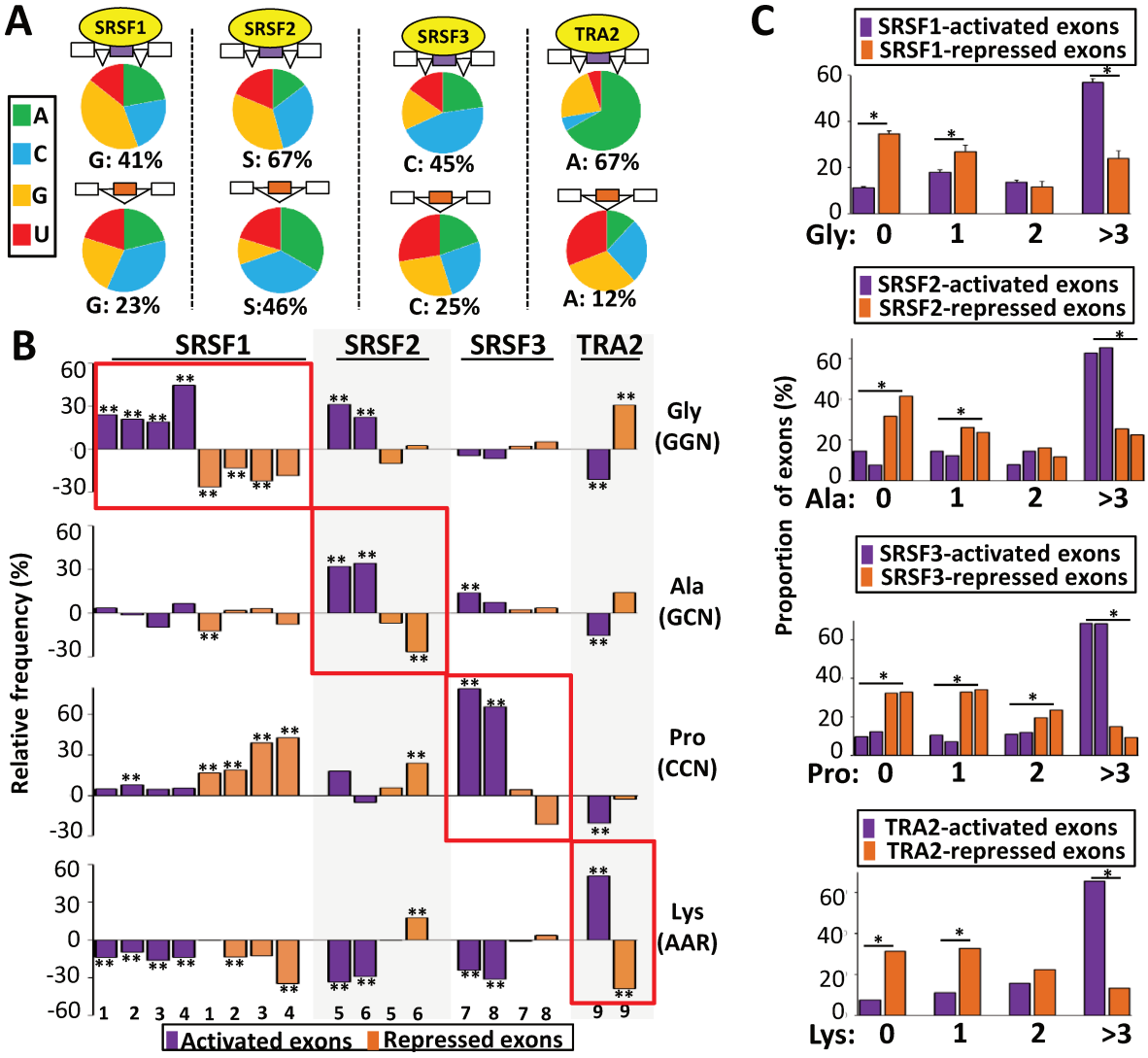
Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, Quesnel-Vallieres M, Tapial J, Raj B, O'Hanlon D et al. 2014. A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**(7): 1511-1523.

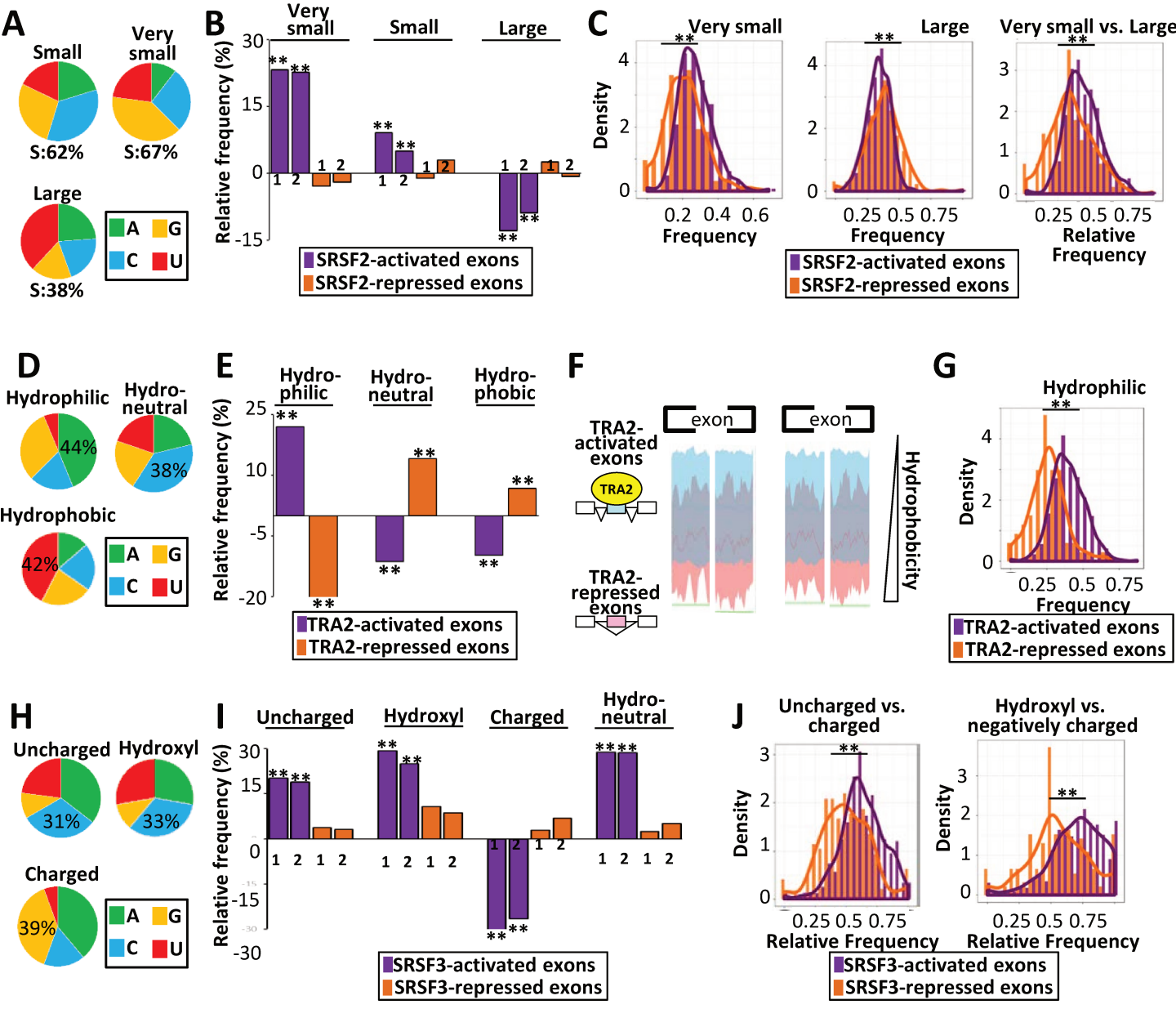
Itzkovitz S, Alon U. 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome research* **17**(4): 405-412.

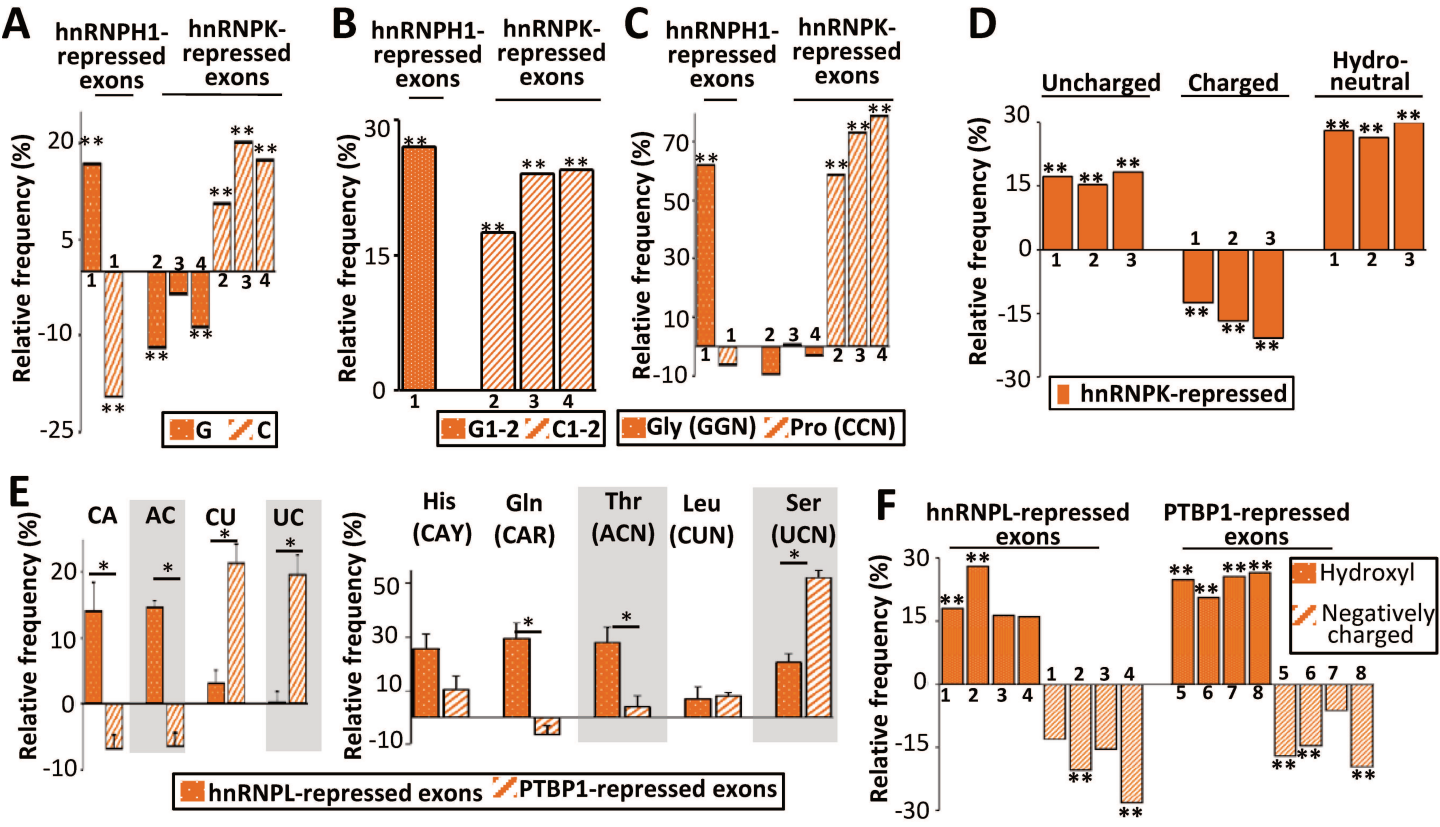
Itzkovitz S, Hodis E, Segal E. 2010. Overlapping codes within protein-coding sequences. *Genome research* **20**(11): 1582-1589.
Jobbins AM, Reichenbach LF, Lucas CM, Hudson AJ, Burley GA, Eperon IC. 2018. The mechanisms of a mammalian splicing enhancer. *Nucleic acids research*.
Katz Y, Wang ET, Airoldi EM, Burge CB. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**(12): 1009-1015.
Klimek-Tomczak K, Wyrwicz LS, Jain S, Bomsztyk K, Ostrowski J. 2004. Characterization of hnRNP K protein-RNA interactions. *J Mol Biol* **342**(4): 1131-1141.
Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**(1): 105-132.
Lin MF, Kheradpour P, Washietl S, Parker BJ, Pedersen JS, Kellis M. 2011. Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. *Genome research* **21**(11): 1916-1928.
Llorian M, Schwartz S, Clark TA, Hollander D, Tan LY, Spellman R, Gordon A, Schweitzer AC, de la Grange P, Ast G et al. 2010. Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. *Nat Struct Mol Biol* **17**(9): 1114-1123.
Luo C, Cheng Y, Liu Y, Chen L, Liu L, Wei N, Xie Z, Wu W, Feng Y. 2017. SRSF2 Regulates Alternative Splicing to Drive Hepatocellular Carcinoma Development. *Cancer Res* **77**(5): 1168-1178.
Mallinjoud P, Villemin JP, Mortada H, Polay Espinoza M, Desmet FO, Samaan S, Chautard E, Tranchevent LC, Auboeuf D. 2014. Endothelial, epithelial, and fibroblast cells exhibit specific splicing programs independently of their tissue of origin. *Genome research* **24**(3): 511-521.
Marfori M, Mynott A, Ellis JJ, Mehdi AM, Saunders NF, Curmi PM, Forwood JK, Boden M, Kobe B. 2011. Molecular basis for specificity of nuclear import and prediction of nuclear localization. *Biochimica et biophysica acta* **1813**(9): 1562-1577.
Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**(12): 1413-1415.
Pandit S, Zhou Y, Shiue L, Coutinho-Mansfield G, Li H, Qiu J, Huang J, Yeo GW, Ares M, Jr., Fu XD. 2013. Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Molecular cell* **50**(2): 223-235.
Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. *PLoS Biol* **5**(2): e14.
Pommie C, Levadoux S, Sabatier R, Lefranc G, Lefranc MP. 2004. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J Mol Recognit* **17**(1): 17-32.
Prilusky J, Bibi E. 2009. Studying membrane proteins through the eyes of the genetic code revealed a strong uracil bias in their coding mRNAs. *Proc Natl Acad Sci U S A* **106**(16): 6662-6666.
Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**(7457): 172-177.
Rossbach O, Hung LH, Khrameeva E, Schreiner S, Konig J, Curk T, Zupan B, Ule J, Gelfand MS, Bindereif A. 2014. Crosslinking-immunoprecipitation (iCLIP) analysis reveals global regulatory roles of hnRNP L. *RNA Biol* **11**(2): 146-155.
Savisaar R, Hurst LD. 2017a. Both Maintenance and Avoidance of RNA-Binding Protein Interactions Constrain Coding Sequence Evolution. *Mol Biol Evol* **34**(5): 1110-1126.
-. 2017b. Estimating the prevalence of functional exonic splice regulatory information. *Hum Genet* **136**(9): 1059-1078.
Shabalina SA, Spiridonov NA, Kashina A. 2013. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. *Nucleic acids research* **41**(4): 2073-2094.
Smithers B, Oates ME, Gough J. 2015. Splice junctions are constrained by protein disorder. *Nucleic acids research* **43**(10): 4814-4822.
Taylor FJ, Coates D. 1989. The code within the codons. *Biosystems* **22**(3): 177-187.

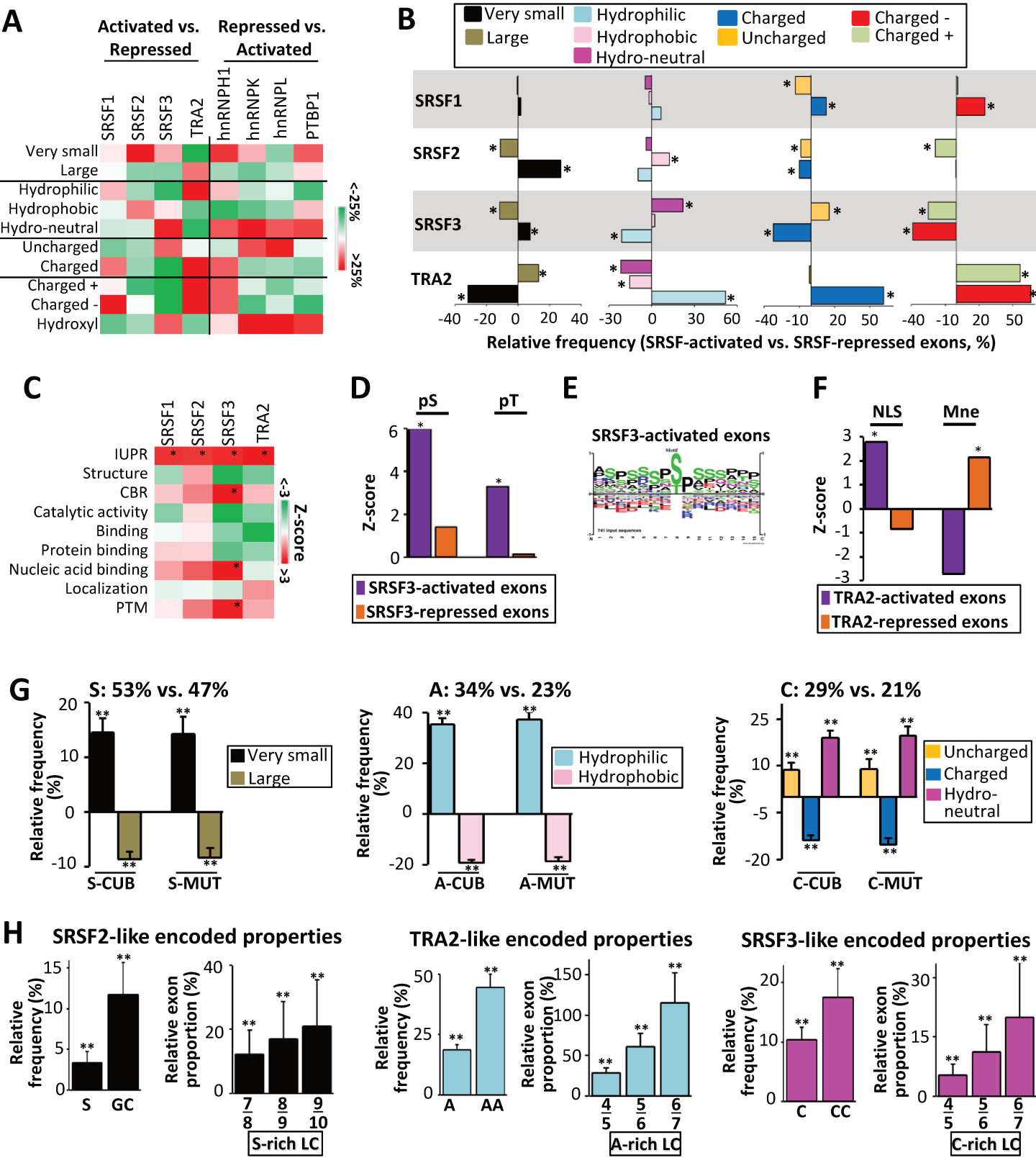
Thapar R. 2015. Structural basis for regulation of RNA-binding proteins by phosphorylation. *ACS chemical biology* **10**(3): 652-666.
Tranchevent LC, Aube F, Dulaurier L, Benoit-Pilven C, Rey A, Poret A, Chautard E, Mortada H, Desmet FO, Chakrama FZ et al. 2017. Identification of protein features encoded by alternative exons using Exon Ontology. *Genome research* **27**(6): 1087-1097.
Tsuda K, Someya T, Kuwasako K, Takahashi M, He F, Unzai S, Inoue M, Harada T, Watanabe S, Terada T et al. 2011. Structural basis for the dual RNA-recognition modes of human Tra2-beta RRM. *Nucleic acids research* **39**(4): 1538-1553.
Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**(7221): 470-476.
Wang X, Wang D, Zhao J, Qu M, Zhou X, He H, He R. 2006. The proline-rich domain and the microtubule binding domain of protein tau acting as RNA binding domains. *Protein and peptide letters* **13**(7): 679-685.
Warnecke T, Parmley JL, Hurst LD. 2008. Finding exonic islands in a sea of non-coding sequence: splicing related constraints on protein composition and evolution are common in intron-rich genomes. *Genome Biol* **9**(2): R29.
Woese CR. 1965. Order in the genetic code. *Proc Natl Acad Sci U S A* **54**(1): 71-75.
Wolfenden RV, Cullis PM, Southgate CC. 1979. Water, protein folding, and the genetic code. *Science* **206**(4418): 575-577.
Zhang C, Lee KY, Swanson MS, Darnell RB. 2013. Prediction of clustered RNA-binding protein motif sites in the mammalian genome. *Nucleic acids research* **41**(14): 6793-6807.
Zhang Z, Yu J. 2011. On the organizational dynamics of the genetic code. *Genomics Proteomics Bioinformatics* **9**(1-2): 21-29.


A
SRSF1
SRSF2
SRSF3
TRA2
GGAGGA
GAGGAG
AGAGGA
UGAGGA
CAGGAG
AGGAGG
AGGAGA
GAGGAU
GGAGCU
GCAGGA
GCUGCU
GAGCAG
CUGCUG
GAGGAG
GGAGGA
AGGAGG
UGCUGC
AGCUGC
UGCAGU
CAGCAG
CCUCCU
CAGCCC
CACCAC
CCCAGC
CCAGCC
UCCUCC
CAGCAG
CCCCAG
CCCCUC
CAGCCU
GAAGAA
AAGAAA
AGAAGA
UGAAGA
AGAAAA
AAGAAG
AAAGAA
GAAGAU
AUGAAG
AAAAGA
Cell line: 1 2 3 4 1 2 3 4
Cell line: 5 6 5 6
Cell line: 7 8 7 8
Cell line: 9 9
<-100%
>+100%
Activated exons
SRSFs
Repressed exons
B
Relative frequency (%)
G
S
C
A
SRSF1
SRSF2
SRSF3
TRA2
Activated exons
Repressed exons
Density
frequency
C
GG
GA
GC
GG
CC
CA
AA
GA
<-20%
>20%
<-25%
>25%
<-30%
>30%
<-40%
>40%
Exon proportion (%)
G-rich LC
S-rich LC
C-rich LC
A-rich LC
D
GGA
AGG
GGG
GAG
GCG
GCU
GGU
GGG
CCC
CCA
CCU
UCA
GAA
AGA
AAA
UUA
<-50%
>50%
<-80%
>80%
<-120%
G1-2
S1-2
C1-2
A1-2


A
SRSF1
SRSF2
SRSF3
TRA2
A
C
G
U
G: 41%
S: 67%
C: 45%
A: 67%
G: 23%
S:46%
C: 25%
A: 12%
B
SRSF1
SRSF2
SRSF3
TRA2
Gly (GGN)
Ala (GCN)
Pro (CCN)
Lys (AAR)
Relative frequency (%)
1 2 3 4 1 2 3 4 5 6 5 6 7 8 7 8 9 9
Activated exons
Repressed exons
C
SRSF1-activated exons
SRSF1-repressed exons
Gly: 0 1 2 >3
SRSF2-activated exons
SRSF2-repressed exons
Ala: 0 1 2 >3
SRSF3-activated exons
SRSF3-repressed exons
Pro: 0 1 2 >3
TRA2-activated exons
TRA2-repressed exons
Lys: 0 1 2 >3
Proportion of exons (%)


A
Small
Very small
S:62%
S:67%
Large
S:38%
A
G
C
U
B
Very small
Small
Large
Relative frequency (%)
30
15
0
-15
SRSF2-activated exons
SRSF2-repressed exons
C
Very small
Large
Very small vs. Large
Density
Frequency
Relative Frequency
SRSF2-activated exons
SRSF2-repressed exons
D
Hydrophilic
Hydro-neutral
Hydrophobic
44%
38%
42%
E
Hydro-philic
Hydro-neutral
Hydro-phobic
Relative frequency (%)
25
10
-5
-20
TRA2-activated exons
TRA2-repressed exons
F
TRA2-activated exons
TRA2
exon
TRA2-repressed exons
Hydrophobicity
G
Hydrophilic
Density
Frequency
TRA2-activated exons
TRA2-repressed exons
H
Uncharged
Hydroxyl
31%
33%
Charged
39%
I
Uncharged
Hydroxyl
Charged
Hydro-neutral
Relative frequency (%)
SRSF3-activated exons
SRSF3-repressed exons
J
Uncharged vs. charged
Hydroxyl vs. negatively charged
Density
Relative Frequency
SRSF3-activated exons
SRSF3-repressed exons


A
hnRNPH1-repressed exons
hnRNPK-repressed exons
Relative frequency (%)
G
C
B
hnRNPH1-repressed exons
hnRNPK-repressed exons
Relative frequency (%)
G1-2
C1-2
C
hnRNPH1-repressed exons
hnRNPK-repressed exons
Relative frequency (%)
Gly (GGN)
Pro (CCN)
D
Uncharged
Charged
Hydro-neutral
Relative frequency (%)
hnRNPK-repressed
E
CA
AC
CU
UC
His (CAY)
Gln (CAR)
Thr (ACN)
Leu (CUN)
Ser (UCN)
Relative frequency (%)
hnRNPL-repressed exons
PTBP1-repressed exons
F
hnRNPL-repressed exons
PTBP1-repressed exons
Relative frequency (%)
Hydroxyl
Negatively charged


A
Activated vs. Repressed
Repressed vs. Activated
SRSF1
SRSF2
SRSF3
TRA2
hnRNPH1
hnRNPK
hnRNPL
PTBP1
Very small
Large
Hydrophilic
Hydrophobic
Hydro-neutral
Uncharged
Charged
Charged +
Charged -
Hydroxyl
>25%
<25%
B
Very small
Large
Hydrophilic
Hydrophobic
Hydro-neutral
Charged
Uncharged
Charged -
Charged +
SRSF1
SRSF2
SRSF3
TRA2
Relative frequency (SRSF-activated vs. SRSF-repressed exons, %)
C
SRSF1
SRSF2
SRSF3
TRA2
IUPR
Structure
CBR
Catalytic activity
Binding
Protein binding
Nucleic acid binding
Localization
PTM
>3
<-3
Z-score
D
pS
pT
Z-score
SRSF3-activated exons
SRSF3-repressed exons
E
SRSF3-activated exons
F
NLS
Mne
Z-score
TRA2-activated exons
TRA2-repressed exons
G
S: 53% vs. 47%
Relative frequency (%)
Very small
Large
S-CUB
S-MUT
A: 34% vs. 23%
Relative frequency (%)
Hydrophilic
Hydrophobic
A-CUB
A-MUT
C: 29% vs. 21%
Relative frequency (%)
Uncharged
Charged
Hydro-neutral
C-CUB
C-MUT
H
SRSF2-like encoded properties
Relative frequency (%)
S
GC
Relative exon proportion (%)
7/8
8/9
9/10
S-rich LC
TRA2-like encoded properties
Relative frequency (%)
A
AA
Relative exon proportion (%)
4/5
5/6
6/7
A-rich LC
SRSF3-like encoded properties
Relative frequency (%)
C
CC
Relative exon proportion (%)
4/5
5/6
6/7
C-rich LC

Supplementary Table S1: GEO number of publicly available RNA-seq datasets and lists of splicing factor-regulated exons.

Supplementary Tables S2: Frequencies of hexanucleotides, nucleotides, dinucleotides, codons, codon nucleotide position, amino acids, and encoded physicochemical amino acid properties in splicing factor-regulated exons.

Supplementary Figure S1

A. Venn diagrams of exons whose inclusion is activated (left) or repressed (right) by SRSF1, SRSF2, and SRSF3 across different cell lines.

B. Venn diagrams of exons whose inclusion is activated (left) or repressed (right) by hnRNPK, hnRNPL, and PTBP1 across different cell lines.

C. Relative frequency (%) when compared to sets of control exons of the G nucleotide in SRSF3-activated and -repressed exons identified in HepG2 (1), GM19238 (2). (*) Randomization test FDR-adjusted P-value <0.05.

Supplementary Figure S2

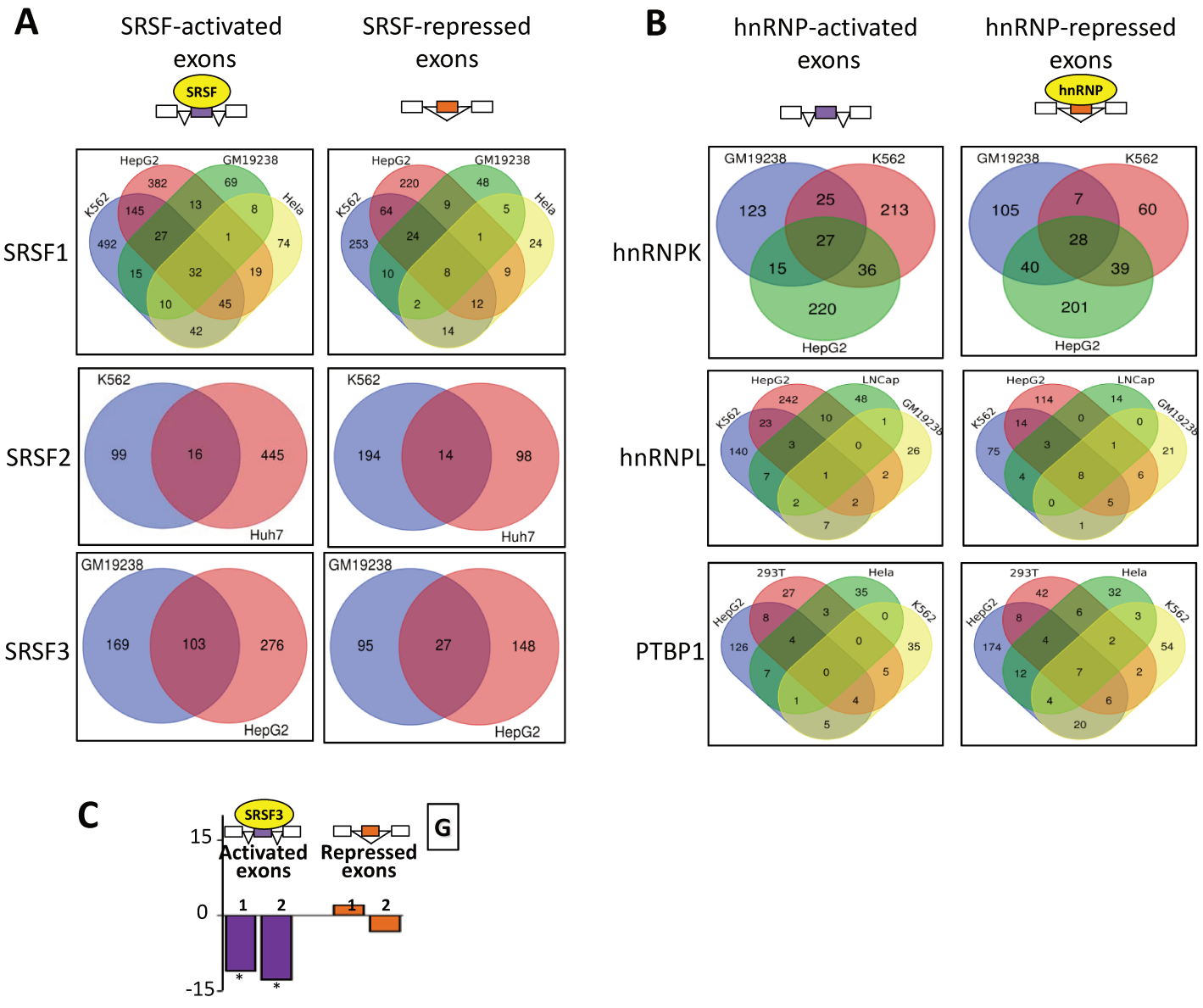
A. Color-code of the relative frequency, when compared to sets of control exons, of the codons of SRSF1-, SRSF2-, SRSF3-, and TRA2-activated or -repressed exons using datasets generated from different cell lines as indicated. The frequency of each codon was calculated in activated and repressed exons and expressed as the % of the average frequency calculated in 10 000 sets of control exons. Red and green colors indicate when the frequency is higher and lower, respectively, in the set of regulated exons compared to sets of control exons. The “activated vs repressed” column shows the relative frequency of each codon calculated in activated exons compared to the frequency calculated in repressed exons.

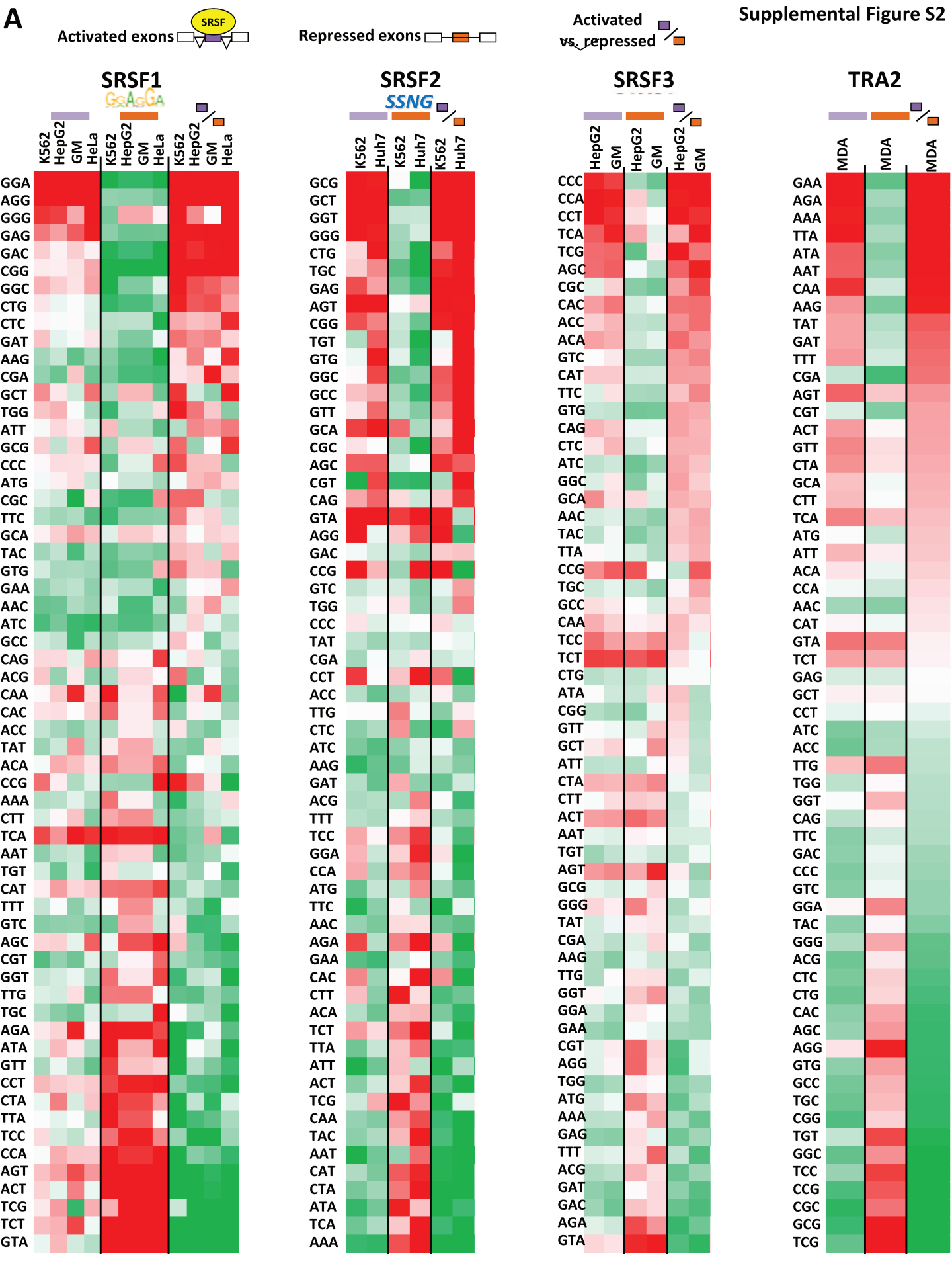
B. Same as above using exons regulated by hnRNPH1, hnRNPK, hnRNPL, or PTBP1. The “repressed vs activated” column shows the relative frequency of each codon calculated in repressed exons compared to the frequency calculated in activated exons.

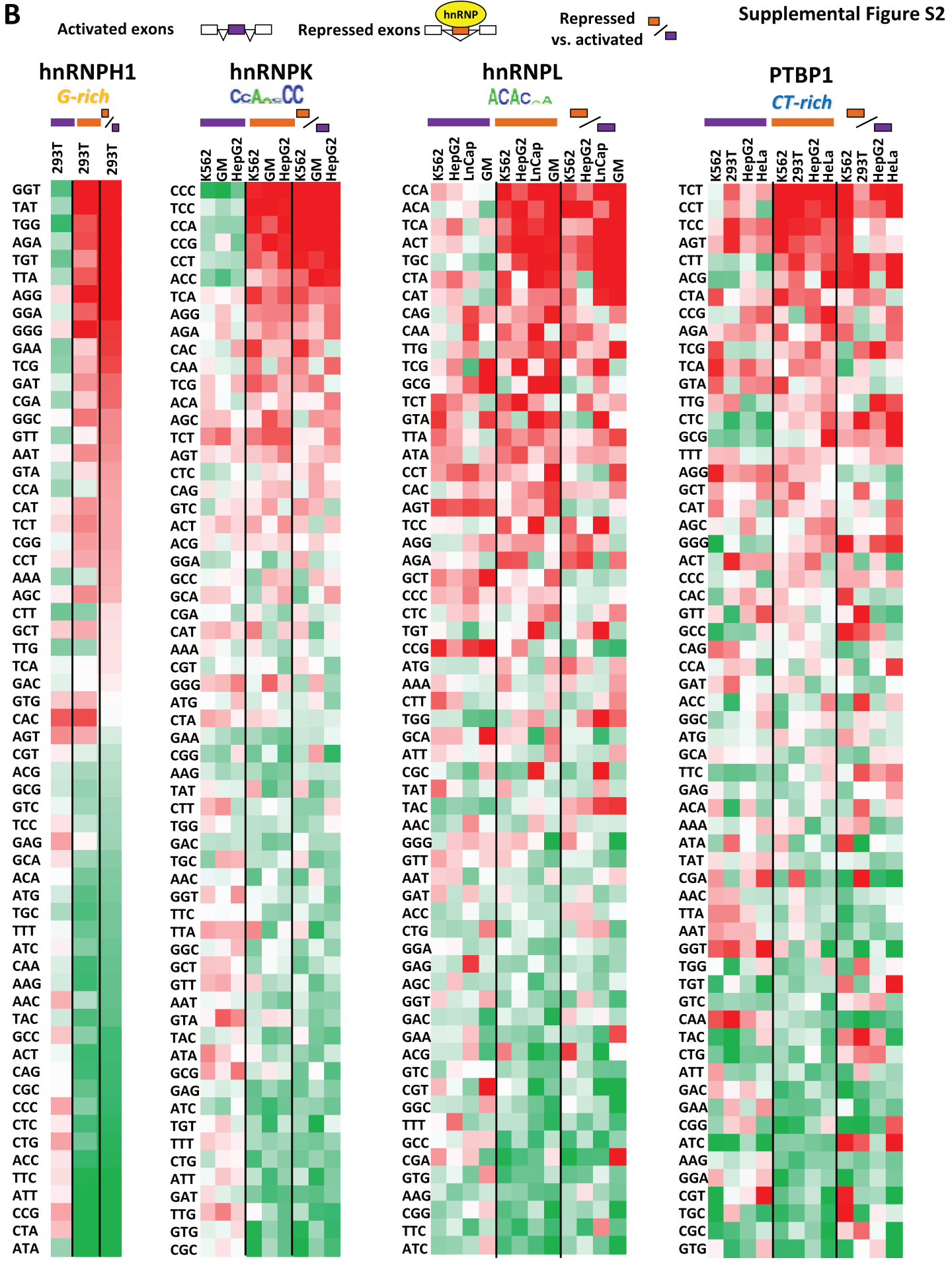
Supplementary Figure S3

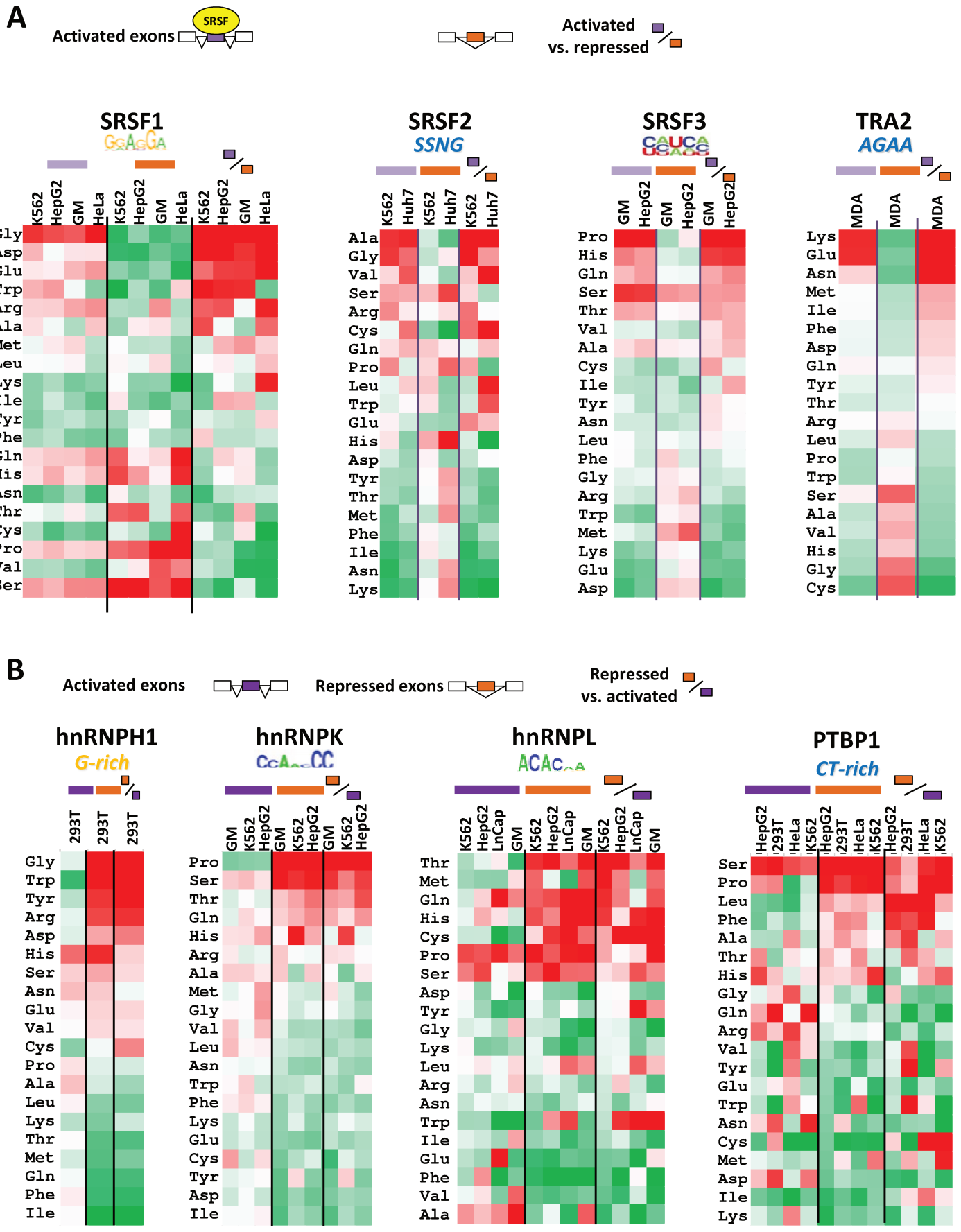
A. Color-code of the relative frequency, when compared to sets of control exons, of amino acids encoded by SRSF1-, SRSF2-, SRSF3-, and TRA2-activated or -repressed exons using datasets generated from different cell lines as indicated. The frequency of each amino acids was calculated in activated and repressed exons and expressed as the % of the average frequency calculated in 10 000 sets of control exons. Red and green colors indicate when the frequency is higher and lower, respectively, in the set of regulated exons compared to sets of control exons. The “activated vs. repressed” column shows the relative frequency of each amino acid calculated in activated exons compared to the frequency calculated in repressed exons.

B. Same as above using exons regulated by hnRNPH1, hnRNPK, hnRNPL, or PTBP1. The “repressed vs activated” column shows the relative frequency of each amino acid calculated in repressed exons compared to the frequency calculated in activated exons.


A
SRSF-activated exons
SRSF-repressed exons
SRSF
SRSF1
HepG2
GM19238
K562
Hela
382
69
13
145
8
492
27
1
74
15
32
19
10
45
42
220
48
64
9
5
253
24
1
24
10
8
9
2
12
14
SRSF2
K562
99
16
445
Huh7
194
14
98
SRSF3
GM19238
169
103
276
HepG2
95
27
148
B
hnRNP-activated exons
hnRNP-repressed exons
hnRNP
hnRNPK
GM19238
K562
123
25
213
27
15
36
220
HepG2
105
7
60
28
40
39
201
hnRNPL
HepG2
LNCap
K562
GM19238
242
10
48
23
1
140
3
0
26
7
1
2
2
2
7
114
14
0
14
0
75
3
1
21
4
8
6
0
5
1
PTBP1
293T
Hela
HepG2
K562
27
35
8
3
0
126
4
0
35
7
0
5
1
4
5
42
32
6
8
3
174
4
2
54
12
7
2
4
6
20
C
SRSF3
G
15
0
-15
Activated exons
Repressed exons
1
2
1
2
*
*


A
Activated exons
SRSF
Repressed exons
Activated vs. repressed
SRSF1
SRSF2
SRSF3
TRA2
K562
HepG2
GM
HeLa
Huh7
MDA


B
Activated exons
Repressed exons
hnRNP
Repressed
vs. activated
Supplemental Figure S2
hnRNPH1
G-rich
hnRNPK
hnRNPL
PTBP1
CT-rich
293T
K562
GM
HepG2
LnCap
HeLa
GGT TAT TGG AGA TGT TTA AGG GGA GGG GAA TCG GAT CGA GGC GTT AAT GTA CCA CAT TCT CGG CCT AAA AGC CTT GCT TTG TCA GAC GTG CAC AGT CGT ACG GCG GTC TCC GAG GCA ACA ATG TGC TTT ATC CAA AAG AAC TAC GCC ACT CAG CGC CCC CTC CTG ACC TTC ATT CCG CTA ATA
CCC TCC CCA CCG CCT ACC TCA AGG AGA CAC CAA TCG ACA AGC TCT AGT CTC CAG GTC ACT ACG GGA GCC GCA CGA CAT AAA CGT GGG ATG CTA GAA CGG AAG TAT CTT TGG GAC TGC AAC GGT TTC TTA GGC GCT GTT AAT GTA TAC ATA GCG GAG ATC TGT TTT CTG ATT GAT TTG GTG CGC
CCA ACA TCA ACT TGC CTA CAT CAG CAA TTG TCG GCG TCT GTA TTA ATA CCT CAC AGT TCC AGG AGA GCT CCC CTC TGT CCG ATG AAA CTT TGG GCA ATT CGC TAT TAC AAC GGG GTT AAT GAT ACC CTG GGA GAG AGC GGT GAC GAA ACG GTC CGT GGC TTT GCC CGA GTG AAG CGG TTC ATC
TCT CCT TCC AGT CTT ACG CTA CCG AGA TCG TCA GTA TTG CTC GCG TTT AGG GCT CAT AGC GGG ACT CCC CAC GTT GCC CAG CCA GAT ACC GGC ATG GCA TTC GAG ACA AAA ATA TAT CGA AAC TTA AAT GGT TGG TGT GTC CAA TAC CTG ATT GAC GAA CGG ATC AAG GGA CGT TGC CGC GTG


A
Activated exons
SRSF
Activated
vs. repressed
SRSF1
SRSF2
SSNG
SRSF3
TRA2
AGAA
K562
HepG2
GM
HeLa
Huh7
MDA
B
Activated exons
Repressed exons
Repressed
vs. activated
hnRNPH1
G-rich
hnRNPK
hnRNPL
PTBP1
CT-rich
293T
LnCap

Conclusion de l’Article # 2

Dans le travail présenté ci-dessus, nous avons analysé la composition nucléotidique d’exons co-régulés par les mêmes facteurs d’épissage ainsi que la composition en acides aminés des peptides codés par ces exons. Nous avons aussi analysé la nature et les propriétés physico-chimiques des acides aminés codés par ces exons. Sur la base de ces analyses, nous faisons état d'un lien direct entre le processus de régulation de l’épissage et les conséquences fonctionnelles de cette régulation. Ce lien repose sur deux principes simples : 1) les facteurs d'épissage se lient à des séquences exoniques qui ont un biais de composition nucléotidique, ce qui a des conséquences sur la composition nucléotidique des codons au sein des exons et ce à toutes les positions nucléotidiques des codons ; et 2) les codons ayant le même biais de composition nucléotidique correspondent à des acides aminés ayant des propriétés physico-chimiques similaires. Ceci a des conséquences sur des caractéristiques protéiques codées par les exons co-régulés par un même facteur d'épissage. Cela signifie que des exons co-régulés par un facteur d’épissage codent pour des régions protéiques ayant les mêmes propriétés.

En conclusion, la régulation de l'épissage d'un exon par un facteur d'épissage qui repose sur l'affinité de ce facteur pour un ou plusieurs nucléotides spécifiques est étroitement liée aux propriétés physico-chimiques codées en raison du caractère non aléatoire du code génétique. Nos travaux permettent donc de comprendre comment un phénomène complexe (la relation entre régulation de l'épissage et ses conséquences biologiques) peut s'appuyer sur des principes simples.

Discussion

Lorsque j'ai commencé ma thèse, notre objectif était de rechercher des propriétés et des caractéristiques communes à des groupes d'exons co-régulés par un même facteur d'épissage. Pour cela, j'ai développé différentes approches bio-informatiques permettant d'analyser, de visualiser et d'interpréter les résultats obtenus. Nous n'avions aucune idée préconçue relativement à la composition nucléotidique des exons et des introns. Dans ce contexte, l'observation de biais de composition nucléotidique partagés par des exons co-régulés est très certainement l'un des résultats majeurs de ma thèse. L'objectif de cette discussion générale est de tenter d'intégrer ces observations dans un contexte beaucoup plus général. Pour cela, je discuterai dans un premier temps de la relation entre l'organisation 3D du génome et la régulation de l'épissage. Je proposerai un modèle selon lequel, ces deux phénomènes bien que se réalisant à des échelles différentes (de centaines de milliers à une centaine de nucléotides) sont reliés directement par le biais de composition nucléotidique. Je discuterai dans un deuxième temps de la relation entre les biais de composition nucléotidique, contribuant à la régulation de l'expression des gènes comme l'épissage et les biais de séquence en termes d'acides aminés contribuant à la fonction biologique des gènes.

1. Biais de composition nucléotidique : de l'organisation du génome à la régulation de l'épissage.

Un acide nucléique est un polymère de nucléotides. L'agencement de ces nucléotides dans la séquence nucléique détermine les propriétés physico-chimiques du polymère. Ainsi, selon la nature des bases constituant un ADN double brin, les deux brins seront liés l'un à l'autre avec plus ou moins de

force. En effet, les deux brins sont liés par la complémentarité de base, ou « base pairing ». Or, les bases G et C établissent trois liaisons hydrogènes et les bases A et T établissent deux liaisons hydrogènes seulement (Fig.22A). Par conséquent, les interactions entre bases contribuent à la stabilité du double brin d'ADN. Par ailleurs, les atomes composant deux bases successives d'un même brin d'ADN peuvent établir des contacts, ce qui définit la notion de « base stacking » . La nature de ces interactions détermine également certaines propriétés du polymère d'ADN, comme sa flexibilité (Fig.22B). Par exemple, un « empilement » de nucléotides GC autorise une plus grande flexibilité de courbure qu'un empilement de paires AT (Segal and Widom, 2009; Trifonov, 2011). Si l'on considère une région génomique de plusieurs milliers de paires de bases, la conformation que ce polymère adoptera dépendra donc en partie de sa composition nucléotidique qui détermine ces propriétés physiques. Par ailleurs, la composition nucléotidique interfère sur la nature des interactions ADN-protéines. En particulier, la disposition des nucléosomes est influencée par la capacité du segment d'ADN à se courber (Segal and Widom, 2009; Trifonov, 2011). Les séquences riches en GC étant flexibles, elles autorisent plus souvent la formation d'un nucléosome alors que les séquences riches en AT, notamment les poly-dT et les poly-dA constituent des barrières énergétiques à la formation d'un nucléosome (Segal and Widom, 2009; Trifonov, 2011). La présence et le positionnement des

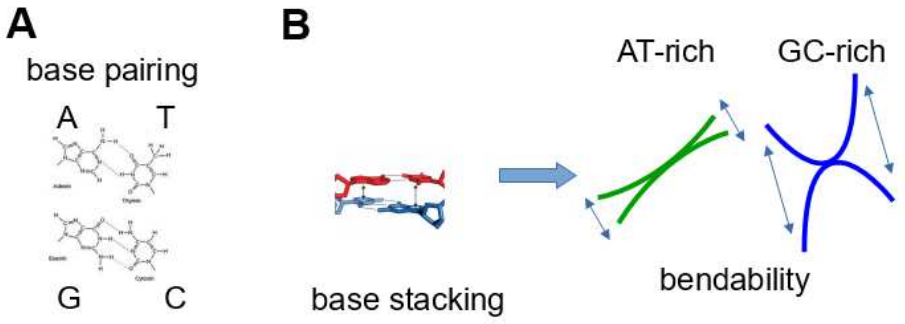


Figure 22 : A. Le "base pairing" est l'appariement des base, adénine (A) avec thymine (T) par deux liaisons hydrogène, guanine (G) avec cytosine (C) par trois liaisons hydrogènes. B. L'empilement de ces paires ("base stacking" de AT et GC) définit la capacité d'un fragment l'ADN à se courber ("bendability").

nucléosomes sont donc, au moins en partie, déterminés par la composition en GC de l'ADN. Les propriétés physiques de l'ADN (comme sa flexibilité) (Vinogradov, 2003) ainsi que la disposition des nucléosomes sur le polymère détermine l'organisation chromatinienne.

La composition nucléotidique a également un impact sur la résistance de l'ADN à des contraintes physiques (Dans et al., 2014; Naughton et al., 2013; Reymer et al., 2018; Shin et al., 2016; Vinogradov, 2003). Par exemple, il a été proposé que des régions riches en GC sont plus résistantes aux contraintes liées à la transcription. Ceci serait la conséquence du fait que l'ADN riche en GC est plus flexible et plus « polymorphe » (Dans et al., 2014; Shin et al., 2016; Vinogradov, 2003). En effet, les deux brins d'ADN, dans leur association, prennent spontanément une conformation en double hélice qui varie dans sa forme selon le taux de GC de la séquence nucléique. Des régions riches en GC peuvent passer plus facilement de la forme B de l'ADN, qui représente une forme structurelle « régulière » et très répandue, à une forme structurelle plus « irrégulière » ou forme Z (Ghosh and Bansal, 2003; Shin et al., 2016; Vinogradov, 2003). De façon intéressante, il a été montré que la forme « Z-DNA » peut « absorber » (en termes énergétique) plus de contraintes physiques, comme les sur-enroulements négatifs et positifs, générés lors de la transcription (Shin et al., 2016; Vinogradov, 2003). Dans ce contexte, nous observons que les gènes riches en GC ont une densité de RNAPII plus importante. Par ailleurs, il a été montré que des régions génomiques (de dizaines de paires de base à quelques centaines de milliers) riches en GC sont composées de nombreux petits gènes eux-mêmes riches en GC et étant fortement transcrits (Frousios et al., 2013; Gul et al., 2018; International Human Genome Sequencing Consortium, 2001; Rogier Versteeg et al., 2003). A l'inverse, des régions génomiques riches en AT sont composées de grands gènes eux-mêmes riches en AT et étant généralement faiblement transcrits. Pour conclure, les propriétés physiques d'un polymère d'ADN (par exemple, sa flexibilité) dépendent de sa composition nucléotidique et donc des biais de composition (par exemple fort taux de GC sur une région génomique donnée). Ceci est équivalent au fait que les propriétés

physico-chimiques des protéines, constituant une autre famille de polymère sont déterminées par la nature des acides aminés qui les composent.

Or, dans le cas de l'ADN, il est devenu évident que les propriétés physiques de ce polymère jouent un rôle majeur dans son organisation au sein des cellules et que l'organisation en 3D de l'ADN a des conséquences sur l'expression des gènes (Naughton et al., 2013; Nguyen and Bosco, 2015; Tsochatzidou et al., 2017). Par conséquent, l'organisation de la chromatine et les activités liées à l'expression des gènes impliquent des propriétés physiques qui reposent directement ou indirectement sur la composition nucléotidique de l'ADN (Berná et al., 2012).

Dans ce contexte, il est important de souligner que les biais de composition nucléotidique au sein des génomes sont homogènes sur de très grandes distances génomiques. L'observation de cette régularité a conduit à la définition des isochores (Bernardi, 2015). Les isochores correspondent à des régions génomiques ayant un contenu en GC homogène sur toute leur longueur ; composition qui diffère de celle des régions qui l'encadrent (Fig.23). Cette notion de biais de composition homogène sur de grandes régions génomiques (de quelques dizaines à plusieurs centaines de milliers de paires de bases) est très importante. En effet, cette homogénéité signifie qu'un biais de composition nucléotidique sur une très grande région (des dizaines de milliers de paires de bases) se retrouve aussi à plus petite échelle dans des sous-régions (une centaine de paires de bases) (Fig.23). Ces notions sont en accord avec nos observations montrant que la composition nucléotidique des exons corrèle avec celles des gènes auxquels ils appartiennent. Ainsi des exons riches en GC appartiennent à des gènes également riches en GC.

Sur la base de ces observations, nous proposons que des contraintes physiques liées à l'organisation du génome et/ou à la transcription imposent des biais de séquences sur de grandes régions génomiques. Ceci implique des biais de composition nucléotidique au sein de ces grandes régions, donc des gènes qui composent ces régions et donc des exons qui composent ces gènes. Dans ce

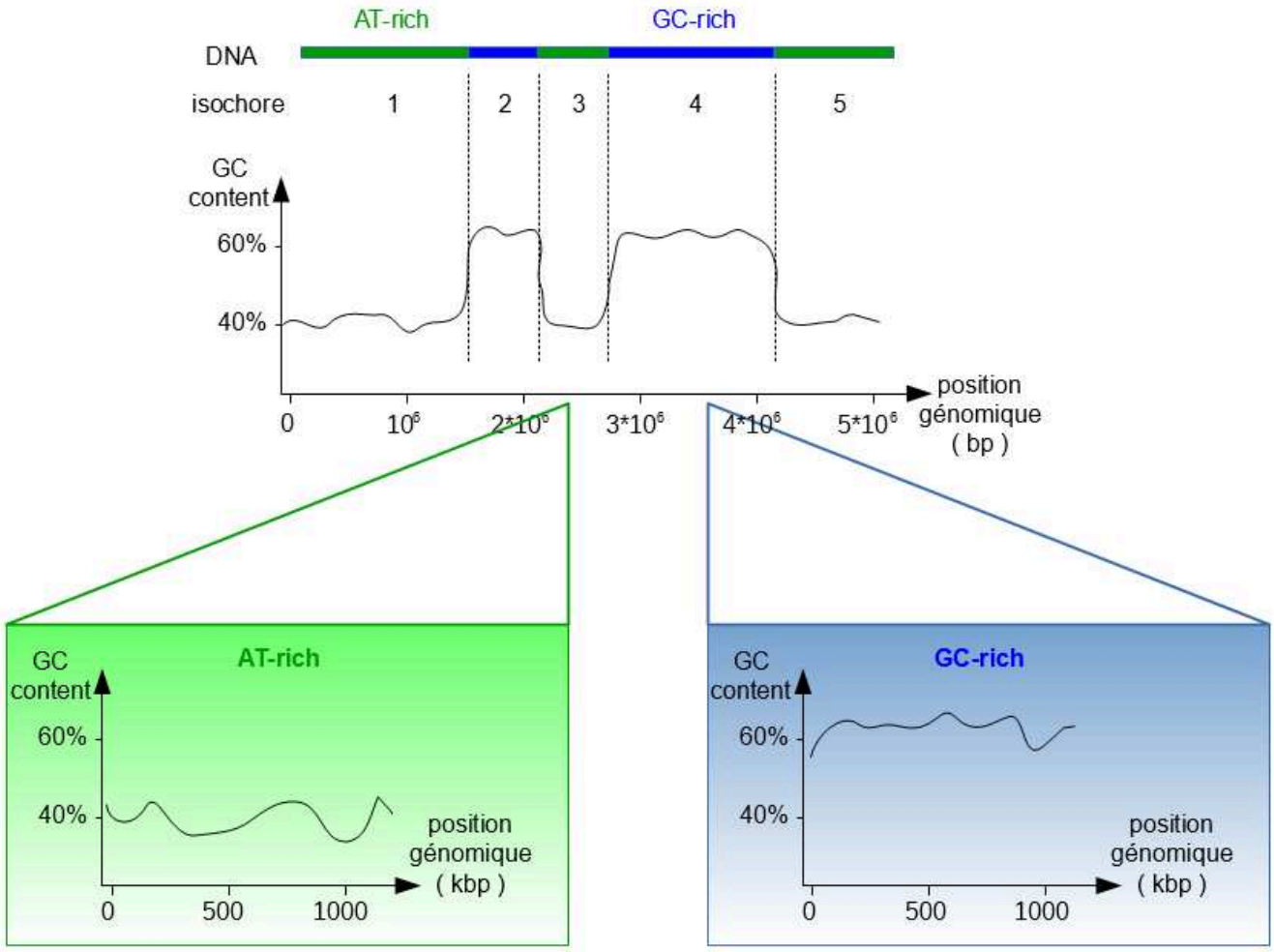


Figure 23 : Haut. Les isochores sont définis sur l'homogénéité du contenu en GC sur une grande région de l'ADN (centaines de milliers de bp). Deux types d'isochores existent dans le génome humain, les isochores riches en AT (vert) et ceux riches en GC (bleu). Bas. Le contenu en GC d'un isochore est respecté à plus petite échelle (centaines de bp).

contexte, nous montrons, que des biais de composition nucléotidique sont associés localement à des mécanismes spécifiques de l'organisation chromatinienne et de l'épissage. Ainsi, le biais de GC à l'échelle des introns et des exons est fortement lié à la localisation des nucléosomes dans les gènes. En effet, les nucléosomes sont mieux positionnés au niveau des exons quand ceux-ci sont localisés dans des régions riches en AT. Cela pourrait être en partie dû au fait que de longs poly-dT sont présents en amont de ces exons et forment une barrière énergétique à la présence d'un nucléosome (Schwartz et al.,

2009; Segal and Widom, 2009; Tanaka et al., 2010). Un nucléosome positionné sur un exon pourrait constituer un obstacle à la RNAPII (Teves et al., 2014), ce qui pourrait renforcer la reconnaissance de l'exon au cours de l'épissage. Par ailleurs, si un exon est localisé dans une région riche en AT, nous montrons que la région intronique en amont de l'exon est aussi enrichie en adénines et en thymines, ce qui augmente la probabilité de fréquences de sites ressemblant à des BPs et/ou à des sites de fixation de SF1 ou U2AF65 (Corvelo et al., 2010). Nous pensons que cette abondance de sites potentiels (ou leurres) pourrait nuire à l'efficacité de détection du BP et donc de l'épissage. Dans ce contexte, une hypothèse est que des facteurs d'épissage reconnaissant des séquences riches en AT pourraient se fixer sur les « leurres » et aider au « bon » positionnement de U2AF65 et SF1 et donc à définir plus efficacement l'exon en aval (Howard et al., 2018; Pineda and Bradley, 2018; Tavanez et al., 2012; Wu and Maniatis, 1993).

Dans le cas d'exons riches en GC présents dans des régions riches en GC, les nucléosomes pourraient être moins spécifiquement positionnés sur les exons. Néanmoins, le fait que la RNAPII est plus lente dans des régions riches en GC pourrait aider à synchroniser la transcription et l'épissage. Cependant, la richesse en GC augmenterait la probabilité de former des structures secondaires ARN au niveau du 5'SS, diminuant la probabilité d'hybridation entre ce site et U1 snRNA. Des facteurs d'épissage se fixant sur des régions riches en GC pourraient aider à ce processus et donc augmenter la reconnaissance des exons.

Collectivement, ces observations suggèrent que le biais de composition nucléotidique permet de faire un lien direct entre l'organisation en 3D de l'ADN qui a des conséquences sur la régulation de l'expression des gènes et l'épissage, qui a des conséquences sur la sélection des exons au cours de la transcription.

Dans ce contexte, il est également intéressant de souligner qu'une étude récente montre une relation entre le biais de composition nucléotidique (taux de GC) et des mécanismes de dégradation et

de traduction des ARNm (Courel et al., 2018). Par exemple, les ARNm pauvres en GC sont plus fréquemment associés aux « corps P » (P-bodies) concentrant des facteurs protéiques particuliers associés à la régulation de la dégradation et de la traduction des ARNm. Les ARNm riches en GC seraient moins fréquemment associés à ces « corps P », et leur dégradation et traduction dépendraient d'autres facteurs protéiques.

2. Relation entre régulation et conséquences fonctionnelles

Dans la première partie de cette discussion, je me suis attaché à montrer que l'analyse des biais en composition nucléotidique permettait de comprendre les propriétés physiques des polymères que sont l'ADN avec des conséquences importantes sur les phénomènes de régulation à plus petite échelle comme la transcription et l'épissage.

De la même manière, les protéines sont des polymères dont les propriétés physico-chimiques dépendent de leur composition en acides aminés. Il est clairement établi que les fonctions biologiques des protéines dépendent des propriétés physico-chimiques des acides aminés qui composent ces protéines. Par exemple, la composition en acides aminés hydrophobes a des conséquences sur la localisation membranaire des protéines (Engelman et al., 1986). Or, la composition en acides aminés des protéines dépend à son tour de la composition nucléotidique des gènes qui codent pour ces protéines. En effet, le code génétique n'est pas aléatoire puisque des acides aminés qui partagent des propriétés physicochimiques correspondent à des codons qui ont une composition nucléotidique similaire. Ces observations suggèrent donc qu'il existe un lien entre la régulation de l'expression des gènes, dépendante de leur composition nucléotidique, et les fonctions biologiques de leurs produits, dépendantes de leur composition en acides aminés. Supportant ce modèle, des études ont établis que les gènes se trouvant à proximité dans l'espace du noyau sont co-régulés (Nguyen and Bosco, 2015;

Szczepińska and Pawłowski, 2013; Tsochatzidou et al., 2017) et les protéines codées par ces gènes participent aux même fonctions biologiques (Karathia et al., 2016; Paz et al., 2014). Par ailleurs, des gènes dont les produits partagent les mêmes fonctions biologiques ont une composition nucléotidique similaire (Berná et al., 2012; Paz et al., 2014).

Dans ce contexte, nos travaux montrent également qu'il existe une relation directe entre régulation à l'échelle d'un exon (l'épissage) et la fonction biologique du domaine protéique codée par cet exon grâce aux biais de composition nucléotidiques. En effet, le biais de composition nucléotidique contribue à la fois à la reconnaissance de cet exon par un facteur d'épissage et à la nature des acides aminés qu'il code.

En conclusion, j'ai développé au cours de ma thèse un certain nombre « d'outils informatiques », qui m'ont permis d'explorer la relation entre biais de composition nucléotidique des gènes et épissage. Comme les biais de composition nucléotidiques sont aussi associés à l'organisation 3D du génome, mes observations suggèrent que le biais de composition nucléotidique établit un lien entre des phénomènes biologiques portant sur des centaines de milliers de bases comme l'organisation du génome et ceux portant sur quelques centaines de bases, à savoir l'épissage.

Bibliographie

Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C., and Komorowski, J. (2009). Nucleosomes are well positioned in exons and carry characteristic histone modifications. Genome Res. *19*, 1732–1741.

Änkö, M.-L. (2014). Regulation of gene expression programmes by serine–arginine rich splicing factors. Seminars in Cell & Developmental Biology *32*, 11–21.

Auboeuf, D., Hönig, A., Berget, S.M., and O'Malley, B.W. (2002). Coordinate Regulation of Transcription and Splicing by Steroid Receptor Coregulators. Science *298*, 416–419.

Ast, G. (2004). How did alternative splicing evolve? Nature Reviews Genetics *5*, 773–782.

Bai, L., and Morozov, A.V. (2010). Gene regulation by nucleosome positioning. Trends in Genetics *26*, 476–483.

Baralle, F.E., and Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. Nature Reviews Molecular Cell Biology *18*, 437–451.

Baralle, M., Skoko, N., Knezevich, A., Conti, L.D., Motti, D., Bhuvanagiri, M., Baralle, D., Buratti, E., and Baralle, F.E. (2006). NF1 mRNA biogenesis: Effect of the genomic milieu in splicing regulation of the NF1 exon 37 region. FEBS Letters *580*, 4449–4456.

Baraniak, A.P., Lasda, E.L., Wagner, E.J., and Garcia-Blanco, M.A. (2003). A Stem Structure in Fibroblast Growth Factor Receptor 2 Transcripts Mediates Cell-Type-Specific Splicing by Approximating Intronic Control Elements. Molecular and Cellular Biology *23*, 9327–9337.

Bechara, E.G., Sebestyén, E., Bernardis, I., Eyras, E., and Valcárcel, J. (2013). RBM5, 6, and 10 Differentially Regulate NUMB Alternative Splicing to Control Cancer Cell Proliferation. Molecular Cell *52*, 720–733.

Beck, S., Rhee, C., Song, J., Lee, B.-K., LeBlanc, L., Cannon, L., and Kim, J. (2018). Implications of CpG islands on chromosomal architectures and modes of global gene regulation. Nucleic Acids Res *46*, 4382–4391.

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

Berk, A.J. (2016). Discovery of RNA splicing and genes in pieces. PNAS *113*, 801–805.

Berná, L., Chaurasia, A., Angelini, C., Federico, C., Saccone, S., and D'Onofrio, G. (2012). The footprint of metabolism in the organization of mammalian genomes. BMC Genomics *13*, 174.

Bernardi, G. (2015). Genome Organization and Chromosome Architecture. Cold Spring Harb Symp Quant Biol *80*, 83–91.

Bernardi, G. (2018). The formation of chromatin domains involves a primary step based on the 3-D structure of DNA. Scientific Reports *8*, 17821.

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

Boutz, P.L., Chawla, G., Stoilov, P., and Black, D.L. (2007). MicroRNAs regulate the expression of the alternative splicing factor nPTB during muscle development. Genes Dev. *21*, 71–84.

Buratti, E., and Baralle, F.E. (2004). Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process. Molecular and Cellular Biology *24*, 10505–10514.

Cáceres, J.F., and Kornblihtt, A.R. (2002). Alternative splicing: multiple control mechanisms and involvement in human disease. Trends in Genetics *18*, 186–193.

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

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

Corvelo, A., Hallegger, M., Smith, C.W.J., and Eyras, E. (2010). Genome-Wide Association between Branch Point Properties and Alternative Splicing. PLOS Computational Biology *6*, e1001016.

Costantini, M., Cammarano, R., and Bernardi, G. (2009). The evolution of isochore patterns in vertebrate genomes. BMC Genomics *10*, 146.

Courel, M., Clement, Y., Foretek, D., Vidal, O., Yi, Z., Kress, M., Vindry, C., Benard, M., Bossevain, C., Antoniewski, C., et al. (2018). GC content shapes mRNA decay and storage in human cells. BioRxiv 373498.

Cozzi, P., Milanesi, L., and Bernardi, G. (2015). Segmenting the Human Genome into Isochores. Evol Bioinform Online *11*, EBO.S27693.

Cramer, P., Cáceres, J.F., Cazalla, D., Kadener, S., Muro, A.F., Baralle, F.E., and Kornblihtt, A.R. (1999). Coupling of Transcription with Alternative Splicing: RNA Pol II Promoters Modulate SF2/ASF and 9G8 Effects on an Exonic Splicing Enhancer. Molecular Cell *4*, 251–258.

Dans, P.D., Faustino, I., Battistini, F., Zakrzewska, K., Lavery, R., and Orozco, M. (2014). Unraveling the sequence-dependent polymorphic behavior of d(CpG) steps in B-DNA. Nucleic Acids Res *42*, 11304–11320.

Dardenne, E., Polay Espinoza, M., Fattet, L., Germann, S., Lambert, M.-P., Neil, H., Zonta, E., Mortada, H., Gratadou, L., Deygas, M., et al. (2014). RNA Helicases DDX5 and DDX17 Dynamically Orchestrate Transcription, miRNA, and Splicing Programs in Cell Differentiation. Cell Reports *7*, 1900–1913.

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

Davie, J.R., Xu, W., and Delcuve, G.P. (2015). Histone H3K4 trimethylation: dynamic interplay with pre-mRNA splicing. Biochem. Cell Biol. *94*, 1–11.

Engelman, D.M., Steitz, T.A., and Goldman, A. (1986). Identifying Nonpolar Transbilayer Helices in Amino Acid Sequences of Membrane Proteins. Annual Review of Biophysics and Biophysical Chemistry *15*, 321–353.

Expert-Bezançon, A., Caer, J.P.L., and Marie, J. (2002). Heterogeneous Nuclear Ribonucleoprotein (hnRNP) K Is a Component of an Intronic Splicing Enhancer Complex That Activates the Splicing of the Alternative Exon 6A from Chicken β-Tropomyosin Pre-mRNA. J. Biol. Chem. *277*, 16614–16623.

Fiszbein, A., Giono, L.E., Quaglino, A., Berardino, B.G., Sigaut, L., von Bilderling, C., Schor, I.E., Steinberg, J.H.E., Rossi, M., Pietrasanta, L.I., et al. (2016). Alternative Splicing of G9a Regulates Neuronal Differentiation. Cell Reports *0*.

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

Frousios, K., Iliopoulos, C.S., Tischler, G., Kossida, S., Pissis, S.P., and Arhondakis, S. (2013). Transcriptome map of mouse isochores in embryonic and neonatal cortex. Genomics *101*, 120–124.

Furlanis, E., and Scheiffele, P. (2018). Regulation of Neuronal Differentiation, Function, and Plasticity by Alternative Splicing. Annual Review of Cell and Developmental Biology *34*, 451–469.

Gao, K., Masuda, A., Matsuura, T., and Ohno, K. (2008). Human branch point consensus sequence is yUnAy. Nucleic Acids Res *36*, 2257–2267.

Gaykalova, D.A., Kulaeva, O.I., Volokh, O., Shaytan, A.K., Hsieh, F.-K., Kirpichnikov, M.P., Sokolova, O.S., and Studitsky, V.M. (2015). Structural analysis of nucleosomal barrier to transcription. PNAS *112*, E5787–E5795.

Geuens, T., Bouhy, D., and Timmerman, V. (2016). The hnRNP family: insights into their role in health and disease. Hum Genet *135*, 851–867.

Ghosh, A., and Bansal, M. (2003). A glossary of DNA structures from A to Z. Acta Cryst D *59*, 620–626.

Gilbert, W.V., Bell, T.A., and Schaening, C. (2016). Messenger RNA modifications: Form, distribution, and function. Science *352*, 1408–1412.

Goina, E., Skoko, N., and Pagani, F. (2008). Binding of DAZAP1 and hnRNPA1/A2 to an Exonic Splicing Silencer in a Natural BRCA1 Exon 18 Mutant. Molecular and Cellular Biology *28*, 3850–3860.

Graveley, B.R. (2000). Sorting out the complexity of SR protein functions. RNA *6*, 1197–1211.

Graveley, B.R. (2005). Mutually Exclusive Splicing of the Insect Dscam Pre-mRNA Directed by Competing Intronic RNA Secondary Structures. Cell *123*, 65–73.

Gul, I.S., Staal, J., Hulpiau, P., De Keuckelaere, E., Kamm, K., Deroo, T., Sanders, E., Staes, K., Driege, Y., Saeys, Y., et al. (2018). GC Content of Early Metazoan Genes and Its Impact on Gene Expression Levels in Mammalian Cell Lines. Genome Biol Evol *10*, 909–917.

Gurung, S.P., Schwarz, C., Hall, J.P., Cardin, C.J., and Brazier, J.A. (2015). The importance of loop length on the stability of i-motif structures †Electronic supplementary information (ESI) available: Experimental details, UV melting curves at 295 nm, and CD spectra at pH 5 and 8. See DOI: 10.1039/c4cc07279k Click here for additional data file. Chem Commun (Camb) *51*, 5630–5632.

Han, S.P., Tang, Y.H., and Smith, R. (2010). Functional diversity of the hnRNPs: past, present and perspectives. Biochemical Journal *430*, 379–392.

Hang, J., Wan, R., Yan, C., and Shi, Y. (2015). Structural basis of pre-mRNA splicing. Science *349*, 1191–1198.

Hatje, K., Rahman, R.-U., Vidal, R.O., Simm, D., Hammesfahr, B., Bansal, V., Rajput, A., Mickael, M.E., Sun, T., Bonn, S., et al. (2017). The landscape of human mutually exclusive splicing. Molecular Systems Biology *13*, 959.

Hertel, K.J. (2008). Combinatorial Control of Exon Recognition. J. Biol. Chem. *283*, 1211–1215.

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

Howard, J.M., Lin, H., Wallace, A.J., Kim, G., Draper, J.M., Haeussler, M., Katzman, S., Toloue, M., Liu, Y., and Sanford, J.R. (2018). HNRNPA1 promotes recognition of splice site decoys by U2AF2 in vivo. Genome Res. *28*, 689–698.

Hua, Y., Vickers, T.A., Okunola, H.L., Bennett, C.F., and Krainer, A.R. (2008). Antisense Masking of an hnRNP A1/A2 Intronic Splicing Silencer Corrects SMN2 Splicing in Transgenic Mice. The American Journal of Human Genetics *82*, 834–848.

Huang, H., Zhang, J., Harvey, S.E., Hu, X., and Cheng, C. (2017). RNA G-quadruplex secondary structure promotes alternative splicing via the RNA-binding protein hnRNPF. Genes Dev. *31*, 2296–2309.

Huppert, J.L., and Balasubramanian, S. (2005). Prevalence of quadruplexes in the human genome. Nucleic Acids Res *33*, 2908–2916.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

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

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

Kadener, S., Cramer, P., Nogués, G., Cazalla, D., Mata, M. de la, Fededa, J.P., Werbajh, S.E., Srebrow, A., and Kornblihtt, A.R. (2001). Antagonistic effects of T-Ag and VP16 reveal a role for RNA pol II elongation on alternative splicing. The EMBO Journal *20*, 5759–5768.

Kanopka, A., Mühlemann, O., and Akusjärvi, G. (1996). Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA. Nature *381*, 535–538.

Kar, A., Fushimi, K., Zhou, X., Ray, P., Shi, C., Chen, X., Liu, Z., Chen, S., and Wu, J.Y. (2011). RNA Helicase p68 (DDX5) Regulates tau Exon 10 Splicing by Modulating a Stem-Loop Structure at the 5′ Splice Site. Molecular and Cellular Biology *31*, 1812–1821.

Karathia, H., Kingsford, C., Girvan, M., and Hannenhalli, S. (2016). A pathway-centric view of spatial proximity in the 3D nucleome across cell lines. Scientific Reports *6*, 39279.

Kent, O.A., Reayi, A., Foong, L., Chilibeck, K.A., and MacMillan, A.M. (2003). Structuring of the 3′ Splice Site by U2AF65. J. Biol. Chem. *278*, 50572–50577.

Khan, D.H., Gonzalez, C., Tailor, N., Hamedani, M.K., Leygue, E., and Davie, J.R. (2016). Dynamic Histone Acetylation of H3K4me3 Nucleosome Regulates MCL1 Pre-mRNA Splicing. J. Cell. Physiol. n/a-n/a.

Kim, M.-S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature *509*, 575–581.

Kornblihtt, A.R. (2005). Promoter usage and alternative splicing. Current Opinion in Cell Biology *17*, 262–268.

Labena, A.A., Guo, H.-X., Dong, C., Li, L., and Guo, F.-B. (2018). The Topologically Associated Domains (TADs) of a Chromatin Correlated with Isochores Organization of a Genome.

Lareau, L.F., and Brenner, S.E. (2015). Regulation of Splicing Factors by Alternative Splicing and NMD Is Conserved between Kingdoms Yet Evolutionarily Flexible. Mol Biol Evol *32*, 1072–1079.

Latham, J.A., and Dent, S.Y.R. (2007). Cross-regulation of histone modifications. Nat Struct Mol Biol *14*, 1017–1024.

Lavigueur, A., Branche, H.L., Kornblihtt, A.R., and Chabot, B. (1993). A splicing enhancer in the human fibronectin alternate ED1 exon interacts with SR proteins and stimulates U2 snRNP binding. Genes Dev. *7*, 2405–2417.

Lee, Y., and Rio, D.C. (2015). Mechanisms and Regulation of Alternative Pre-mRNA Splicing. Annu Rev Biochem *84*, 291–323.

Liu, Z.-R. (2002). p68 RNA Helicase Is an Essential Human Splicing Factor That Acts at the U1 snRNA-5′ Splice Site Duplex. Mol Cell Biol *22*, 5443–5450.

Liu, J., Yue, Y., Han, D., Wang, X., Fu, Y., Zhang, L., Jia, G., Yu, M., Lu, Z., Deng, X., et al. (2014). A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. Nat Chem Biol *10*, 93–95.

Liu, N., Dai, Q., Zheng, G., He, C., Parisien, M., and Pan, T. (2015). N6-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. Nature *518*, 560–564.

Liu, S., Zhang, L., Quan, H., Tian, H., Meng, L., Yang, L., Feng, H., and Gao, Y.Q. (2018). From 1D sequence to 3D chromatin dynamics and cellular functions: a phase separation perspective. Nucleic Acids Res *46*, 9367–9383.

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

Luco, R.F., Allo, M., Schor, I.E., Kornblihtt, A.R., and Misteli, T. (2011). Epigenetics in Alternative Pre-mRNA Splicing. Cell *144*, 16–26.

de la Mata, M., and Kornblihtt, A.R. (2006). RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20. Nature Structural & Molecular Biology *13*, 973–980.

Matera, A.G., and Wang, Z. (2014). A day in the life of the spliceosome. Nat Rev Mol Cell Biol *15*, 108–121.

Millevoi, S., Moine, H., and Vagner, S. (2012). G-quadruplexes in RNA biology. WIREs RNA *3*, 495–507.

Misteli, T., and Spector, D.L. (1999). RNA Polymerase II Targets Pre-mRNA Splicing Factors to Transcription Sites In Vivo. Molecular Cell *3*, 697–705.

Monsalve, M., Wu, Z., Adelmant, G., Puigserver, P., Fan, M., and Spiegelman, B.M. (2000). Direct Coupling of Transcription and mRNA Processing through the Thermogenic Coactivator PGC-1. Molecular Cell *6*, 307–316.

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

Nasim, F.-U.H., Hutchison, S., Cordeau, M., and Chabot, B. (2002). High-affinity hnRNP A1 binding sites and duplex-forming inverted repeats have similar effects on 5′ splice site selection in support of a common looping out and repression mechanism. RNA *8*, 1078–1089.

Naughton, C., Avlonitis, N., Corless, S., Prendergast, J.G., Mati, I.K., Eijk, P.P., Cockroft, S.L., Bradley, M., Ylstra, B., and Gilbert, N. (2013). Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. Nature Structural & Molecular Biology *20*, 387–395.

Nguyen, H.Q., and Bosco, G. (2015). Gene Positioning Effects on Expression in Eukaryotes. Annual Review of Genetics *49*, 627–646.

Nogués, G., Kadener, S., Cramer, P., de la Mata, M., Fededa, J.P., Blaustein, M., Srebrow, A., and Kornblihtt, A. (2003). Control of Alternative Pre-mRNA Splicing by RNA Pol II Elongation: Faster is Not Always Better. IUBMB Life *55*, 235–241.

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

Paz, A., Frenkel, S., Snir, S., Kirzhner, V., and Korol, A.B. (2014). Implications of human genome structural heterogeneity: functionally related genes tend to reside in organizationally similar genomic regions. BMC Genomics *15*, 252.

Pervouchine, D.D., Khrameeva, E.E., Pichugina, M.Y., Nikolaienko, O.V., Gelfand, M.S., Rubtsov, P.M., and Mironov, A.A. (2012). Evidence for widespread association of mammalian splicing and conserved long-range RNA structures. RNA *18*, 1–15.

Pineda, J.M.B., and Bradley, R.K. (2018). Most human introns are recognized via multiple and tissue-specific branchpoints. Genes Dev. *32*, 577–591.

Ping, X.-L., Sun, B.-F., Wang, L., Xiao, W., Yang, X., Wang, W.-J., Adhikari, S., Shi, Y., Lv, Y., Chen, Y.-S., et al. (2014). Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. Cell Research *24*, 177–189.

Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K.C., Grüning, B.A., Villaveces, J., Habermann, B., Akhtar, A., and Manke, T. (2018). High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nature Communications *9*, 189.

Reymer, A., Zakrzewska, K., and Lavery, R. (2018). Sequence-dependent response of DNA to torsional stress: a potential biological regulation mechanism. Nucleic Acids Res *46*, 1684–1694.

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

Rogier Versteeg, Schaik, B.D.C. van, Batenburg, M.F. van, Roos, M., Monajemi, R., Caron, H., Bussemaker, H.J., and Kampen, A.H.C. van (2003). The Human Transcriptome Map Reveals Extremes in Gene Density, Intron Length, GC Content, and Repeat Pattern for Domains of Highly and Weakly Expressed Genes. Genome Res. *13*, 1998–2004.

Rowley, M.J., and Corces, V.G. (2018). Organizational principles of 3D genome architecture. Nature Reviews Genetics *19*, 789.

Sahebi, M., Hanafi, M.M., van Wijnen, A.J., Azizi, P., Abiri, R., Ashkani, S., and Taheri, S. (2016). Towards understanding pre-mRNA splicing mechanisms and the role of SR proteins. Gene *587*, 107–119.

Saint-André, V., Batsché, E., Rachez, C., and Muchardt, C. (2011). Histone H3 lysine 9 trimethylation and HP1γ favor inclusion of alternative exons. Nature Structural & Molecular Biology *18*, 337–344.

Sánchez-Hernández, N., Boireau, S., Schmidt, U., Muñoz-Cobo, J.P., Hernández-Munain, C., Bertrand, E., and Suñé, C. (2016). The in vivo dynamics of TCERG1, a factor that couples transcriptional elongation with splicing. RNA.

Schor, I.E., Rascovan, N., Pelisch, F., Alló, M., and Kornblihtt, A.R. (2009). Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. PNAS *106*, 4325–4330.

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

Scotti, M.M., and Swanson, M.S. (2016). RNA mis-splicing in disease. Nat Rev Genet *17*, 19–32.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.-P.Z., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772–778.

Segal, E., and Widom, J. (2009). Poly(dA:dT) tracts: major determinants of nucleosome organization. Current Opinion in Structural Biology *19*, 65–71.

Sequeira-Mendes, J., and Gutierrez, C. (2016). Genome architecture: from linear organisation of chromatin to the 3D assembly in the nucleus. Chromosoma *125*, 455–469.

Shin, S.-I., Ham, S., Park, J., Seo, S.H., Lim, C.H., Jeon, H., Huh, J., and Roh, T.-Y. (2016). Z-DNA-forming sites identified by ChIP-Seq are associated with actively transcribed regions in the human genome. DNA Res.

Snoussi, K., Nonin-Lecomte, S., and Leroy, J.-L. (2001). The RNA i-motif11Edited by J. A. Wells. Journal of Molecular Biology *309*, 139–153.

Soemedi, R., Cygan, K.J., Rhine, C.L., Glidden, D.T., Taggart, A.J., Lin, C.-L., Fredericks, A.M., and Fairbrother, W.G. (2017). The effects of structure on pre-mRNA processing and stability. Methods *125*, 36–44.

Song, J., Perreault, J.-P., Topisirovic, I., and Richard, S. (2016). RNA G-quadruplexes and their potential regulatory roles in translation. Translation *4*, e1244031.

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

Srebrow, A., and Kornblihtt, A.R. (2006). The connection between splicing and cancer. Journal of Cell Science *119*, 2635–2641.

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

Suzuki, H., and Matsuoka, M. (2017). hnRNPA1 autoregulates its own mRNA expression to remain non-cytotoxic. Mol Cell Biochem *427*, 123–131.

Szafranski, K., Fritsch, C., Schumann, F., Siebel, L., Sinha, R., Hampe, J., Hiller, M., Englert, C., Huse, K., and Platzer, M. (2014). Physiological state co-regulates thousands of mammalian mRNA splicing events at tandem splice sites and alternative exons. Nucleic Acids Res *42*, 8895–8904.

Taggart, A.J., Lin, C.-L., Shrestha, B., Heintzelman, C., Kim, S., and Fairbrother, W.G. (2017). Large-scale analysis of branchpoint usage across species and cell lines. Genome Res.

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

Tan, J., Ho, J.X.J., Zhong, Z., Luo, S., Chen, G., and Roca, X. (2016). Noncanonical registers and base pairs in human 5′ splice-site selection. Nucl. Acids Res. gkw163.

Tanaka, Y., Yamashita, R., Suzuki, Y., and Nakai, K. (2010). Effects of Alu elements on global nucleosome positioning in the human genome. BMC Genomics *11*, 309.

Tavanez, J.P., Madl, T., Kooshapur, H., Sattler, M., and Valcárcel, J. (2012). hnRNP A1 Proofreads 3′ Splice Site Recognition by U2AF. Molecular Cell *45*, 314–329.

Teves, S.S., Weber, C.M., and Henikoff, S. (2014). Transcribing through the nucleosome. Trends in Biochemical Sciences *39*, 577–586.

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

Trifonov, E.N. (2011). Cracking the chromatin code: Precise rule of nucleosome positioning. Physics of Life Reviews *8*, 39–50.

Tsochatzidou, M., Malliarou, M., Papanikolaou, N., Roca, J., and Nikolaou, C. (2017). Genome urbanization: clusters of topologically co-regulated genes delineate functional compartments in the genome of Saccharomyces cerevisiae. Nucleic Acids Res *45*, 5818–5828.

Vinogradov, A.E. (2003). DNA helix: the importance of being GC-rich. Nucleic Acids Res *31*, 1838–1844.

Wang, Z., and Burge, C.B. (2008). Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. RNA *14*, 802–813.

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

Wang, Y., Ma, M., Xiao, X., and Wang, Z. (2012). Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. Nature Structural & Molecular Biology *19*, 1044–1052.

Wilhelm, B.T., Marguerat, S., Aligianni, S., Codlin, S., Watt, S., and Bähler, J. (2011). Differential patterns of intronic and exonic DNA regions with respect to RNA polymerase II occupancy, nucleosome density and H3K36me3 marking in fission yeast. Genome Biology *12*, R82.

Wong, J.J.-L., Gao, D., Nguyen, T.V., Kwok, C.-T., Geldermalsen, M. van, Middleton, R., Pinello, N., Thoeng, A., Nagarajah, R., Holst, J., et al. (2017). Intron retention is regulated by altered MeCP2-mediated splicing factor recruitment. Nature Communications *8*, 15134.

Wu, J.Y., and Maniatis, T. (1993). Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. Cell *75*, 1061–1070.

Xiao, W., Adhikari, S., Dahal, U., Chen, Y.-S., Hao, Y.-J., Sun, B.-F., Sun, H.-Y., Li, A., Ping, X.-L., Lai, W.-Y., et al. (2016). Nuclear m6A Reader YTHDC1 Regulates mRNA Splicing. Molecular Cell *61*, 507–519.

Xie, W.J., Meng, L., Liu, S., Zhang, L., Cai, X., and Gao, Y.Q. (2017). Structural Modeling of Chromatin Integrates Genome Features and Reveals Chromosome Folding Principle. Scientific Reports *7*, 2818.

Zhao, X., Yang, Y., Sun, B.-F., Shi, Y., Yang, X., Xiao, W., Hao, Y.-J., Ping, X.-L., Chen, Y.-S., Wang, W.-J., et al. (2014). FTO-dependent demethylation of N6-methyladenosine regulates mRNA splicing and is required for adipogenesis. Cell Res *24*, 1403–1419.

Annexes

DOI: 10.15252/embj.201592849

Nucleosome eviction in mitosis assists condensin loading and chromosome condensation

Esther Toselli-Mollereau[1,3], Xavier Robellet[1,3], Lydia Fauque[1,3], Sébastien Lemaire[1], Christoph Schiklenk[2], Carlo Klein[2], Clémence Hocquet[1], Pénélope Legros[1], Lia N'Guyen[1], Léo Mouillard[1], Emilie Chautard[1], Didier Auboeuf[1], Christian Haering[2] and Pascal Bernard[1,4]

1. Univ Lyon, ENS de Lyon, Univ Claude Bernard, CNRS UMR 5239, INSERM U1210, Laboratory of Biology and Modelling of the Cell, 46 allée d'Italie, Site Jacques Monod, F-69007, Lyon, France
2. Cell Biology and Biophysics Unit, EMBL, D-69117 Heidelberg, Germany
3. Equal contribution
4. Correspondence: pascal.bernard@ens-lyon.fr

Running title: nucleosomes removal facilitates condensin loading

Total characters including spaces and excluding references: 55 652.

Abstract

Condensins associate with DNA and shape mitotic chromosomes. Condensins are enriched nearby highly expressed genes during mitosis, but how this binding is achieved and what features associated with transcription attract condensins remain unclear. Here, we report that condensin accumulates at or in the immediate vicinity of nucleosome-depleted regions during fission yeast mitosis. Two transcriptional coactivators, the Gcn5 histone acetyltransferase and the RSC chromatin remodelling complex, bind to promoters adjoining condensin binding sites and locally evict nucleosomes to facilitate condensin binding and allow efficient mitotic chromosome condensation. The function of Gcn5 is closely linked to condensin positioning, since neither the localization of Topoisomerase II nor that of the cohesin loader Mis4 is altered in *gcn5* mutant cells. We propose that nucleosomes act as a barrier for the initial binding of condensin and that nucleosome-depleted regions formed at highly expressed genes by transcriptional coactivators constitute access points into chromosomes where condensin binds free genomic DNA.

Introduction

In most eukaryotes, chromatin fibres metamorphose into compact and individualized rod-shaped chromosomes during mitosis and meiosis. This profound reorganisation, called chromosome condensation, is a strict prerequisite for the accurate segregation of chromosomes. From yeasts to human, chromosome condensation relies upon condensin complexes and Topoisomerase II α (Topo II). It is widely accepted that Topo II ensures decatenation of sister-chromatids and chromosomes (Baxter *et al*, 2011; Charbin *et al*, 2014). In contrast, how condensins reconfigure chromosome structure in a cell-cycle regulated manner remains poorly understood.

Condensins are ring-shaped ATPases that belong to the family of SMC (Structural Maintenance of Chromosomes) protein complexes, which also include cohesin, responsible for sister chromatid cohesion, and the Smc5/Smc6 complex, which is implicated in DNA damage repair (Thadani *et al*, 2012; Aragon *et al*, 2013; Hirano, 2016). Eukaryotic condensins are composed of the Smc2 and Smc4 ATPase subunits, called Cut14 and Cut3 in fission yeast, and three non-SMC subunits. Smc2 and Smc4 form a V-shaped heterodimer in which two ATPase heads face each other at the apices of two 50 nm-long coiled coil arms. A kleisin subunit (called Cnd2 in fission yeast) associates with the ATPase heads, thereby creating a tripartite ring-like structure, and recruits two HEAT-repeat containing subunits. Most eukaryotes possess two condensins, called condensin I and II, which are made of the same Smc2/Smc4 heterodimer but contain different sets of the three non-SMC subunits. Budding and fission yeasts possess a single condensin complex, which is similar to condensin I by protein sequence. Condensin II is nuclear throughout the cell cycle and accumulates on chromosomes at the beginning of prophase. In contrast, condensin I is cytoplasmic during interphase and gains access to chromosomes only from prometaphase to telophase (Hirota *et al*, 2004; Ono *et al*, 2003). Fission yeast condensin shows a similar localization pattern as condensin I, with the bulk of the complexes binding chromosomes from prophase to telophase (Sutani *et al*, 1999). However, a fraction seems to persist on chromosomes during interphase (Aono *et al*, 2002; Nakazawa *et al*, 2015).

Studies performed in a wide range of systems have substantiated the idea that condensins modify the topology of chromosomal DNA by introducing positive supercoils (Kimura & Hirano, 1997), by topological entrapment of DNA molecules within their ring-like structure (Cuylen *et al*, 2011), and/or by promoting the reannealing of unwound genomic DNA segments (Sutani *et al*, 2015). It remains unclear, however, to what extent each these molecular activities contribute to the shaping of chromosomes. Equally unclear is how condensins associate with and manipulate chromosomal DNA in the intricate context of the chromatin fiber.

From yeasts to mammals, condensins are enriched at centromeres, telomeres, and, along chromosome arms, nearby genes that are highly transcribed by either of the three RNA polymerases (D'Ambrosio *et al*, 2008; Schmidt *et al*, 2009; Dowen *et al*, 2013; Kim *et al*, 2013; Kranz *et al*, 2013; Nakazawa *et al*, 2015; Sutani *et al*, 2015). Several factors have been implicated in the recruitment of condensins along chromosomes. In budding yeast, the cohesin loading factor Scc2/4 has been reported to promote the full level of association of condensin at tRNA genes and at genes encoding ribosomal proteins (D'Ambrosio *et al*, 2008). The replication fork blocking protein Fob1, in a complex with the monopolin subunits Csm1 and Lrs4, recruits condensin at rDNA repeats (Johzuka & Horiuchi, 2009). Fission yeast monopolin associates with condensin and contributes to its localization at the rDNA repeats and also at the kinetochore, but plays no role along chromosome arms (Tada *et al*, 2011). In chicken DT40 cells, the chromokinesin Kif4 has been implicated in the localization of condensins (mostly condensin I) along the longitudinal axes of mitotic chromosomes (Samejima *et al*, 2012). In human cells, the zinc finger protein AKAP95 directly interacts with the kleisin CAP-H and takes part is the recruitment of condensin I onto chromatin (Steen *et al*, 2000; Eide *et al*, 2002). More recently, it has been reported that condensin I and II bind the N-terminal tail of histones H2A and H4, respectively (Liu *et al*, 2010; Tada *et al*, 2011). However, contacts with histones do not explain the pattern of condensins along chromosomes, and contrasting results have been obtained regarding the role played by histone tails, if any, in the chromosomal association of condensin I (Tada *et al*, 2011; Shintomi *et al*, 2015). Thus, the mechanisms through which condensins associate with chromatin remain unclear.

The enrichment of condensin nearby highly expressed genes suggests a crucial link between the localization of condensins and a feature associated with high transcription levels. In line with this, chemical inhibition of RNA Pol II reduces condensin association with mitotic chromosomes during early mitosis in fission yeast (Sutani *et al*, 2015), and transcription factors have been implicated in the loading of condensin (D'Ambrosio *et al*, 2008; Nakazawa *et al*, 2015; Iwasaki *et al*, 2015). Paradoxically, transcription of DNA repeats by RNA Pol I or Pol II obstructs the stable association of condensin in budding yeast (Johzuka & Horiuchi, 2007; Clemente-Blanco *et al*, 2009, 2011). Fission yeast condensin appears excluded from gene bodies and accumulates towards the 3' ends of highly expressed genes (Sutani *et al*, 2015; Nakazawa *et al*, 2015). Moreover, transcription by all three RNA Pols is usually repressed during mitosis in most eukaryotes (Gottesfeld & Forbes, 1997), when the association of condensin with chromosomes reaches its maximum. In budding yeast, the inhibition of RNA Pol I and Pol II by the Cdc14 phosphatase during anaphase is necessary for condensin binding (Clemente-Blanco *et al*, 2009, 2011). Thus, the association of condensin with chromatin appears both positively and negatively linked with transcription. What remains unclear, however, is the molecular

102 determinant(s) that defines condensin association sites and what feature(s) associated with
103 transcription takes part in condensin binding.
104 Active gene promoters are associated with histone H3 acetylated at lysines 9 and 14 (H3K9ac and
105 H3K14ac) (Pokholok *et al*, 2005; Roh *et al*, 2005; Wiren *et al*, 2005). Although the bulk of chromatin is
106 deacetylated during mitosis in mammals (Kruhlak et al., 2001; Patzlaff et al., 2010), traces of H3K9ac
107 and H3K14ac persist at some gene promoters (Wang & Higgins, 2013). Gcn5 is the histone
108 acetyltransferase (HAT) subunit of the SAGA complex, an evolutionarily-conserved modular
109 transcription coactivator that acetylates nucleosomes, notably H3K9, H3K14 and H3K18 (Koutelou *et*
110 *al*, 2010; Weake & Workman, 2012). Gcn5-containing SAGA occupies the promoters and coding
111 regions of active genes. At promoters, Gcn5 occupancy increases with transcription rate (Robert *et*
112 *al*, 2004; Govind *et al*, 2007; Johnsson *et al*, 2009; Xue-Franzen *et al*, 2013; Bonnet *et al*, 2014). By
113 acetylating nucleosomes, SAGA promotes the local formation of an "open" chromatin structure
114 where the transcription pre-initiation complex assembles.
115 Using fission yeast as a model system, we show that condensin binding to chromosomes and mitotic
116 chromosome condensation rely upon Gcn5 HAT activity. Although the majority of Gcn5 is transiently
117 displaced from chromosomes in early mitosis and the bulk of chromatin is deacetylated, Gcn5 and
118 acetylated H3 persist at promoters adjoining a number of highly expressed genes. There, Gcn5 and
119 the ATP-dependent chromatin remodelling complex RSC (Remodels the Structure of Chromatin) evict
120 nucleosomes and promote the efficient binding of condensin at these nucleosome-depleted regions.
121 Our results suggest that nucleosomes constitute a barrier for the localization of condensin, which is
122 overcome by Gcn5-mediated histone acetylation and chromatin remodelling. Besides providing
123 unanticipated insights into the mechanism of condensin binding to chromatin, our study suggests
124 that the presence of exposed, non-nucleosomal DNA may be an important feature that attracts
125 condensins to highly expressed genes in eukaryotes.
126

Results

Gcn5 takes part in mitotic chromosome condensation

Fission yeast cells carrying the thermo-sensitive *cut3-477* mutation in the Smc4 condensin subunit cease to divide at 36°C, but continue to proliferate at the semi permissive temperature of 32°C, even though condensin binding to chromosomes is reduced and mitotic chromosome condensation is partly impaired (Saka *et al*, 1994; Tada *et al*, 2011; Robellet *et al*, 2014). To identify factors that collaborate with condensin, we screened for mutations synthetically lethal with *cut3-477* at 32°C (Robellet *et al*, 2014). We isolated *gcn5-47*, a nonsense G765A mutation in the *gnc5* open reading frame that is predicted to eliminate the C-terminal bromodomain of the protein (Fig. 1A-B). Deletion of the complete *gcn5* open reading frame was also colethal with *cut3-477* at 32°C (Fig. 1A) and with the *top2-250* allele of Topo II (Appendix Fig. S1A). In sharp contrast, neither *gcn5-47* nor *gcn5Δ* lowered the restrictive temperature of mutations *eso1-H17* and *rad21-K1*, which affect sister-chromatid cohesion (Appendix Fig. S1B). This indicates that lack of Gcn5 function does not confer a blind hypersensitivity to any perturbation in the structure of chromosomes. Thus, Gcn5 interacts positively and specifically with key chromosome condensation factors.

To assess whether Gcn5 plays a role in mitotic chromosome condensation, we used a quantitative condensation assay that measures the three dimensional distances between two fluorescently labeled loci located on the left arm of chromosome I as cells pass through mitosis (Petrova *et al*, 2013). In wild type cells, the distance between two loci separated by 0.5 or 1.0 Mb of DNA decreased about 2-fold from G2 phase until late anaphase (Fig. 1C). In *gcn5-47* mutant cells, the distances between the two loci combinations remained larger throughout the mitotic time course (Fig. 1C), even though condensation was not as severely affected as in the *cut14-208* condensin mutant. The distances between the fluorescently labelled loci were also slightly enlarged during interphase in *gcn5-47* cells (Fig. 1C). Reduced acetylation of nucleosomes might relax chromatin fibers during interphase. Alternatively, or additionally, lack of Gcn5 might impair condensin-mediated chromosome shaping throughout the cell cycle.

Consistent with impaired condensation, *gcn5-47* mutant cells exhibited frequent chromatin bridges or chromatin trailing in anaphase (Fig. 1D) and failed to efficiently disentangle the rDNA repeats located on the arms of chromosome III (Fig. EV1). These phenotypes are frequently observed as a consequence of defects in mitotic chromosome condensation (Tada *et al*, 2011), for example in the *cut3-477* condensin mutant (Fig. 1D and EV1A). However, most chromatin bridges disappeared during late anaphase B in *gcn5-47* mutant cells, unlike in the more severe *cut3-477* mutant. Plotting the number of chromatin bridges as a function of spindle length suggested that chromosome arms

eventually achieved complete separation in *gcn5-47* cells, at a time when the mitotic spindle was 25% longer compared to wild-type cells (Fig. 1E). Taken together, these data indicate that Gcn5 is required for the efficient condensation of chromosome arms during mitosis. Because condensation is partly impaired in the absence of Gcn5 function, the complete separation of chromosome arms necessitates a longer mitotic spindle.

The role of Gcn5 in chromosome condensation relies on its acetyltransferase activity

The function of Gcn5 as a transcriptional coactivator raised the possibility that its effect on chromosome condensation might be indirect, i.e. that Gcn5 controls the transcription of a *bona fide* condensation factor. However, our data suggest that this is unlikely. The mRNA levels of Topo II and all five condensin subunits were not notably reduced in *gcn5-47* or *gcn5Δ* cells (Fig. EV2A), nor were Cut3, Cnd2 or Topo II protein levels (Fig. EV2B-C). Moreover, efficient co-immunoprecipitation of Cut3-HA with Cnd2-GFP suggested that the integrity of the condensin complex was not affected in the absence of Gcn5 (Fig. EV2C). We also re-analyzed available transcriptome data for wild type and *gcn5Δ* mutant cells (Helmlinger *et al*, 2008) by comparing the mRNA levels of 236 genes that have been reported in pombase (www.pombase.org) to be required for chromosome segregation and/or condensation, or to genetically or physically interact with condensin. Using a threshold of at least 1.5 fold up- or down-regulation, we found a single hit: the *cnp3* gene, which encodes a centromeric protein (Fig. EV2D). However, restoring the level of *cnp3* mRNA by ectopic expression did not suppress the colethality between *gcn5-47* and *cut3-477* at 32°C (Fig. EV2E-F), indicating that the functional interaction between Gcn5 and condensin was independent of *cnp3*. We conclude that Gcn5 plays a genuine role in mitotic chromosome condensation.

The Gcn5-containing SAGA complex consists of three independent functional modules (Weake & Workman, 2012; see Fig. 2A): the HAT module, which is composed of Gcn5, Ada2 and Ada3, the Spt module, which includes the Spt8 subunit implicated in the recruitment of the TATA Binding Protein to certain promoters, and the ubiquitin protease module, which contains Sgf11 and Sgf73. To test whether SAGA components other than Gcn5 genetically interact with condensin, we combined deletions of representative subunits of each of the three modules with *cut3-477* (Fig. 2A). Lack of the *ada2* or *ada3* subunits of the HAT module was co-lethal with *cut3-477* at 32°C, but deletion of subunits of the two other modules had no effect. This finding strongly suggests that condensin is closely linked to the Gcn5 HAT activity. We confirmed this conclusion by testing the catalytically inactive *gcn5-E191Q* mutant (Helmlinger *et al*, 2008) (Fig. 2B).

Different HATs can have overlapping functions and Mst2 and Elp3 are partially redundant with Gcn5 (Nugent *et al*, 2010). Of five additional HATs – Hat1, Naa40, Rtt109, Mst2 and Elp3 – tested, none showed an obvious negative genetic interaction with *cut3-477* (Fig. 2C). However, simultaneous

deletion of Mst2 and Gcn5 reduced viability of the *cut3-477* mutant even further than depletion of Gcn5 alone, which suggests that the two HATs might be partially redundant (Fig. 2D). Thus, Gcn5 functionally interacts with condensin through its acetyltransferase activity, possibly in the context of the SAGA complex, and Mst2 is partly redundant with Gcn5 for this function.

Gcn5 and Mst2 facilitate condensin binding to mitotic chromosomes

To test whether Gcn5 and Mst2 might affect the association of condensin with chromosomes, we assessed chromosomal condensin levels using chromatin immunoprecipitation (ChIP) against the kleisin subunit Cnd2 tagged with GFP. Previous genome-wide mapping experiments have shown that the condensin binding profile along chromosome arms in mitosis consists of low-occupancy binding sites and hot spots of association, which correlate with highly expressed genes (D'Ambrosio *et al*, 2008; Schmidt *et al*, 2009; Nakazawa *et al*, 2015; Sutani *et al*, 2015). We quantified Cnd2-GFP binding using qPCR at the kinetochore (*cnt1*), pericentric heterochromatin (*dh1*) and 14 loci along chromosome arms that represent 9 high-occupancy and 5 low-occupancy binding sites. Since highly expressed genes can be vulnerable to misleading ChIP enrichments (Teytelman *et al*, 2013), we first verified that the association of Cnd2-GFP with these genes was reduced in the condensin mutant *cut3-477* (Fig. EV3A-C), as we would expect for *bona fide* condensin binding sites. We then arrested wild type or *gcn5* mutant cells in pro/metaphase by the *nda3-KM311* tubulin mutation and determined the chromosomal association of Cnd2-GFP by ChIP. In cells lacking Gcn5, binding of Cnd2-GFP was significantly reduced at the kinetochore domain and at all high-occupancy condensin binding sites (Fig. 3A). ChIP signals were even further decreased in cells lacking both Gcn5 and Mst2 (Fig. 3A). The reduced ChIP signals were not due to a decrease in Cnd2-GFP protein levels in *gcn5Δ mst2Δ* cells (Fig. 3B). In sharp contrast, Cnd2-GFP occupancy at pericentric heterochromatin (dh1) and at all low-occupancy condensin binding sites remained unchanged.

To assess the total levels of chromosome-bound condensin, we measured the amount of Cut3-HA on mitotic chromosome spreads. To identify mitotic chromosomes, we used Dam1-GFP, which is recruited to kinetochores specifically during mitosis (Liu *et al*, 2005). Cut3-HA was markedly enriched on chromosomes positive for Dam1-GFP in wild type cells (Fig. 3C). The amount of Cut3-HA bound to mitotic chromosomes was reduced by ~ 30% in the absence of Gcn5 and by ~ 40% in the absence of both Gcn5 and Mst2, even though total Cut3-HA protein levels were not reduced (Fig. 3D). This result confirms the ChIP experiment and, importantly, rules out the possibility that condensin merely relocates from its canonical high-occupancy binding sites in the absence of Gcn5 and Mst2. Moreover, we observed no reduction in the association of Topo II or the cohesin loader Mis4 with mitotic chromosomes of *gcn5Δ mst2Δ* cells (Fig. EV4), which reinforces the conclusion that Gcn5 is specifically linked to condensin. Together, these data indicate that Gcn5 and Mst2 specifically assist

condensin binding to mitotic chromosomes at core centromeres and at its high-occupancy sites near genes highly transcribed by RNA Pol II.

Gcn5 and acetylated H3 co-occupy condensin binding sites during mitosis

To investigate how Gcn5 could act to regulate condensin binding during mitosis, we assessed transcription of condensin binding sites by RNA Pol II, since active transcription is believed to antagonize condensin binding (Johzuka & Horiuchi, 2007; Clemente-Blanco *et al*, 2009, 2011). We observed no increase in the occupancy of transcriptionally engaged RNA Pol II throughout condensin binding sites during mitosis in the absence of Gcn5 or both Gcn5 and Mst2 (Fig. EV3D and E), arguing against this possibility. Next, we investigated the localization of Gcn5 during mitosis. Immunofluorescence studies have shown that Gcn5 dissociates from mitotic chromosomes in vertebrate cells (Orpinell *et al*, 2010). Using chromosome spreading we measured the amount of Gcn5-myc and Cut14-HA (condensin) on interphase and mitotic chromosomes (Fig. 4A). We categorized spread nuclei as interphase, prophase or prometaphase/metaphase by judging the Cut14-HA/DAPI ratio. We detected less Gcn5 associated with prophase chromosomes as compared to interphase chromosomes, but levels on prometaphase/metaphase chromosomes were comparable to those of interphase ones. These results imply that Gcn5 might only temporarily dissociate from chromosomes during entry into mitosis but then reassociate at a time when condensin levels further increase. We confirmed by ChIP the presence of Gcn5 on mitotic chromosomes by comparing Gcn5-myc levels at promoters adjoining condensin binding sites in asynchronous cells (80% of which are in G2) and cells arrested in pro/metaphase (Fig. 4B). 5 promoters were chosen among high-occupancy condensin binding sites and 4 among low-occupancy binding sites. In asynchronous cells, we detected Gcn5 in variable amounts at all promoters. In cells arrested in mitosis, the levels of promoter-bound Gnc5 were considerably higher at all condensin high-occupancy sites compared to condensin low-occupancy sites. Remarkably, Gcn5 binding decreased at all low-occupancy sites during mitosis. In sharp contrast, Gcn5 levels remained unchanged or even increased at all promoters adjoining high-occupancy sites (Fig. 4B), where condensin binding depends on functional Gcn5 (Fig.3A). Whether Gcn5 occupancy increases or drops at promoters during mitosis is therefore linked to condensin occupancy, but unrelated to its absolute binding levels in interphase (compare Gcn5-myc levels at the *slp1* high-occupancy binding site to the four low-occupancy sites, Fig. 4B). Despite its enrichment at promoters, we failed to detect Gcn5-myc within adjacent gene bodies or at their 3' ends during mitosis (Fig. EV5A), which might reflect a more dynamic association of Gcn5 with transcribed regions (Bonnet *et al*, 2014). These data suggest that Gcn5 is specifically retained or even enriched during mitosis at promoters adjoining high-occupancy condensin association sites, where condensin binding in return relies upon Gcn5.

To explore the state of chromatin at these promoters, we monitored H3K9ac, H3K18ac and H3K14ac in asynchronous and mitotic cells. Reminiscent of mammalian cells, the bulk of H3K9ac and H3K18ac was reduced in fission yeast cells arrested in pro/metaphase, whilst steady state levels of H3K14ac remained unaltered (Fig. 4C). We used ChIP to assess whether traces of H3K9ac and H3K18ac might persist at condensin binding sites co-occupied by Gcn5. To control for nucleosome occupancy, we performed in parallel ChIP against total H3 using an antibody directed against the C-terminal part of H3 (H3-Ct) and determined the amount of acetylated H3 (H3ac) as a ratio of H3ac/H3-Ct in cells arrested in mitosis (Fig. 4D-E). We found that H3K9ac and H3K18ac levels at promoters correlated with the occupancy of Gcn5, being markedly enriched at the promoters of high-occupancy condensin binding sites when compared to the heterochromatic *dh1* site used as negative control (Fig. 4E). H3K9ac and H3K18ac were also detected at promoters of low-occupancy condensin binding sites, but acetylation ratios were considerably lower. Moreover, Gcn5 was required for the enrichment of H3K9ac and H3K18ac at all promoters tested (Fig. 4E). Deletion of Mst2 did not further reduce H3K9ac and H3K18ac, indicating that these histone modifications were mostly deposited by Gcn5. H3K14ac was also present at promoters in mitosis, but, unlike for H3K9ac and H3K18ac, levels were regulated by both Gcn5 and Mst2 (Fig. EV5B). Taken together, these data suggest that the bulk of Gcn5 dissociates from chromosomes during prophase, but a fraction of Gcn5 persists at, or is recruited during prometaphase to, promoters adjoining RNA Pol II-transcribed genes that are strongly co-occupied by condensin, where it ensures the persistence of histone acetylation.

Gcn5 plays a role in nucleosome depletion at condensin binding sites

Given that Gcn5 takes part in the eviction of nucleosomes at active genes (Govind et al., 2007; Xue-Franzen et al., 2013), we reasoned that Gcn5 might assist condensin association during mitosis by controlling nucleosome occupancy. To test this idea, we arrested cells in pro/metaphase (Fig. 5A) and assessed nucleosome occupancy at condensin binding sites by MNase digestion of mitotic chromatin (Fig. 5A and EV6A) and massive parallel sequencing of the resulting mononucleosomal DNA fragments (MNase-seq). We considered as nucleosome depleted regions (NDRs) any chromosomal region of at least 150 bp in length exhibiting a normalized MNase-seq coverage depth smaller than 0.4 (Soriano *et al*, 2013). With those settings, we identified 6816 ± 732 NDRs of an average size of 275 ± 6 bp (n=3 replicates, see Appendix Table S2). Given the enrichment of Gcn5 at gene promoters, we assessed nucleosome occupancy upstream of transcription start sites (TSS) of genes bound by condensin. We found that all the 47 high-occupancy condensin binding sites overlapping with Pol II-transcribed genes, identified by Sutani *et al.* (Sutani *et al*, 2015), were situated in the immediate vicinity of a NDR preceding a TSS (Fig. 5B, EV6B and EV7 and Appendix Table S2). Moreover, 40 out of these 47 high-occupancy condensin binding sites clearly overlapped with NDRs located at the 3' end

of Pol II genes (Fig. 5B, EV6B and EV7 and Appendix Table S2). Permutation tests confirmed that the co-localisation was highly statistically significant (Fig. 5C). Furthermore, by re-assigning all the reads from condensin ChIP-seq experiments (Sutani *et al*, 2015) to NDR versus non-NDR DNA sequences, we found NDR DNA sequences specifically enriched in the fraction co-immunoprecipitated with condensin (Fig. EV6B). This suggests that the vast majority of condensin binding sites, and not solely the high-occupancy ones, overlaps with an NDR and/or resides in the immediate vicinity of an NDR.

The nucleosome pattern significantly changed in *gcn5* mutant cells. We analysed nucleosome occupancy at promoters upstream of the 48 condensin binding sites and at the 41 NDRs covered by condensin at the 3' end of genes. In both cases, we found that nucleosome occupancy increased within the core of the NDR, and that the positioning of the two nucleosomes delimitating the NDR also markedly increased (Fig. 5B, D-E and Fig. EV7). Again, we observed a cumulative effect of *gcn5Δ* and *mst2Δ* mutations. We confirmed the increased nucleosome occupancy and the cumulative effect by MNase-qPCR (Fig. EV6C) and by ChIP against histone H3 (Fig. 5F). Note that nucleosome occupancy appeared unchanged at the 5S rRNA, *cnd1* and *uge1* genes in cells lacking Gcn5 or both Gcn5 and Mst2 (Fig. EV6D and Fig. 5F), where the binding of condensin remained unchanged (Fig. 3A). These data indicate that Gcn5 and Mst2 collaborate during mitosis to evict nucleosomes from NDRs, both at gene promoters and at the 3' end of genes, which constitute high-occupancy condensin binding sites.

Nucleosome eviction assists condensin binding

If nucleosome eviction were important for condensin binding, then increasing nucleosome occupancy by means other than Gcn5 or Mst2 inactivation should similarly reduce condensin binding. Arp9 and Snf21 are two subunits of RSC (Remodels the Structure of Chromatin), an ATP-dependent chromatin remodelling complex that evicts nucleosome from promoters (Monahan *et al*, 2008; Lorch *et al*, 2014). We had previously identified *arp9-127* and *snf21-129* loss-of-function mutations by screening for synthetic lethality with *cut3-477* at 32°C (Robellet *et al*, 2014). We therefore asked whether RSC could regulate condensin binding to chromosomes during mitosis through its function as nucleosome remodeller. We arrested wild type, *arp9Δ* and *snf21-129* cells in mitosis (Fig. 6A) and processed chromatin for simultaneous ChIPs against histones H3, H4 and Cnd2-GFP. We found that the occupancy of H3 and H4 increased at the promoters of both condensin high-occupancy (*prl53*, *exg1*, *ecm33*, *cdc22*, *slp1* and *snoU14*) and low-occupancy (*5S* and *gly05*) sites in the *arp9Δ* and *snf21-129* mutants (Fig. 6B). The occupancy of histone H3 increased also at the 3' end of some but not all tested genes, though the effects were less dramatic (Fig. EV8). This suggests that RSC evicts nucleosomes mainly at a broad range of gene promoters. Remarkably, the binding of Cnd2-GFP in the

arp9Δ and *snf21-129* mutants concomitantly dropped at all condensin binding sites, including *5S* and *gly05*, though the reductions were much less dramatic at these sites compared to the other sites (Fig. 6C). Note that the steady state level of Cnd2-GFP remained unaffected in the *arp9Δ* or *snf21-129* genetic background (Fig. 6D). These data strongly support the conclusion that nucleosome eviction facilitates condensin binding during mitosis.

Given that the bromodomain-containing Snf21 protein binds H3K14ac (Wang *et al*, 2012), we asked whether Gcn5 and Mst2 might recruit RSC at condensin binding sites in mitosis. Even though binding of chromatin remodellers is notoriously difficult to assay by ChIP, we detected Snf21-flag at promoters upstream of condensin binding sites in cells arrested in mitosis (Fig. 6E and F). The association of Snf21 at the promoters of *snoU14* was reduced in cells lacking Gcn5 or both Gcn5 and Mst2, but remained unchanged at the seven other tested condensin binding sites. Thus, RSC is present during mitosis at gene promoters adjoining condensin binding sites and its localization near most, but not all, of these sites is independent of Gcn5 and Mst2.

Discussion

Condensin binding to DNA is instrumental for chromosome condensation, but how binding is achieved in the context of a chromatin environment has remained elusive. Here, we provide evidence that nucleosome eviction, promoted by the histone acetyltransferases Gcn5 and Mst2 as well as the RSC chromatin-remodelling complex, is necessary for condensin binding to chromosomes during mitosis and for proper mitotic chromosome condensation. Thus, nucleosomes must constitute a barrier for the association of condensin. Nucleosome-depleted regions generated at least in part by transcription cofactors, such as Gcn5, Mst2 and RSC, might therefore constitute access points into the chromosome where condensin first associates with exposed double-stranded DNA helices.

One key finding of our study is the preferred localization of condensin at or in the immediate vicinity of NDRs (Fig. 5B and EV6). This is particularly true for high-occupancy condensin binding sites (see Fig. 5A). In itself, this pattern of localization suggests that nucleosomes constitute an obstacle for the localization of condensin. In the case of transcription, the nucleosome barrier is overcome by histone acetylation and nucleosome removal (Owen-Hughes & Gkikopoulos, 2012). We found that most Gcn5 transiently dissociates from chromatin during prophase in fission yeast, and that the bulk of H3K9ac and H3K18ac is deacetylated, reminiscent of the chromatin modifications that occurs during mitosis in mammalian cells. However, a fraction of Gcn5 remains on chromosomes during mitosis and is enriched at a number of promoters of Pol II-transcribed genes. At these sites, Gcn5 maintains acetylation of H3K9 and H3K18 and ensures condensin binding. Thus, nucleosome acetylation by Gcn5 might play a role in regulating condensin association with chromatin in mitosis. The finding that

lack of both Gcn5 and Mst2 causes a cumulative reduction of H3K14ac in addition to H3K9ac and H3K18ac, and reduces condensin occupancy further, strengthens this conclusion.

The acetylation of histone H3 is a transcription conductive modification. Gcn5 and Mst2 may therefore facilitate condensin binding by promoting transcription. Although we cannot formally rule out this possibility, we think this is unlikely, because we identified five out of nine condensin binding sites where the association of condensin with chromatin was reduced in the absence of Gcn5 whilst the occupancy of RNA Pol II phosphorylated on serine 2 remained unchanged (see *rds1*, *prl53*, *ecm33*, *snoU14* and *cdc22* in Fig. 3A and Fig. EV3D). This suggests that transcription and condensin binding can be uncoupled.

Acetylated nucleosomes are targeted for removal by bromodomain-containing nucleosome remodellers (Owen-Hughes & Gkikopoulos, 2012). In line with this, nucleosome occupancy is increased at 3' NDRs that overlap with condensin binding sites, and at NDR upstream of promoters, in *gcn5* mutant cells, and is further increased when both Gcn5 and Mst2 are missing. Crucially, the increase in nucleosome occupancy is accompanied by a proportional reduction in condensin binding. Moreover, increasing nucleosome occupancy by directly altering RSC activity is sufficient to decrease condensin binding. Thus, nucleosome eviction most likely plays a key role in condensin binding to chromatin during mitosis. Although our results do not exclude the possibility that Gcn5 assists condensin binding at least in part by acetylating condensin and/or other non-histone proteins, they strongly suggest that Gcn5 and Mst2 promote condensin binding in mitosis by evicting nucleosomes from condensin binding sites.

Like Gcn5, RSC is present at promoters adjoining condensin binding sites during mitosis and is necessary for condensin binding at the 3' end of genes. However, RSC deficiency increases nucleosome occupancy strongly at gene promoters but only moderately at the 3' end of genes (see Fig. 6B and EV8). This suggests that nucleosome eviction at gene promoters plays a crucial role in the binding of condensin at the 3' end of genes. Thus, given the enrichment of Gcn5 at gene promoters, and the physical and functional interactions between condensin and the TATA Binding Protein 1 (Iwasaki *et al*, 2015), it is tempting to speculate that condensin rings first associate with chromosomes at promoter NDRs and subsequently translocate towards the 3' end of genes, as proposed for the related cohesin complex (Lengronne *et al*, 2004).

The central domain of centromeres (*cnt1*) is transcribed by RNA Pol II and is a site of high nucleosome turn-over (Choi *et al*, 2011; Sadeghi *et al*, 2014). The reduced association of condensin at *cnt1* in the absence of Gcn5 (Fig. 3A) might therefore indicate that nucleosome eviction and/or dynamics contribute to the association of condensin, along with Monopollin (Tada *et al*, 2011), at centromeres.

Chromatin modifying activities in addition to the activities of Gcn5, Mst2 and RSC are likely to take part in condensin's binding to chromosomes. Although the vast majority of condensin binding sites coincides with NDRs (see Fig. EV6B), Gcn5 and Mst2 seem to assist condensin binding mainly at high-occupancy sites but play only a negligible role (if any) at low-occupancy binding sites. RSC, in contrast, seems to play a more general role (Fig.6B-C). The fact that nucleosome residence increases at high-occupancy condensin binding sites in the absence of Gcn5 and Mst2, despite the presence of RSC, implies that Gcn5 and Mst2 promote condensin binding at these sites by recruiting additional chromatin remodellers, which are at least partly redundant with RSC. The recent finding that budding yeast Gcn5 acts cooperatively, and often redundantly, with the Swi/Snf nucleosome remodelling enzyme to evict promoter nucleosomes (Qiu *et al*, 2016), supports our conclusion.

Note, however, that nucleosome depletion is unlikely to drive condensin binding by itself. We identified ~7000 NDRs in mitotic chromosomes in cells arrested in pro/metaphase, but solely ~400 condensin peaks (48 high and 340 low-occupancy) have been identified by ChIP-seq at a similar cell cycle stage (Sutani *et al*, 2015). This suggests the existence of NDRs devoid of condensin during mitosis. The corollary, therefore, is that nucleosome eviction is necessary but not sufficient for condensin binding. Hence, additional features/activities must attract condensin. Budding yeast condensin has been shown to preferentially bind free over nucleosomal DNA in a sequence independent manner (Piazza *et al*, 2014). It has also been shown that RSC recruits the cohesin loader Scc2/Scc4 complex at NDRs of active promoters in budding yeast (Lopez-Serra *et al*, 2014). The role of nucleosome eviction may therefore be to provide access to free genomic DNA for condensin and/or for DNA binding factors important for condensin association. Also, condensin, cohesin and perhaps all SMC complexes may contact and subsequently entrap chromosomal DNA at NDRs.

Several lines of evidence suggest that the model of condensin loading at NDRs applies to most eukaryotes. In most species, an NDR is present around the transcription start site of active genes (Sadeh & Allis, 2011), the size of which increases with transcription rates to culminate with a disruption of chromatin for the most highly expressed genes (Lantermann *et al*, 2010; Soriano *et al*, 2013), which are occupied by condensin. Moreover, the fact that nuclease sensitivity patterns are preserved during mitosis suggests that NDRs persist despite transcriptional shut down (Gottesfeld & Forbes, 1997). Thus, the positions of condensins should, in principle, coincide with NDRs at highly expressed genes in a wide range of species. In good agreement, condensin preferentially occupies genes that are co-occupied by RSC and Gcn5 in budding yeast (Venters *et al*, 2011), it colocalizes with NDRs at tRNA genes (Piazza *et al*, 2014), and relies upon the histone chaperone Asf1 for its chromosomal association, which, together with FACT, regulates nucleosome turnover (Dewari & Bhargava, 2014). Nucleosome remodellers are components of mitotic chromosomes in vertebrates (MacCallum *et al*, 2002; Ohta *et al*, 2010) and a recent study indicates that the mobilisation of

embryonic nucleosomes by the histone chaperones Nap1 and FACT is necessary for the assembly of Xenopus mitotic chromosomes (Shintomi *et al*, 2015).

Altogether, those observations are consistent with the idea that nucleosomes constitute a barrier for the initial binding of condensins. The enrichment of condensins nearby highly expressed genes observed from yeasts to mammals may thus reflects the fact that condensin rings make their first contact with free, exposed chromosomal DNA at nucleosome depleted regions created by transcription-coupled nucleosome eviction.

Materials and Methods

Media, molecular genetics, and strains

Media and molecular genetics methods were as described previously (Moreno *et al*, 1991). Complete medium was YES+A. Gene deletions or tagging were performed using a polymerase chain reaction (PCR)-based method (Bahler *et al*, 1998). All deletions were confirmed by PCR on genomic DNA. Tagging was validated par Western blotting and Sanger sequencing. Strains used in this study are listed in Appendix Table S3.

Mitotic arrest

All mitotic arrests were performed at 19°C using the cold-sensitive *nda3-KM311* mutation. Unless otherwise stated, mitotic indexes were measured by scoring the accumulation of Cnd2-GFP in the nucleus (Sutani *et al*, 1999).

Chromosome condensation assay

Cells were grown at 25°C to 0.5 – 1 x 10^7 cells/mL. Early G2 cells were purified by centrifugation through a 7%-30% (w/v) lactose step gradient and pipetted onto a 35 mm glass bottom dish (MatTek, No 1.5 P35G-1.5-10-C) covered with 2 mg/mL BS1 lectin (Sigma, L2380) and the dish was shifted to 34°C. The glass bottom dish was filled with 2 mL YES+A pre-warmed to 34°C and placed onto the sample stage of the microscope. Imaging was performed using a 100x objective (Olympus) on a DeltaVision system equipped with a Photometrics CoolSnap HQ camera (Roper Scientific) binning 2x2 pixels. Pixel size was about 130 x 130 nm. A GFP-DsRed Dual dichroic mirror (art F51-019) was used in combination with two bandpass filters to switch between excitation bands. Imaging was started 1 h after lactose gradient centrifugation and Z-stacks to cover a depth of 400 nm were acquired every 45 s for 6 fields of view (512 x 512 px) for 1 h. Imaging data was analysed by determining foci centroids

using a custom ImageJ plugin as described (Petrova *et al*, 2013). Anaphase onset was defined as the frame at which the centromere-proximal foci split. Distance time series were aligned based on anaphase onset. Average and standard deviation were calculated for each time point.

Immunofluorescence

Immunofluorescence was performed as described (Robellet *et al*, 2014). Images were processed and distances measured using Image J.

Chromosome spreading

Chromosome spreads were performed as described (Bahler *et al*, 1993). 5 x 10^7 cells were digested with 2mg of Lysing enzymes (Sigma L-1412) and nuclear spreading was performed in the presence of 1% formaldehyde and 1.5% Lipsol. Immunofluorescence was carried out in PBS supplemented with 1% fish skin gelatin (v/v) and 0.5% BSA (w/v) using monoclonal anti-HA 12CA5 (1/800), polyclonal rabbit anti-GFP A11122 (1/800), or polyclonal anti-Myc A-14 (1/500). Images were acquired using an axioimager Z1 microscope and immunofluorescence signals quantified with Image J software.

Co-immunoprecipitations

Co-immunoprecipitations were performed as described (Vanoosthuyse *et al*, 2014).

Chromatin Immunoprecipitation and quantitative qPCR

ChIPs were performed as described (Vanoosthuyse *et al*, 2014). 2x10^8 cells were fixed with 1% formaldehyde at 19°C for 30 min, washed with PBS and lysed using acid-wash glass beads in a Precellys homogenizer (3 times 10" at full speed with 30 sec pauses in ice). Chromatin was sheared in 300-900 bp fragments by sonication of whole cell extracts at 4°C using a Diagenode bioruptor [10 cycles 30s on / 30s off], max power. Clarified chromatin was split in two equivalent fractions subjected to parallel immunoprecipitations using magnetic Dynabeads previously incubated with the appropriate antibody. Total and immuno-precipitated DNA was purified using the NucleoSpin PCR clean-up kit (Macherey-Nagel). DNA was analysed on a Rotor-Gene PCR cycler using QuantiFast SYBR Green mix. Primers are listed in Appendix Table S4.

Antibodies

Antibodies used in this study are listed in Appendix Table S5.

Reverse -transcription and quantitative (q)PCR

Total RNA was extracted from 10^8 cells by standard hot-phenol method. Reverse transcription was performed on 500 ng of total RNA using Superscript II (Life Technologies) and random hexamers in the presence or absence of Reverse Transcriptase. cDNAs were quantified by real time qPCR on a Rotor-Gene PCR cycler using QuantiFast SYBR Green mix.

MNase-seq

MNase digestion of mitotic chromatin was performed as described (Lantermann *et al*, 2009) using cells arrested in early mitosis by the *nda3-KM311* mutation at 19°C and crosslinked with 0.5% formaldehyde for 30 min $2x10^9$. Chromatin was digested with increasing amount of MNase to reach a mononucleosome to di-nucleosome ratio of ~80:20. Mononucleosomal DNA fragments were separated on an agarose gel, extracted from the gel and subjected to massive parallel sequencing on an Illumina NextSeq 500 Apparatus. Between 46,818,494 and 62,258,601 single-end reads of 75 bp in length were obtained per sample and aligned to the *S. pombe* genome (Ensembl ASM294v2, May 2009) using Bowtie 2.2.4 with default parameters. Detection of the NDRs was performed as described (Soriano *et al*, 2013). Each sample was normalized by dividing signals for every base pair by the mean of coverage depth. NDRs were identified as regions of at least 150 bp in length with normalized sequence coverage inferior to 0.4, and were fused together if separated by less than 15 bps.

Bioinformatic analysis of nucleosome occupancy and positioning at condensin binding sites

For each condensin binding site, adjacent NDRs of reference were manually identified and their coordinates determined by an iterative process. Manually-identified NDRs were aligned with their respective counterparts calculated for each biological replicate (n=3). When the reference NDR was overlapped by at least 70% of its length by its calculated counterpart, the two NDRs were merged, generating a new NDR of reference, which was used for the next iteration. Reference NDRs, and calculated NDRs detected in each replicate, are listed in Appendix Table S2. To compare coverage depth (Fig. 5E), 200 bp upstream and 200 bp downstream of the NDR were added to take into account the flanking nucleosomes. The sum of the normalized coverage depth at each base was calculated and divided by the total length. Resulting distributions were tested for gaussian behavior using the Shapiro-Wilk test. The Kruskal-Wallis test was used for comparison if at least one distribution did not follow the normal law.

Metagene generation

NDRs of reference were fractioned in 10 bins, in which the coverage depth was averaged over the bases. The 200 bp upstream and downstream of the NDR were each resumed in 20 bins following the same process as for the NDRs. The mean enrichments were calculated for each bin in each sample.

Permutation test

Positions of the 48 condensin peaks were shuffled 100000 times using the shuffle tool of Bedtools, without any constraint, to obtain a theoretical distribution of the overlaps between condensin peaks and NDRs. The permutation test was performed on this distribution. The p-value was calculated by dividing the number (x) of permutations giving higher numbers of bases in overlaps than the observed value increased by one (the observed value), by the total number of sets of positions (the set of the observed positions and the 100000 permutations of the positions): (x + 1)/(100000 + 1).

Enrichment of condensin in NDRs

Datasets of ChIP-seq against Cut14-pk9 (Sutani *et al*, 2015) n° SRR1564296, SRR1557176, SRR1559300, SRR1559301, SRR1557175, SRR1557178 were aligned to the *S. pombe* genome (Ensembl ASM294v2) using bowtie 1.1.2 with the parameters indicated in the original paper. For each dataset, coverage was divided by a normalization factor, which corresponds to the ratio of the number of alignments over the size of the genome. Note that the length of the reads is unique (48bp) for all samples. Normalized coverage of the IP sample was divided by the normalized coverage of the corresponding input (both increased by 1 to allow taking into account bases with a coverage value of 0 in the input). Thus, any enrichment gives a ratio > 1, and, reciprocally, any impoverishment gives a ratio between 0 and 1.

Statistical methods

Unless otherwise stated, statistics we used two-tailed Wilcoxon and Mann Whitney test when at least 6 independent values were available per sample.

Data access

The MNase-seq data from this publication have been submitted to ArrayExpress database and assigned the identifier E-MTAB-4620.

Acknowledgments

We are very grateful to Vincent Vanoosthuyse for helpful discussions and comments on the manuscript. We thank Dom Helmlinger, Fred Winston, Jean-Paul Javerzat, Susan Forsburg, Blerta

Xhemalce and Michael Keogh for yeast strains, Keith Gull for the anti-tubulin Tat1 antibody, and the EMBL Advanced Light Microscopy Facility for help and advices. This work was supported by funding from the CNRS (ATIP grant to P.B.), the Association pour la Recherche contre le Cancer, grant SFI20111203612 (P.B.), and by a donation from Claude and Antoine Sapone to P.B. Work in Christian Haering's lab was supported by the German Research Foundation grant HA5853/1-2 (C.H.). X.R. was supported by a postdoctoral fellowship from CNRS, L.F. and C.H. by PhD studentships from la Ligue Nationale Contre le Cancer and the Ministère de l'Education Nationale et de la Recherche, respectively. This publication is dedicated to the memory of Antoine Sapone.

Author Contributions

Conceptualization, P.B.; Methodology, C.S., Ch.H. and P.B.; Investigation, E.M., X.R., L.F., C.S., C.K., C.H., P.L., L.N.G., L.M; Formal analysis, S.L., E.C.; Supervision, D.A., Ch.H. and P.B.; Writing original draft, P.B.; Funding Acquisition, P.B.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Aono N, Sutani T, Tomonaga T, Mochida S & Yanagida M (2002) Cnd2 has dual roles in mitotic condensation and interphase. *Nature* **417:** 197–202

Aragon L, Martinez-Perez E & Merkenschlager M (2013) Condensin, cohesin and the control of chromatin states. *Curr Opin Genet Dev* **23:** 204–11

Bahler J, Wu JQ, Longtine MS, Shah NG, McKenzie A 3rd, Steever AB, Wach A, Philippsen P & Pringle JR (1998) Heterologous modules for efficient and versatile PCR-based gene targeting in Schizosaccharomyces pombe. *Yeast* **14:** 943–51

Bahler J, Wyler T, Loidl J & Kohli J (1993) Unusual nuclear structures in meiotic prophase of fission yeast: a cytological analysis. *J. Cell Biol.* **121:** 241–256

Baxter J, Sen N, Martinez VL, De Carandini ME, Schvartzman JB, Diffley JF & Aragon L (2011) Positive supercoiling of mitotic DNA drives decatenation by topoisomerase II in eukaryotes. *Science* **331:** 1328–32

Bonenfant D, Towbin H, Coulot M, Schindler P, Mueller DR & van Oostrum J (2007) Analysis of Dynamic Changes in Post-translational Modifications of Human Histones during Cell Cycle by Mass Spectrometry. *Mol. Cell. Proteomics* **6:** 1917–1932

Bonnet J, Wang C-Y, Baptista T, Vincent SD, Hsiao W-C, Stierle M, Kao C-F, Tora L & Devys D (2014) The SAGA coactivator complex acts on the whole transcribed genome and is required for RNA polymerase II transcription. *Genes Dev.* **28:** 1999–2012

609 Charbin A, Bouchoux C & Uhlmann F (2014) Condensin aids sister chromatid decatenation by
610 topoisomerase II. *Nucleic Acids Res.* **42:** 340–348

611 Choi ES, Stralfors A, Castillo AG, Durand-Dubief M, Ekwall K & Allshire RC (2011) Identification of
612 noncoding transcripts from within CENP-A chromatin at fission yeast centromeres. *J Biol*
613 *Chem* **286:** 23600–7

614 Clemente-Blanco A, Mayan-Santos M, Schneider DA, Machin F, Jarmuz A, Tschochner H & Aragon L
615 (2009) Cdc14 inhibits transcription by RNA polymerase I during anaphase. *Nature* **458:** 219–
616 22

617 Clemente-Blanco A, Sen N, Mayan-Santos M, Sacristan MP, Graham B, Jarmuz A, Giess A, Webb E,
618 Game L, Eick D, Bueno A, Merkenschlager M & Aragon L (2011) Cdc14 phosphatase promotes
619 segregation of telomeres through repression of RNA polymerase II transcription. *Nat Cell Biol*
620 **13:** 1450–6

621 Cuylen S, Metz J & Haering CH (2011) Condensin structures chromosomal DNA through topological
622 links. *Nat Struct Mol Biol* **18:** 894–901

623 D'Ambrosio C, Schmidt CK, Katou Y, Kelly G, Itoh T, Shirahige K & Uhlmann F (2008a) Identification of
624 cis-acting sites for condensin loading onto budding yeast chromosomes. *Genes Dev* **22:** 2215–
625 27

626 D'Ambrosio C, Schmidt CK, Katou Y, Kelly G, Itoh T, Shirahige K & Uhlmann F (2008b) Identification of
627 cis-acting sites for condensin loading onto budding yeast chromosomes. *Genes Dev* **22:** 2215–
628 27

629 Dewari PS & Bhargava P (2014) Genome-wide mapping of yeast histone chaperone anti-silencing
630 function 1 reveals its role in condensin binding with chromatin. *PloS One* **9:** e108652

631 Dowen JM, Bilodeau S, Orlando DA, Hübner MR, Abraham BJ, Spector DL & Young RA (2013) Multiple
632 Structural Maintenance of Chromosome Complexes at Transcriptional Regulatory Elements.
633 *Stem Cell Rep.* **Vol. I:** 371–378

634 Eide T, Carlson C, Taskén KA, Hirano T, Taskén K & Collas P (2002) Distinct but overlapping domains of
635 AKAP95 are implicated in chromosome condensation and condensin targeting. *EMBO Rep.* **3:**
636 426–432

637 Gottesfeld JM & Forbes DJ (1997) Mitotic repression of the transcriptional machinery. *Trends*
638 *Biochem. Sci.* **22:** 197–202

639 Govind CK, Zhang F, Qiu H, Hofmeyer K & Hinnebusch AG (2007) Gcn5 promotes acetylation, eviction,
640 and methylation of nucleosomes in transcribed coding regions. *Mol Cell* **25:** 31–42

641 Helmlinger D, Marguerat S, Villen J, Gygi SP, Bahler J & Winston F (2008) The S. pombe SAGA complex
642 controls the switch from proliferation to sexual differentiation through the opposing roles of
643 its subunits Gcn5 and Spt8. *Genes Dev* **22:** 3184–95

644 Hirano T (2016) Condensin-Based Chromosome Organization from Bacteria to Vertebrates. *Cell* **164:**
645 847–857

646 Hirota T, Gerlich D, Koch B, Ellenberg J & Peters JM (2004) Distinct functions of condensin I and II in
647 mitotic chromosome assembly. *J Cell Sci* **117:** 6435–45

Iwasaki O, Tanizawa H, Kim K-D, Yokoyama Y, Corcoran CJ, Tanaka A, Skordalakes E, Showe LC & Noma K-I (2015) Interaction between TBP and Condensin Drives the Organization and Faithful Segregation of Mitotic Chromosomes. *Mol. Cell* **59:** 755–767

Johnsson A, Durand-Dubief M, Xue-Franzen Y, Ronnerblad M, Ekwall K & Wright A (2009) HAT-HDAC interplay modulates global histone H3K14 acetylation in gene-coding regions during stress. *EMBO Rep* **10:** 1009–14

Johzuka K & Horiuchi T (2007) RNA polymerase I transcription obstructs condensin association with 35S rRNA coding regions and can cause contraction of long repeat in Saccharomyces cerevisiae. *Genes Cells Devoted Mol. Cell. Mech.* **12:** 759–771

Johzuka K & Horiuchi T (2009) The cis element and factors required for condensin recruitment to chromosomes. *Mol. Cell* **34:** 26–35

Kim JH, Zhang T, Wong NC, Davidson N, Maksimovic J, Oshlack A, Earnshaw WC, Kalitsis P & Hudson DF (2013) Condensin I associates with structural and gene regulatory regions in vertebrate chromosomes. *Nat Commun* **4:** 2537

Kimura K & Hirano T (1997) ATP-dependent positive supercoiling of DNA by 13S condensin: a biochemical implication for chromosome condensation. *Cell* **90:** 625–634

Koutelou E, Hirsch CL & Dent SY (2010) Multiple faces of the SAGA complex. *Curr Opin Cell Biol* **22:** 374–82

Kranz AL, Jiao CY, Winterkorn LH, Albritton SE, Kramer M & Ercan S (2013) Genome-wide analysis of condensin binding in Caenorhabditis elegans. *Genome Biol* **14:** R112

Kruhlak MJ, Hendzel MJ, Fischle W, Bertos NR, Hameed S, Yang XJ, Verdin E & Bazett-Jones DP (2001) Regulation of global acetylation in mitosis through loss of histone acetyltransferases and deacetylases from chromatin. *J Biol Chem* **276:** 38307–19

Lantermann A, Strålfors A, Fagerström-Billai F, Korber P & Ekwall K (2009) Genome-wide mapping of nucleosome positions in Schizosaccharomyces pombe. *Methods San Diego Calif* **48:** 218–225

Lantermann AB, Straub T, Strålfors A, Yuan G-C, Ekwall K & Korber P (2010) Schizosaccharomyces pombe genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of Saccharomyces cerevisiae. *Nat. Struct. Mol. Biol.* **17:** 251–257

Lengronne A, Katou Y, Mori S, Yokobayashi S, Kelly GP, Itoh T, Watanabe Y, Shirahige K & Uhlmann F (2004) Cohesin relocation from sites of chromosomal loading to places of convergent transcription. *Nature* **430:** 573–578

Liu W, Tanasa B, Tyurina OV, Zhou TY, Gassmann R, Liu WT, Ohgi KA, Benner C, Garcia-Bassets I, Aggarwal AK, Desai A, Dorrestein PC, Glass CK & Rosenfeld MG (2010) PHF8 mediates histone H4 lysine 20 demethylation events involved in cell cycle progression. *Nature* **466:** 508–12

Liu X, McLeod I, Anderson S, Yates JR. 3rd & He X (2005) Molecular analysis of kinetochore architecture in fission yeast. *Embo J* **24:** 2919–30

Lopez-Serra L, Kelly G, Patel H, Stewart A & Uhlmann F (2014) The Scc2-Scc4 complex acts in sister chromatid cohesion and transcriptional regulation by maintaining nucleosome-free regions. *Nat Genet* **46:** 1147–51

687 Lorch Y, Maier-Davis B & Kornberg RD (2014) Role of DNA sequence in chromatin remodeling and the
688 formation of nucleosome-free regions. *Genes Dev.* **28:** 2492–2497

689 MacCallum DE, Losada A, Kobayashi R & Hirano T (2002) ISWI remodeling complexes in Xenopus egg
690 extracts: identification as major chromosomal components that are regulated by INCENP-
691 aurora B. *Mol Biol Cell* **13:** 25–39

692 Monahan BJ, Villen J, Marguerat S, Bahler J, Gygi SP & Winston F (2008) Fission yeast SWI/SNF and
693 RSC complexes show compositional and functional differences from budding yeast. *Nat*
694 *Struct Mol Biol* **15:** 873–80

695 Moreno S, Klar A & Nurse P (1991) Molecular genetic analysis of fission yeast Schizosaccharomyces
696 pombe. *Methods Enzym.* **194:** 795–823

697 Nakazawa N, Sajiki K, Xu X, Villar-Briones A, Arakawa O & Yanagida M (2015) RNA pol II transcript
698 abundance controls condensin accumulation at mitotically up-regulated and heat-shock-
699 inducible genes in fission yeast. *Genes Cells Devoted Mol. Cell. Mech.* **20:** 481–499

700 Nugent RL, Johnsson A, Fleharty B, Gogol M, Xue-Franzen Y, Seidel C, Wright AP & Forsburg SL (2010)
701 Expression profiling of S. pombe acetyltransferase mutants identifies redundant pathways of
702 gene regulation. *BMC Genomics* **11:** 59

703 Ohta S, Bukowski-Wills JC, Sanchez-Pulido L, Alves Fde L, Wood L, Chen ZA, Platani M, Fischer L,
704 Hudson DF, Ponting CP, Fukagawa T, Earnshaw WC & Rappsilber J (2010) The protein
705 composition of mitotic chromosomes determined using multiclassifier combinatorial
706 proteomics. *Cell* **142:** 810–21

707 Ono T, Losada A, Hirano M, Myers MP, Neuwald AF & Hirano T (2003) Differential contributions of
708 condensin I and condensin II to mitotic chromosome architecture in vertebrate cells. *Cell*
709 **115:** 109–21

710 Orpinell M, Fournier M, Riss A, Nagy Z, Krebs AR, Frontini M & Tora L (2010) The ATAC acetyl
711 transferase complex controls mitotic progression by targeting non-histone substrates. *Embo J*
712 **29:** 2381–94

713 Owen-Hughes T & Gkikopoulos T (2012) Making sense of transcribing chromatin. *Curr. Opin. Cell Biol.*
714 **24:** 296–304

715 Patzlaff JS, Terrenoire E, Turner BM, Earnshaw WC & Paulson JR (2010) Acetylation of core histones
716 in response to HDAC inhibitors is diminished in mitotic HeLa cells. *Exp Cell Res* **316:** 2123–35

717 Petrova B, Dehler S, Kruitwagen T, Heriche JK, Miura K & Haering CH (2013) Quantitative analysis of
718 chromosome condensation in fission yeast. *Mol Cell Biol* **33:** 984–98

719 Piazza I, Rutkowska A, Ori A, Walczak M, Metz J, Pelechano V, Beck M & Haering CH (2014)
720 Association of condensin with chromosomes depends on DNA binding by its HEAT-repeat
721 subunits. *Nat. Struct. Mol. Biol.* **21:** 560–568

722 Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, Bell GW, Walker K, Rolfe PA,
723 Herbolsheimer E, Zeitlinger J, Lewitter F, Gifford DK & Young RA (2005) Genome-wide map of
724 nucleosome acetylation and methylation in yeast. *Cell* **122:** 517–27

Qiu H, Chereji RV, Hu C, Cole HA, Rawal Y, Clark DJ & Hinnebusch AG (2016) Genome-wide cooperation by HAT Gcn5, remodeler SWI/SNF, and chaperone Ydj1 in promoter nucleosome eviction and transcriptional activation. *Genome Res.* **26:** 211–225

Robellet X, Fauque L, Legros P, Mollereau E, Janczarski S, Parrinello H, Desvignes JP, Thevenin M & Bernard P (2014) A genetic screen for functional partners of condensin in fission yeast. *G3 Bethesda* **4:** 373–81

Robert F, Pokholok DK, Hannett NM, Rinaldi NJ, Chandy M, Rolfe A, Workman JL, Gifford DK & Young RA (2004) Global position and recruitment of HATs and HDACs in the yeast genome. *Mol Cell* **16:** 199–209

Roh TY, Cuddapah S & Zhao K (2005) Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev* **19:** 542–52

Sadeghi L, Siggens L, Svensson JP & Ekwall K (2014) Centromeric histone H2B monoubiquitination promotes noncoding transcription and chromatin integrity. *Nat. Struct. Mol. Biol.* **21:** 236–243

Sadeh R & Allis CD (2011) Genome-wide 'Re'-Modeling of Nucleosome Positions. *Cell* **147:** 263–266

Saka Y, Sutani T, Yamashita Y, Saitoh S, Takeuchi M, Nakaseko Y & Yanagida M (1994) Fission yeast cut3 and cut14, members of a ubiquitous protein family, are required for chromosome condensation and segregation in mitosis. *Embo J* **13:** 4938–52

Samejima K, Samejima I, Vagnarelli P, Ogawa H, Vargiu G, Kelly DA, de Lima Alves F, Kerr A, Green LC, Hudson DF, Ohta S, Cooke CA, Farr CJ, Rappsilber J & Earnshaw WC (2012) Mitotic chromosomes are compacted laterally by KIF4 and condensin and axially by topoisomerase IIalpha. *J Cell Biol* **199:** 755–70

Schmidt CK, Brookes N & Uhlmann F (2009) Conserved features of cohesin binding along fission yeast chromosomes. *Genome Biol* **10:** R52

Shintomi K, Takahashi TS & Hirano T (2015) Reconstitution of mitotic chromatids with a minimum set of purified factors. *Nat. Cell Biol.* **17:** 1014–1023

Soriano I, Quintales L & Antequera F (2013) Clustered regulatory elements at nucleosome-depleted regions punctuate a constant nucleosomal landscape in Schizosaccharomyces pombe. *BMC Genomics* **14:** 813

Steen RL, Cubizolles F, Le Guellec K & Collas P (2000) A kinase-anchoring protein (AKAP)95 recruits human chromosome-associated protein (hCAP)-D2/Eg7 for chromosome condensation in mitotic extract. *J. Cell Biol.* **149:** 531–536

Sutani T, Sakata T, Nakato R, Masuda K, Ishibashi M, Yamashita D, Suzuki Y, Hirano T, Bando M & Shirahige K (2015) Condensin targets and reduces unwound DNA structures associated with transcription in mitotic chromosome condensation. *Nat. Commun.* **6:** 7815

Sutani T, Yuasa T, Tomonaga T, Dohmae N, Takio K & Yanagida M (1999) Fission yeast condensin complex: essential roles of non-SMC subunits for condensation and Cdc2 phosphorylation of Cut3/SMC4. *Genes Dev* **13:** 2271–83

763 Tada K, Susumu H, Sakuno T & Watanabe Y (2011) Condensin association with histone H2A shapes
764 mitotic chromosomes. *Nature* **474:** 477–83

765 Teytelman L, Thurtle DM, Rine J & van Oudenaarden A (2013) Highly expressed loci are vulnerable to
766 misleading ChIP localization of multiple unrelated proteins. *Proc. Natl. Acad. Sci. U. S. A.* **110:**
767 18602–18607

768 Thadani R, Uhlmann F & Heeger S (2012) Condensin, chromatin crossbarring and chromosome
769 condensation. *Curr Biol* **22:** R1012-21

770 Valls E, Sanchez-Molina S & Martinez-Balbas MA (2005) Role of histone modifications in marking and
771 activating genes through mitosis. *J Biol Chem* **280:** 42592–600

772 Vanoosthuyse V, Legros P, van der Sar SJA, Yvert G, Toda K, Le Bihan T, Watanabe Y, Hardwick K &
773 Bernard P (2014) CPF-associated phosphatase activity opposes condensin-mediated
774 chromosome condensation. *PLoS Genet.* **10:** e1004415

775 Venters BJ, Wachi S, Mavrich TN, Andersen BE, Jena P, Sinnamon AJ, Jain P, Rolleri NS, Jiang C,
776 Hemeryck-Walsh C & Pugh BF (2011) A comprehensive genomic binding map of gene and
777 chromatin regulatory proteins in Saccharomyces. *Mol Cell* **41:** 480–92

778 Wang F & Higgins JM (2013) Histone modifications and mitosis: countermarks, landmarks, and
779 bookmarks. *Trends Cell Biol* **23:** 175–84

780 Wang Y, Kallgren SP, Reddy BD, Kuntz K, Lopez-Maury L, Thompson J, Watt S, Ma C, Hou H, Shi Y,
781 Yates JR, Bahler J, O'Connell MJ & Jia S (2012) Histone H3 Lysine 14 Acetylation Is Required
782 for Activation of a DNA Damage Checkpoint in Fission Yeast. *J. Biol. Chem.* **287:** 4386–4393

783 Weake VM & Workman JL (2012) SAGA function in tissue-specific gene expression. *Trends Cell Biol.*
784 **22:** 177–184

785 Wiren M, Silverstein RA, Sinha I, Walfridsson J, Lee HM, Laurenson P, Pillus L, Robyr D, Grunstein M &
786 Ekwall K (2005) Genomewide analysis of nucleosome density histone acetylation and HDAC
787 function in fission yeast. *Embo J* **24:** 2906–18

788 Xue-Franzen Y, Henriksson J, Burglin TR & Wright AP (2013) Distinct roles of the Gcn5 histone
789 acetyltransferase revealed during transient stress-induced reprogramming of the genome.
790 *BMC Genomics* **14:** 479

791

792

Figure Legends

Figure 1. Gcn5 is required for mitotic chromosome condensation

A. Genetic interaction between condensin and Gcn5. Fivefold serial dilutions of fission yeast strains were spotted onto complete medium. **B** Scheme of Gcn5-47. KAT: Lysine Acetyl-Transferase. Bromo: Bromodomain. **C.** Condensation assay. 3D distances between two fluorescently-marked loci on chromosome I were measured by live cell microscopy in fission yeast cells progressing from late G2 through mitosis. Time lapse recording for n>32 cells for one (wild type) or two (*gcn5-47* and *cut14-208*) biological replicates were aligned to anaphase onset (t = 0) and average distances (± s.d.) plotted. Values are listed in Appendix Table S1. **D.** Chromosome segregation in *gcn5-47* or *cut3-477* mutant cells. Cells growing at 32°C were fixed and processed for immunofluorescence against Mis6-HA and Cdc11-GFP to reveal centromeres (Cen) and Spindle Pole Bodies (SPB), respectively. DNA was stained with 4',6-diamidino-2-phenylindole (DAPI). Left panel: single mutant cells *cut3-477* or *gcn5-47* exhibiting chromatin bridges (b) or trailing chromatin (t). Bar: 5 µm. Right panel: frequencies of chromatin bridges and trailing chromatin as a function of spindle length (SPB-SPB distance, n> 40 for each category). **E.** Spindle lengths in cells showing chromatin bridges. Boxes indicate 25th, median and 75th percentile. Whiskers are the minimum and maximum (n>50). *** P<0.001.

Figure 2. The link between condensin and Gcn5 relies on its ability to acetylate nucleosomes

Genetic interactions between condensin and **(A)** subunits of the SAGA complex, **(B)** catalytically inactive Gcn5, **(C)** multiples HATs, or **(D)** Gcn5 and Mst2. Fivefold serial dilutions of fission yeast strains were spotted onto complete medium.

Figure 3. Gcn5 and Mst2 assists condensin binding to chromatin in mitosis

A. Condensin binding assessed by ChIP against Cnd2-GFP. Mitotic indexes are indicated in parentheses. %IP are averages and s.d. calculated from 12 ChIPs on 6 biological replicates. For repeated 5S and gly05 genes, qPCR primers were designed within adjacent, unique 5' intergenic sequences. Note the use of different scales in the arm: high-occupancy and arm: low-occupancy panels. *** P<0.001, ** P<0.01, ° P>0.05. **B.** The level of Cnd2-GFP by western blotting. **C.** Condensin binding assessed by chromosome spreading. Cut3-HA immunofluorescence signals were quantified on n> 100 chromosome spreads exhibiting Dam1-GFP at kinetochores (mitotic, M) or lacking Dam1-GFP (interphase, I). For each strain, four representative nuclei extracted from a same image acquired with same settings are shown. Asterisks indicate nuclei corresponding to Cut3-HA surface plots. Bar:

5 µm. Boxes indicate the 25th, median and 75th percentile and whiskers the min and max values from 3 independent experiments. *** P<0.001. **D.** Cut3-HA level by western blotting.

Figure 4. Gcn5 and acetylated H3 persist at condensin binding sites in mitosis

A. Gcn5 binding to chromosomes assessed by chromosome spreading. Fission yeast *nda3-KM311* cells were shifted at 19°C for 4 hours to enrich the population for the early stages of mitosis, and processed for chromosome spreading and immunofluorescence against Gcn5-myc and Cut14-HA. The Cut14/DAPI ratio served as marker for mitotic progression, with interphase (i): ratio<0.5; prophase (p): 0.5<ratio<1.5; and prometaphase/metaphase (pm): ratio>1.5. For each strain, four representative nuclei extracted from a same image acquired with same settings are shown. Bar: 5 µm. Boxes indicate the 25th, median and 75th percentile and whiskers the min and max values (n>150 nuclei per strain). **B.** ChIP against Gcn5-myc on cycling cells or cells arrested in pro/metaphase. Mitotic indexes in parentheses were determined by scoring binucleated cells and hypercondensed nuclei. % IP are averages and s.d. calculated from 6 ChIPs on 3 biological replicates. **C.** Total H3 and acetylated isoforms in cycling or mitotically arrested cells. Whole cell extracts (WCE) prepared from cycling cells or prometaphase cells (mitotic index 89%) were loaded on a same gel and simultaneously revealed. **D-E.** Chromosomal occupancy of acetylated H3 during mitosis. Cells expressing Cnd2-GFP were blocked in early mitosis, mitotic indexes measured (D), and processed for ChIP against total H3, H3K9ac or H3K18ac. Heterochromatin (*dh1*) served as negative control. H3ac/H3 average ratios and s.d. calculated from 3 ChIPs on 3 biological replicates are shown. *** P<0.001, * P<0.05, °P>0.05. High-occup. and Low-occup. refer to condensin high-occupancy and low-occupancy binding sites, respectively.

Figure 5. Condensin accumulates at nucleosome depleted regions dependent upon Gcn5

A-B. Nucleosome occupancy at condensin binding sites during mitosis. Fission yeast cells expressing Cnd2-GFP were arrested in mitosis, mitotic indexes determined (A), and cells processed for MNase-seq. (B) NDRs are shown in purple and condensin binding sites in blue with a star indicating the peak maximum. Nucleosome patterns shown are representative examples of 3 biological replicates. **C.** Permutation test. 100,000 sets of 48 DNA segments of the same size of condensin binding sites were randomly drawn from the genome and the number of bp associated with a NDR by chance compared with the experimental value (purple bar). *** P<10^{-5}. **D.** Metagene analysis of nucleosome occupancy within the NDR and positioning of the flanking nucleosomes. **E.** Coverage depth over NDR sequences plus upstream and downstream 200 bp. Boxes indicate the 25th, median and 75th percentile, whiskers indicate the upper and lower 1.5 IQR and circles outliers. *** P=10^{-14} by Kruskal-Wallis non parametric test. **F.** Cells arrested in mitosis were processed for ChIP against histone H3. Mitotic

indexes are indicated in parentheses. %IP are averages and s.d. calculated from 6 ChIPs on 3 biological replicates. *** P<0.001, ** P<0.01, ° P>0.05. High-occup. and Low-occup. refer to condensin high-occupancy and low-occupancy binding sites, respectively.

Figure 6. Nucleosome eviction by the RSC remodeling complex facilitates condensin binding

A-C. Nucleosome occupancy and condensin binding during mitosis upon RSC loss of function. Fission yeast cells expressing Cnd2-GFP were arrested in early mitosis, mitotic indexes determined (A), and cells processed for ChIP against H3 and H4 (B) and Cnd2-GFP (C). Averages and s.d. calculated from 3 ChIPs per antigen, each performed on 3 biological replicates. Values for *prl53* and *ecm33* were multiplied by 4 to facilitate reading. **D.** Cnd2-GFP levels by western blotting. **E-F.** RSC localization at condensin binding sites during mitosis. Mitotically-arrested cells were subjected to ChIP against Snf21-FLAG. Averages and s.d. calculated from 12 ChIPs on 6 biological replicates are indicated. Values for *5S* and *gly05* were multiplied by 3. *** P<0.001, ** P<0.01, ° P>0.05. High-occup. and Low-occup. refer to condensin high-occupancy and low-occupancy binding sites, respectively.

Expended View Figure Legends

Figure EV1. Chromosomes III remain untangled during anaphase in *gcn5* mutant cells

Scheme of chromosome III in which ~100 copies of 10 kb rDNA repeats, each consisting of 5.8S, 18S and 28S rDNA genes, are located adjacent to the telomeres. Fib1, which binds the rDNA repeats, was used to monitor the segregation of the arms of chromosome III in late anaphase cells. Fission yeast cells exponentially growing at 32°C and expressing Fib1-RFP (rDNA) and Cdc11-GFP (SPB) were fixed, stained with DAPI and examined for chromosome segregation (DAPI) and rDNA segregation (n>70). Bar: 5 microns.

Figue EV2. The link between Gcn5 and Cut3 is direct

A. Steady state level of condensin and Topo II mRNA in cells lacking functional Gcn5. 500 ng of total RNA extracted from fission yeast cells exponentially growing at 32°C was reverse-transcribed in the presence (+) or absence (-) of Reverse Transcriptase (RT) and cDNAs were quantified by real time qPCR. Indicated values correspond to the average and mean deviation from two independent experiments. **B.** Steady state levels of Cut3-GFP and Top2-HA detected by Western blotting. α-Tubulin (Tub) was used a loading control. **C.** Integrity of the condensin complex as judged by co-

immunoprecipitation. Cnd2-GFP was immunoprecipitated from indicated strains arrested in mitosis (septation indexes < 4%). Levels of Cnd2-GFP and Cut3-HA in total and immunoprecipitated fractions were assessed by Western blotting. **D-E.** Steady state level of *cnp3* mRNA measured by RT-qPCR. 500 ng of total RNA was reverse-transcribed in the presence (+) or absence (-) of RT. pCNP3 indicates that the eponym plasmid bearing the *cnd3* gene was inserted into the genome. **F.** Restoring *cnp3* mRNA level does not suppress the negative genetic interaction between *gcn5-47* and *cut3-477*. Cells of indicated genotype were serially diluted fivefold and spotted onto complete media supplemented with phloxin B.

Figure EV3. Condensin association sites in fission yeast and their transcription

A. Condensin binding assessed by ChIP. Fission yeast cells were arrested in pro/metaphase at 19°C by the *nda3-KM311* mutation and processed for ChIP against Cnd2-GFP. %IP correspond to the averages and s.d. calculated from 6 ChIPs performed on 3 biological replicates. **B**. Mitotic indexes. **C.** Steady state level of Cnd2-GFP. Exponentially growing cells were shifted at 36°C for 2.5 hours to inactive *cut3-477*, and whole cell extracts assessed for Cnd2-GFP levels by western blotting using anti-GFP (A11122) antibody. Tubulin (Tub) served as loading control. **D-E.** RNA Pol II occupancy assessed by ChIP in mitotically-arrested cells. Cells expressing Cnd2-GFP were arrested in mitosis and processed for ChIP against RNA Pol II phosphorylated on Serine 2 (Ser2P). %IP correspond to averages and s.d. calculated from 8 ChIPs performed on 4 biological replicates. (E). Mitotic indexes. *** P<0.001, ** P<0.01, * P<0.05 and ° P>0.05.

Figure EV4. Gcn5 and Mst2 are not required for the chromosomal association of Top2 and Mis4

The chromosomal association of Topo II (**A**) or of the cohesin loader Mis4 (**B**) assessed by chromosome spreading. Fission yeast cells exponentially growing at 32°C were processed for chromosome spreading and immunofluorescence against Dam1-GFP and Top2-HA or Mis4-HA. Chromatin was stained with DAPI. Dam1-GFP, recruited at kinetochores during mitosis, was used as a mitotic marker of isolated nuclei. Top2-HA or Mis4-HA signals associated with interphase (Dam1-GFP negative) or mitotic nuclei (Dam1-GFP positive) were quantified and divided by the intensity of the DAPI signal. Ratios were normalized to the median value given by the wt control in mitosis. Boxes indicate the 25^{th} percentile, median and 75^{th} percentile and whiskers the min and max values calculated from n=3 experiments. ° P>0.05 and * P<0.05. Bars: 5 microns.

Figure EV5. Gcn5 localization across transcribed genes during mitosis

A. ChIP against Gcn5-myc on fission yeast cells cycling or arrested in prometaphase. Mitotic indexes indicated in parentheses were determined by scoring binucleated cells and hypercondensed

nuclei. % IP are averages and s.d. calculated from 6 ChIPs performed on 3 biological replicates. **B**. The chromosomal occupancy of H3K14ac during mitosis. Cells expressing Cnd2-GFP were blocked in early mitosis, mitotic indexes measured (see Fig. 4D), and cells processed for ChIP against total H3 of H3K14ac. Transcriptionally silent pericentric heterochromatin (dh1) served as negative control. H3ac/H3 average ratios and standard deviations calculated from 6 ChIPs performed on 3 biological replicates are shown.

Figure EV6. Condensin accumulates at nucleosome depleted regions dependent upon Gcn5

A. Fission yeast cells expressing Cnd2-GFP were arrested in early mitosis, mitotic indexes determined (see Fig. 5A) and processed for MNase digestion. Mitotic chromatin was digested with increasing amounts of MNase to obtain mononucleosomes to dinucleosomes ratios of ~80:20. Mononucleosomal DNA excised from the gel was subjected to massive parallel sequencing (MNase-seq). **B.** Overlap between nucleosome depleted regions (NDR) or non-NDRs identified during mitosis by MNase-seq and condensin peaks identified by ChIP-seq against Cut14-pk9 (Sutani et al. 2015). NDRs (red) are enriched in the population of DNA fragments co-immunoprecipitated with Cut14-pk9 (IP/Input >1), and depleted from the flow through (IP/Input <1). **C.** Nucleosome scanning (MNase-qPCR) assay for nucleosome occupancy at the *prl53* promoter. Mitotic chromatin was digested by MNase as described in (A). Input (undigested) and mononucleosomal DNA was purified and the % of input DNA which was protected from MNase digestion was assessed by qPCR. Averages and s.d. calculated from 3 experiments on 3 biological replicates are shown. Nucleosome patterns obtained by MNase-seq analysis (related to Fig. 5B) are indicated for comparison. **D.** Representative MNase-seq nucleosome patterns at the 5S rRNA, *uge1* and *cnd1* genes (see Fig. 5B). Lack of Gcn5 or both Gcn5 and Mst2 does not significantly modifies nucleosome occupancy at these three low-occupancy condensin binding sites.

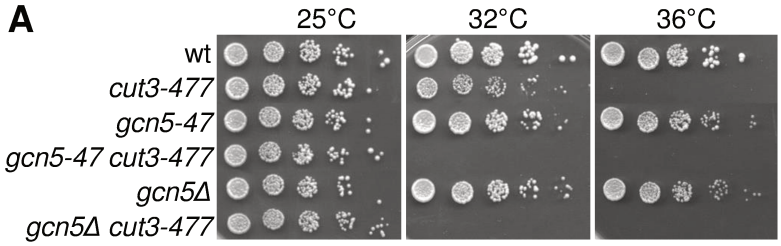
Figure EV7. MNase-seq patterns at three high-occupancy condensin binding sites in replicates

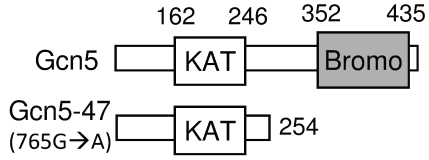
Related to Fig. 5A and EV6. MNase-seq nucleosome patterns at the *snoU14*, *ecm33* and *cdc22* genes in the three wt, *gcn5Δ* or *gcn5Δ mst2Δ* biological and technical replicates. For each gene and each replicate, results shown in Fig. 5A and EV6 correspond to the first line.

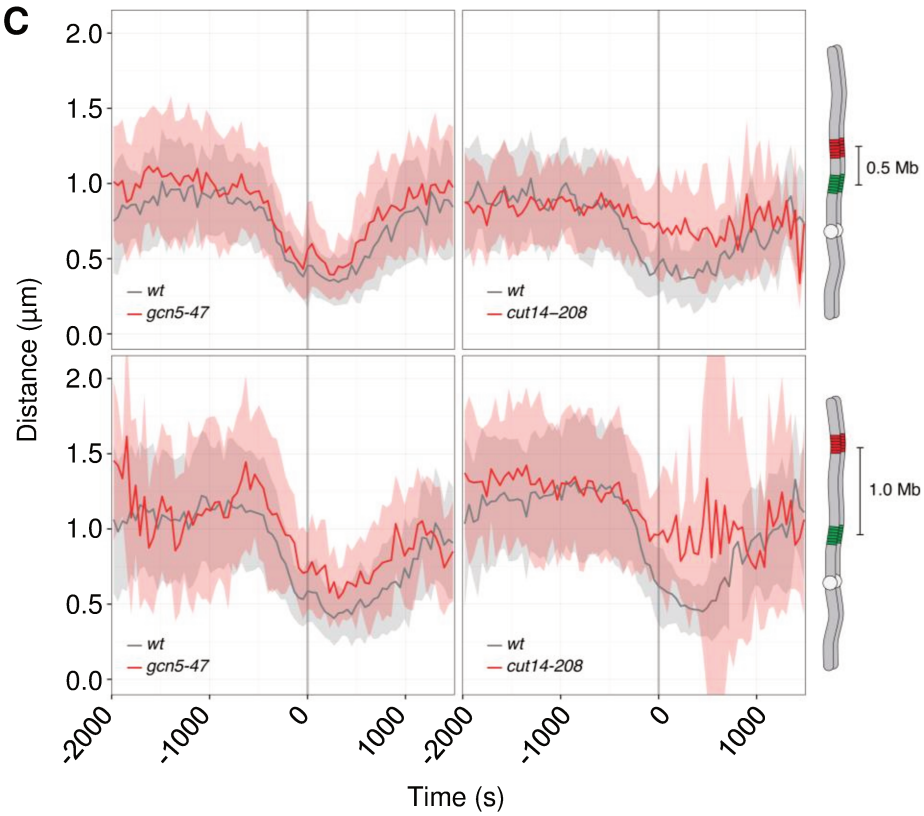
Figure EV8. The impact of RSC loss of function on histone H3 occupancy at the 3' end of genes bound by condensin.

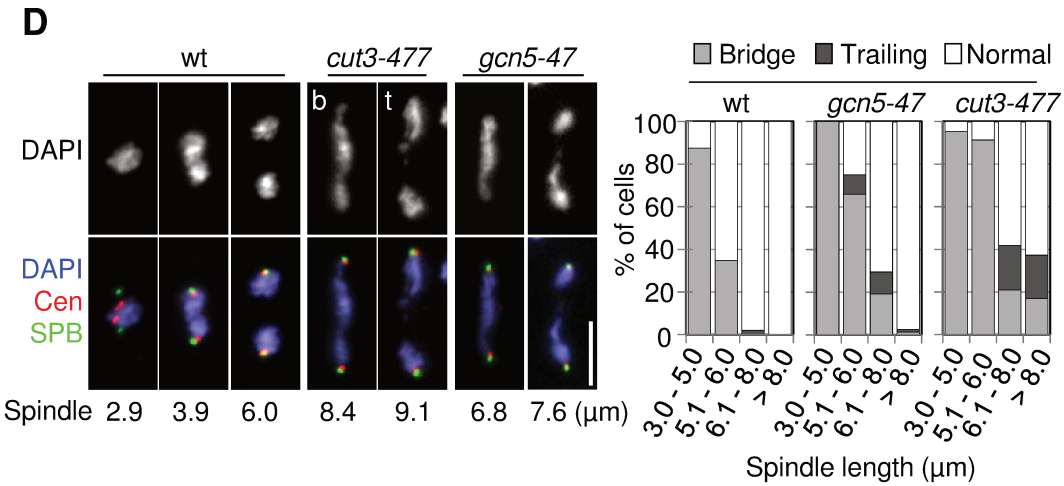
Fission yeast cells expressing Cnd2-GFP were arrested in early mitosis, mitotic indexes determined (see Fig. 6A), and cells processed for ChIP against H3. % IP correspond to averages and s.d. calculated from 3 ChIPs performed on 3 biological replicates. See Fig. 6 for additional information.

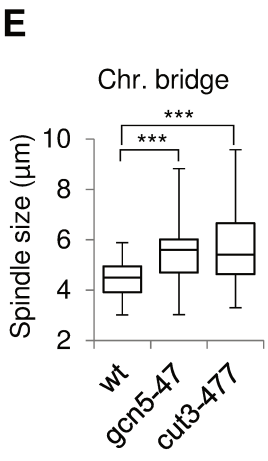
965

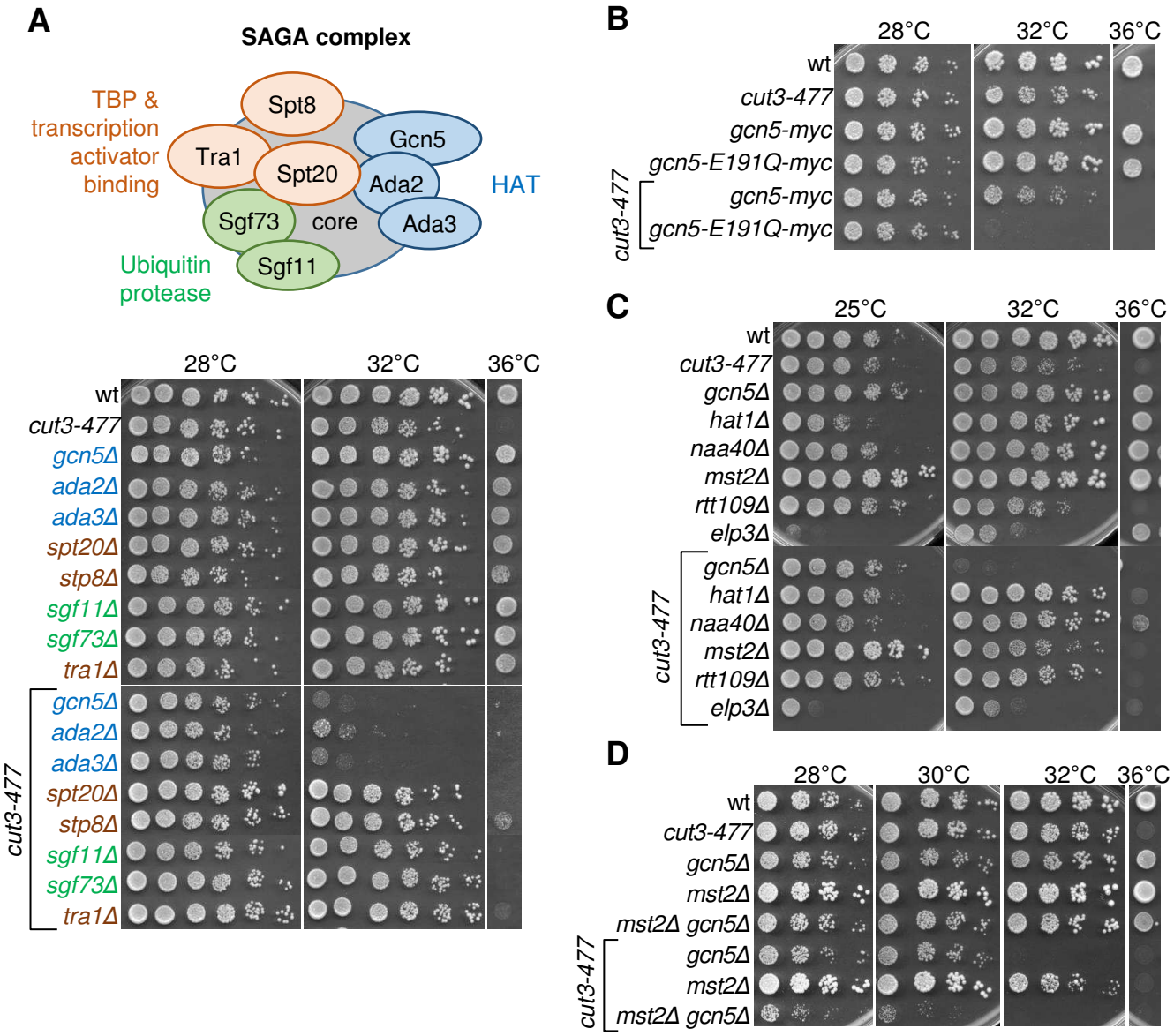

A
25°C
32°C
36°C
wt
cut3-477
gcn5-47
gcn5-47 cut3-477
gcn5Δ
gcn5Δ cut3-477

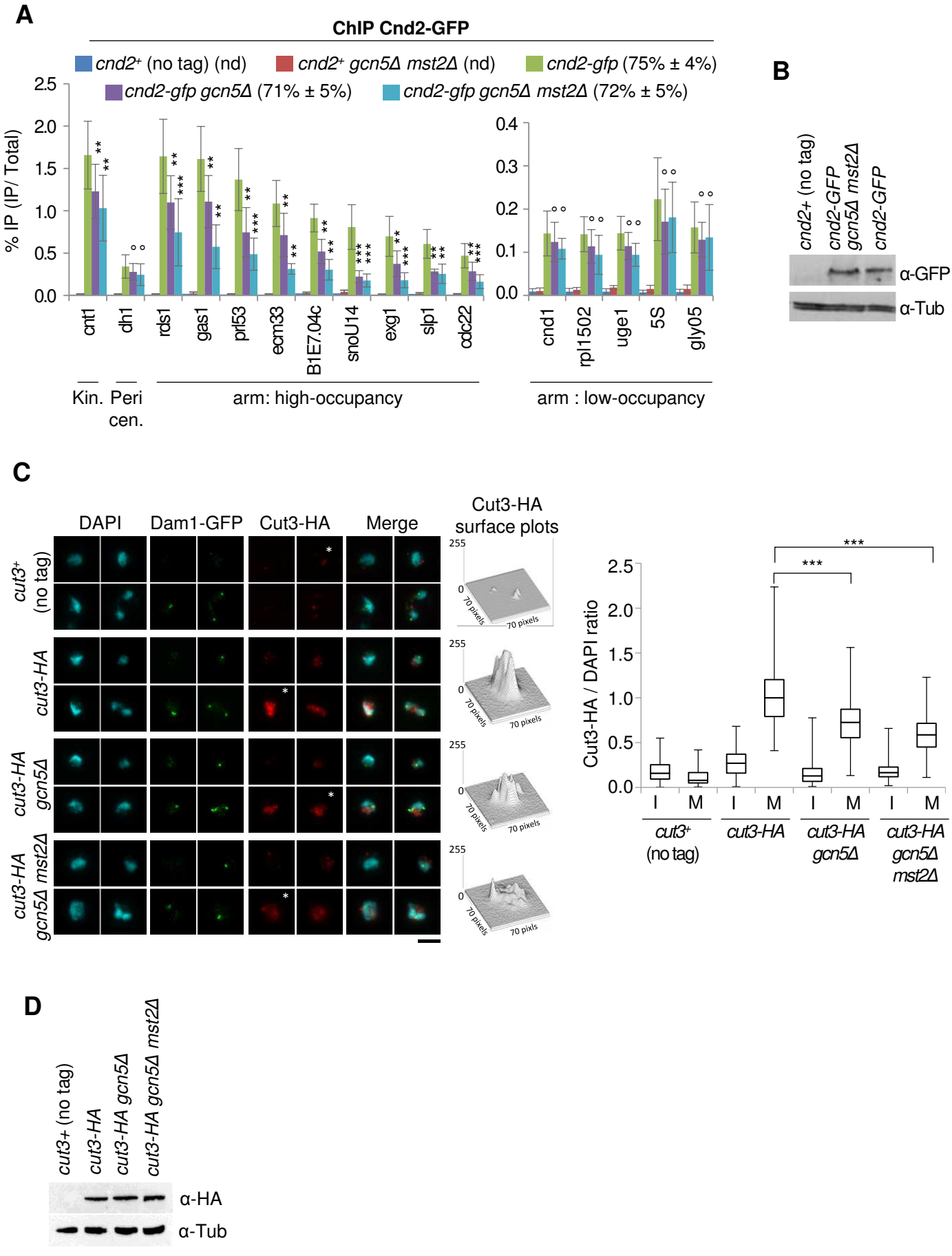

B
162
246
352
435
Gcn5
KAT
Bromo
Gcn5-47
(765G→A)
KAT
254

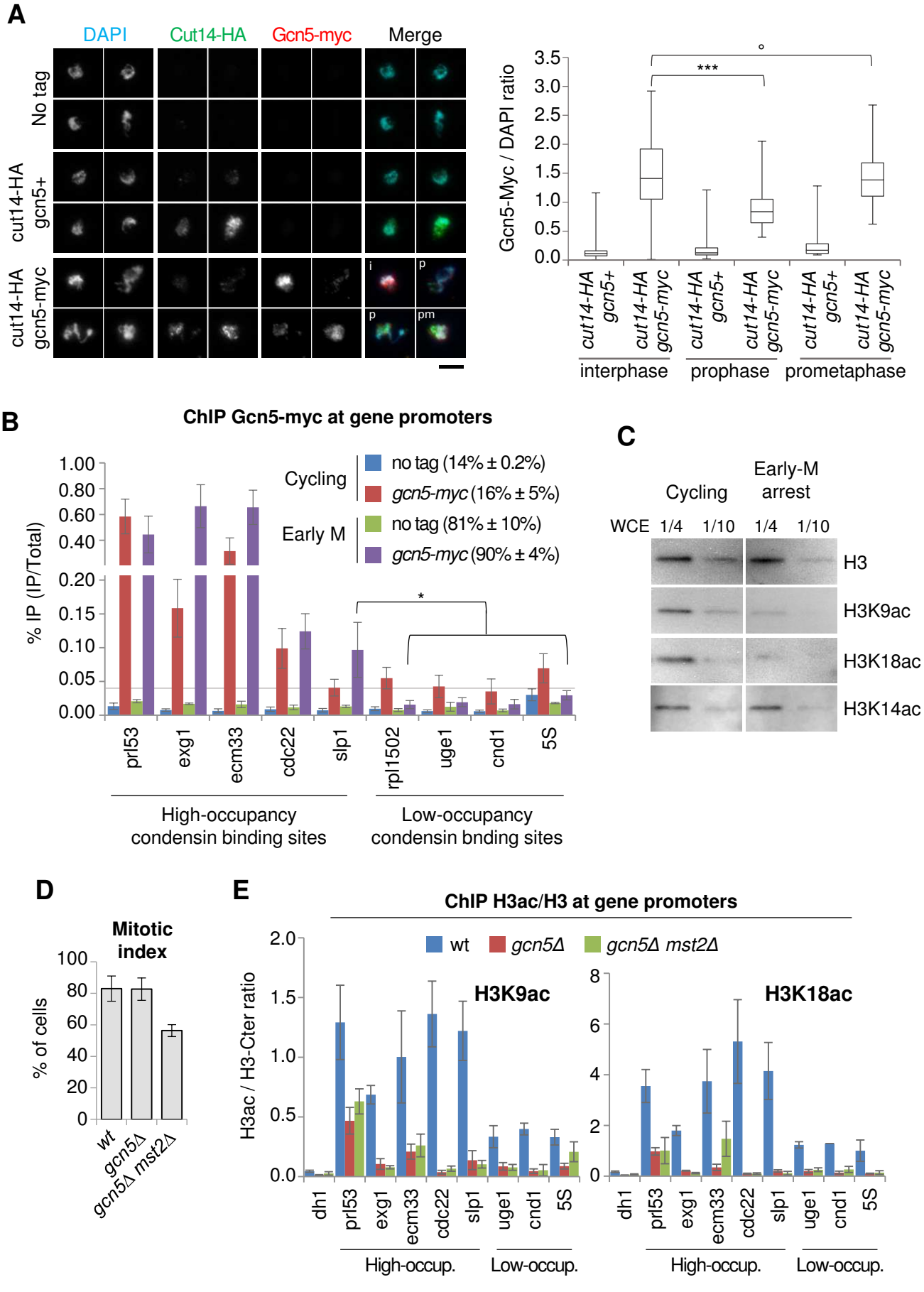

C
Distance (µm)
2.0
1.5
1.0
0.5
0.0
wt
gcn5-47
wt
cut14–208
0.5 Mb
wt
gcn5-47
wt
cut14-208
1.0 Mb
-2000
-1000
0
1000
Time (s)

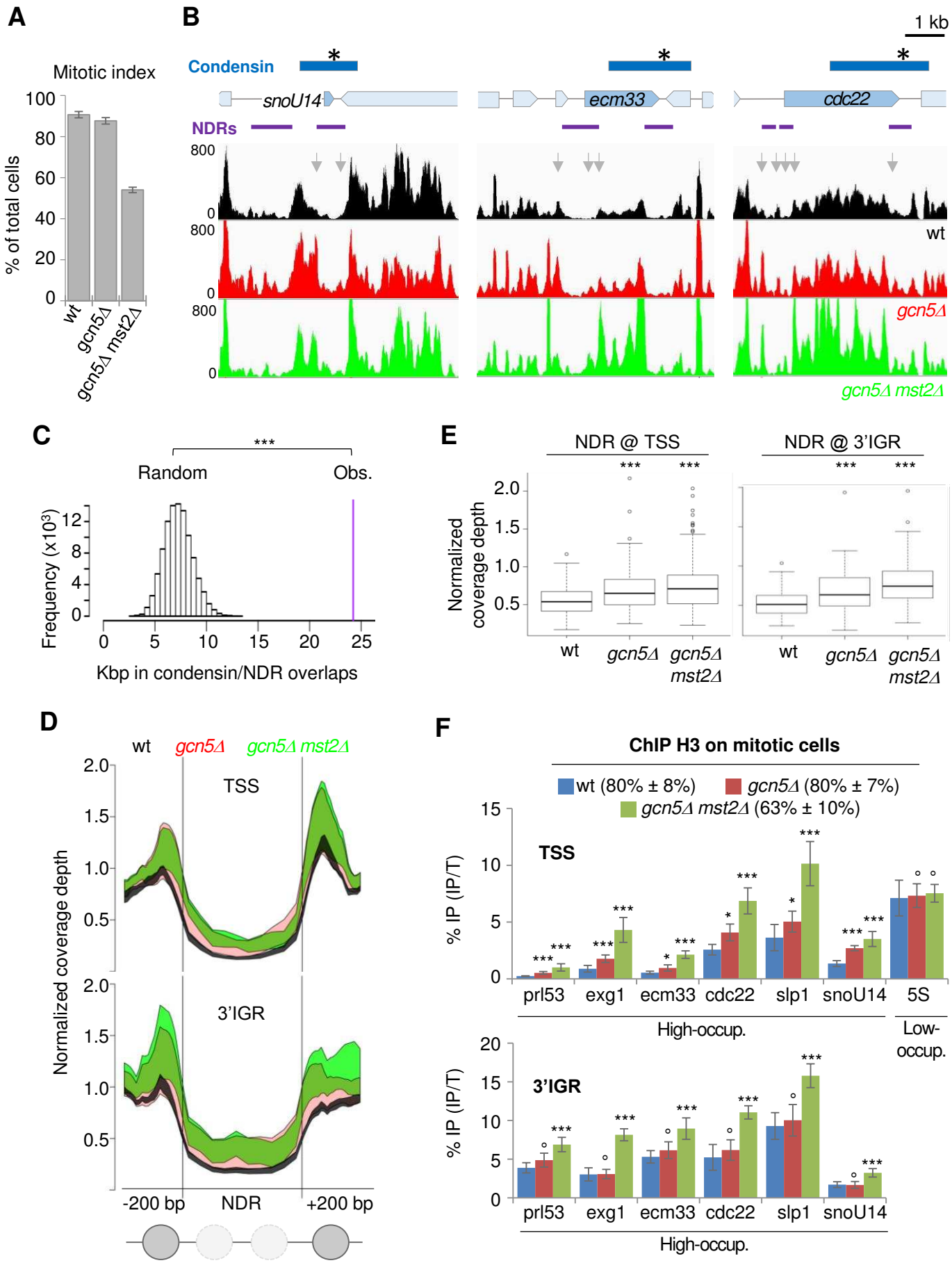

D
wt
cut3-477
gcn5-47
DAPI
DAPI
Cen
SPB
b
t
Spindle 2.9 3.9 6.0 8.4 9.1 6.8 7.6 (µm)
Bridge
Trailing
Normal
wt
gcn5-47
cut3-477
% of cells
100
80
60
40
20
0
3.0 - 5.0
5.1 - 6.0
6.1 - 8.0
> 8.0
Spindle length (µm)

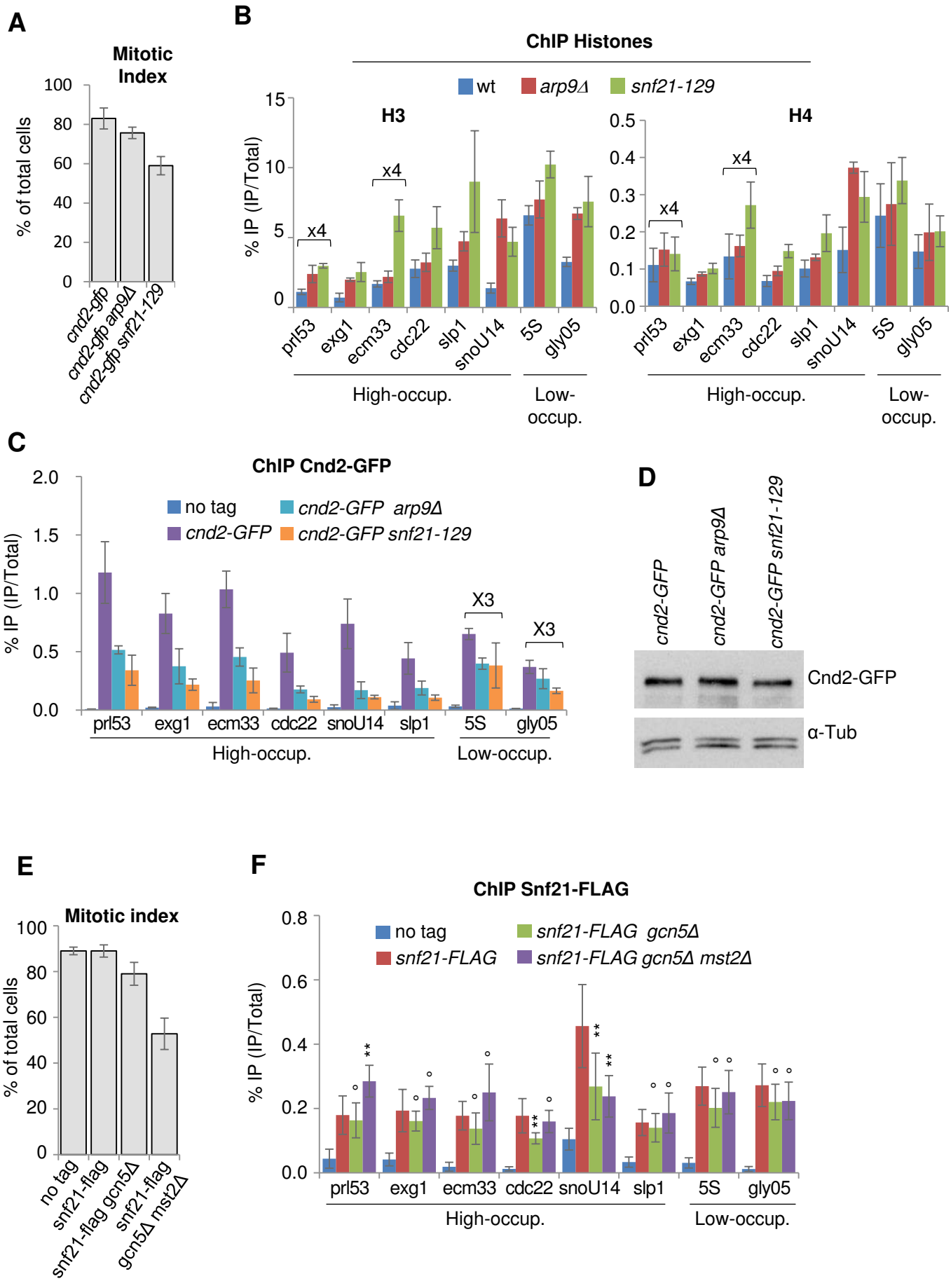

E
Chr. bridge
***
***
Spindle size (µm)
10
8
6
4
2
wt
gcn5-47
cut3-477

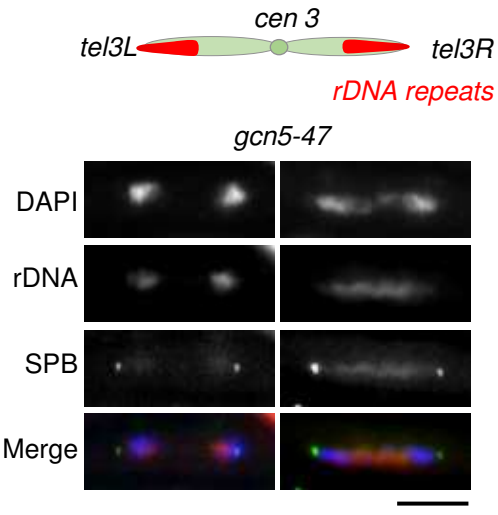

A
SAGA complex
TBP & transcription activator binding
Spt8
Tra1
Spt20
Gcn5
Ada2
HAT
Sgf73
core
Ada3
Ubiquitin protease
Sgf11
28°C
32°C
36°C
wt
cut3-477
gcn5Δ
ada2Δ
ada3Δ
spt20Δ
stp8Δ
sgf11Δ
sgf73Δ
tra1Δ
cut3-477
B
28°C
32°C
36°C
wt
cut3-477
gcn5-myc
gcn5-E191Q-myc
cut3-477
gcn5-myc
gcn5-E191Q-myc
C
25°C
32°C
36°C
wt
cut3-477
gcn5Δ
hat1Δ
naa40Δ
mst2Δ
rtt109Δ
elp3Δ
cut3-477
D
28°C
30°C
32°C
36°C
wt
cut3-477
gcn5Δ
mst2Δ
mst2Δ gcn5Δ
cut3-477
gcn5Δ
mst2Δ
mst2Δ gcn5Δ

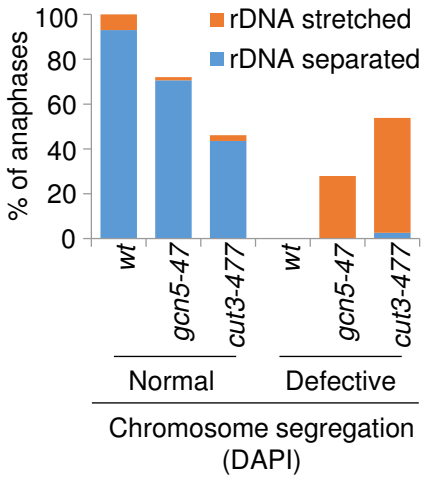

A
ChIP Cnd2-GFP
cnd2+ (no tag) (nd)
cnd2+ gcn5Δ mst2Δ (nd)
cnd2-gfp (75% ± 4%)
cnd2-gfp gcn5Δ (71% ± 5%)
cnd2-gfp gcn5Δ mst2Δ (72% ± 5%)
% IP (IP/ Total)
cnt1
dh1
rds1
gas1
prl53
ecm33
B1E7.04c
snoU14
exg1
slp1
cdc22
cnd1
rpl1502
uge1
5S
gly05
Kin.
Peri cen.
arm: high-occupancy
arm : low-occupancy
B
cnd2+ (no tag)
cnd2-GFP
cnd2-GFP gcn5Δ mst2Δ
α-GFP
α-Tub
C
DAPI
Dam1-GFP
Cut3-HA
Merge
Cut3-HA surface plots
cut3+ (no tag)
cut3-HA
cut3-HA gcn5Δ
cut3-HA gcn5Δ mst2Δ
70 pixels
Cut3-HA / DAPI ratio
I
M
D
cut3+ (no tag)
cut3-HA
cut3-HA gcn5Δ
cut3-HA gcn5Δ mst2Δ
α-HA
α-Tub

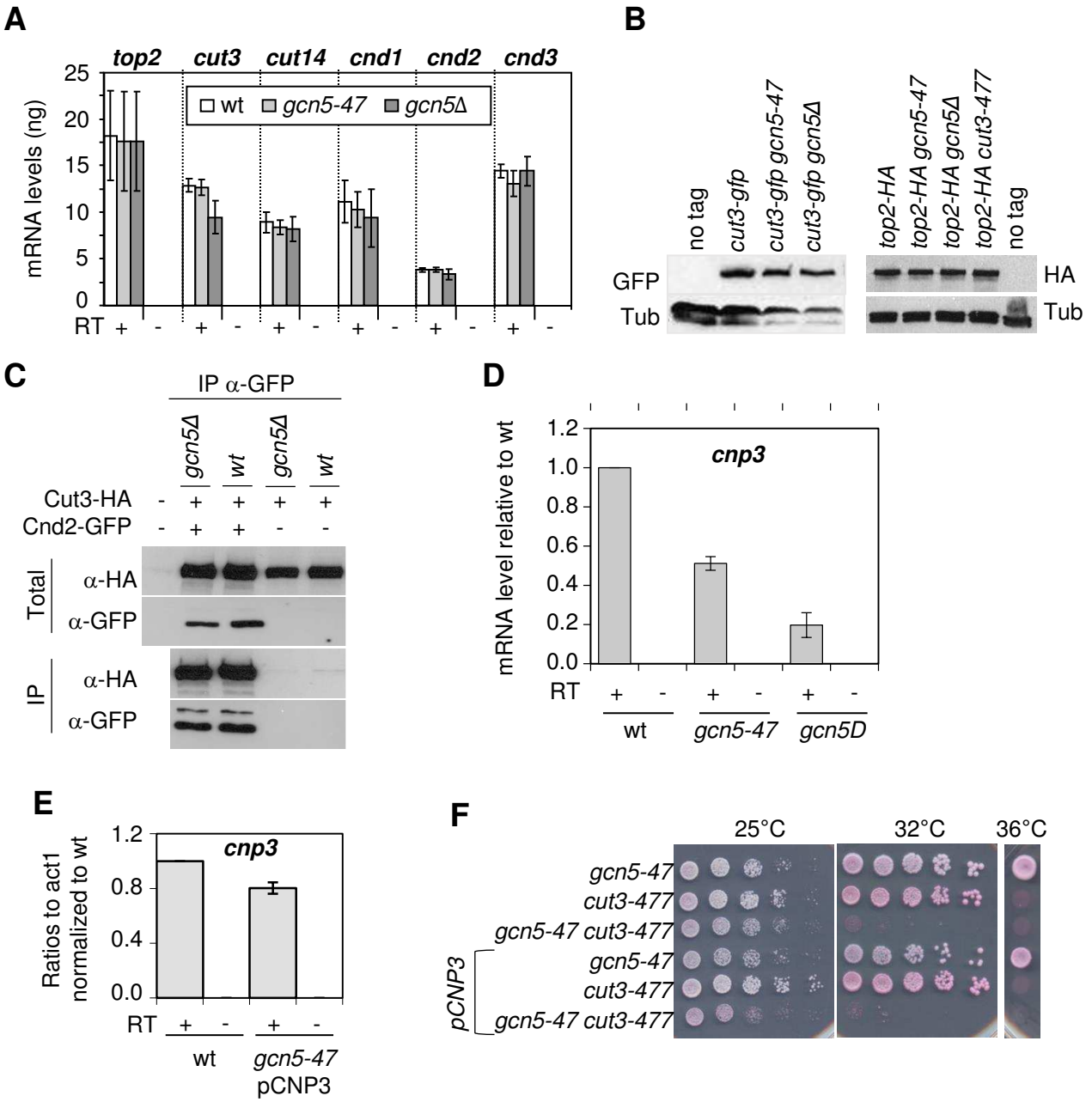

A
DAPI
Cut14-HA
Gcn5-myc
Merge
No tag
cut14-HA gcn5+
cut14-HA gcn5-myc
i
p
p
pm
Gcn5-Myc / DAPI ratio
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
***
°
cut14-HA gcn5+
cut14-HA gcn5-myc
interphase
prophase
prometaphase
B
ChIP Gcn5-myc at gene promoters
% IP (IP/Total)
1.00
0.80
0.60
0.40
0.20
0.15
0.10
0.05
0.00
Cycling
no tag (14% ± 0.2%)
gcn5-myc (16% ± 5%)
Early M
no tag (81% ± 10%)
gcn5-myc (90% ± 4%)
*
prl53
exg1
ecm33
cdc22
slp1
rpl1502
uge1
cnd1
5S
High-occupancy condensin binding sites
Low-occupancy condensin bnding sites
C
Cycling
Early-M arrest
WCE 1/4 1/10 1/4 1/10
H3
H3K9ac
H3K18ac
H3K14ac
D
Mitotic index
% of cells
100
80
60
40
20
0
wt
gcn5Δ
gcn5Δ mst2Δ
E
ChIP H3ac/H3 at gene promoters
wt
gcn5Δ
gcn5Δ mst2Δ
H3ac / H3-Cter ratio
2.0
1.5
1.0
0.5
0.0
H3K9ac
H3K18ac
8
6
4
2
0
dh1
prl53
exg1
ecm33
cdc22
slp1
uge1
cnd1
5S
High-occup.
Low-occup.

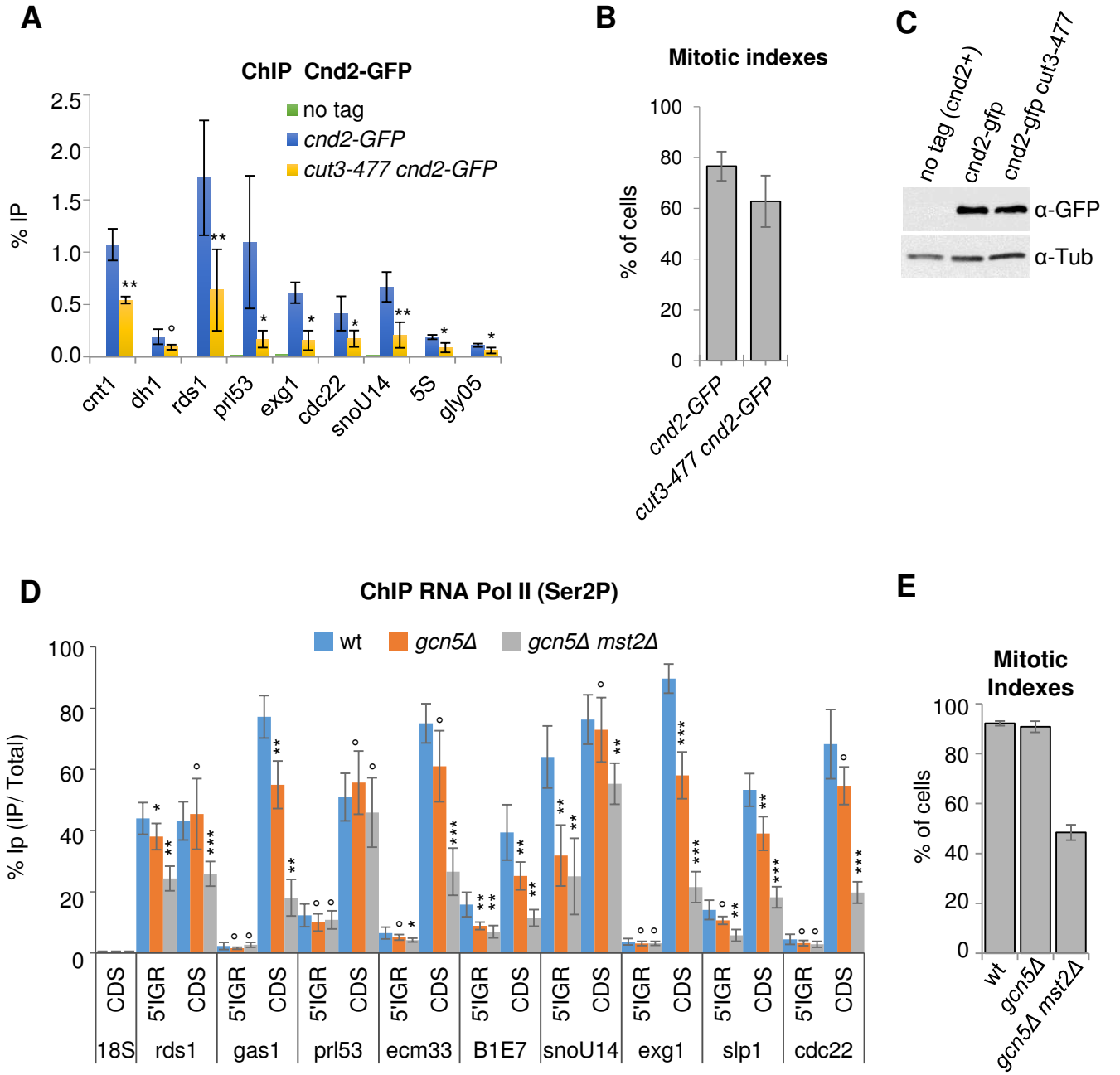

A
Mitotic index
% of total cells
100
80
60
40
20
0
wt
gcn5Δ
gcn5Δ mst2Δ
B
1 kb
Condensin
snoU14
ecm33
cdc22
NDRs
800
0
wt
gcn5Δ
gcn5Δ mst2Δ
C
***
Random
Obs.
Frequency (x10³)
12
8
4
0
0
5
10
15
20
25
Kbp in condensin/NDR overlaps
D
wt
gcn5Δ
gcn5Δ mst2Δ
TSS
3'IGR
Normalized coverage depth
2.0
1.5
1.0
0.5
-200 bp
NDR
+200 bp
E
NDR @ TSS
NDR @ 3'IGR
***
Normalized coverage depth
2.0
1.5
1.0
0.5
wt
gcn5Δ
gcn5Δ mst2Δ
F
ChIP H3 on mitotic cells
wt (80% ± 8%)
gcn5Δ (80% ± 7%)
gcn5Δ mst2Δ (63% ± 10%)
TSS
% IP (IP/T)
15
10
5
0
prl53
exg1
ecm33
cdc22
slp1
snoU14
5S
High-occup.
Low-occup.
3'IGR
% IP (IP/T)
20
15
10
5
0
prl53
exg1
ecm33
cdc22
slp1
snoU14
High-occup.


A
Mitotic Index
% of total cells
cnd2-gfp
cnd2-gfp arp9Δ
cnd2-gfp snf21-129
B
ChIP Histones
wt
arp9Δ
snf21-129
H3
H4
% IP (IP/Total)
x4
prl53
exg1
ecm33
cdc22
slp1
snoU14
5S
gly05
High-occup.
Low-occup.
C
ChIP Cnd2-GFP
no tag
cnd2-GFP
cnd2-GFP arp9Δ
cnd2-GFP snf21-129
% IP (IP/Total)
X3
D
cnd2-GFP
cnd2-GFP arp9Δ
cnd2-GFP snf21-129
Cnd2-GFP
α-Tub
E
Mitotic index
% of total cells
no tag
snf21-flag
snf21-flag gcn5Δ
snf21-flag gcn5Δ mst2Δ
F
ChIP Snf21-FLAG
no tag
snf21-FLAG
snf21-FLAG gcn5Δ
snf21-FLAG gcn5Δ mst2Δ
% IP (IP/Total)

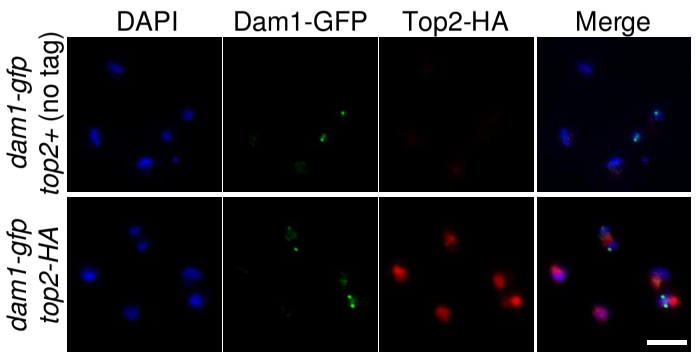

cen 3
tel3L
tel3R
rDNA repeats
gcn5-47
DAPI
rDNA
SPB
Merge

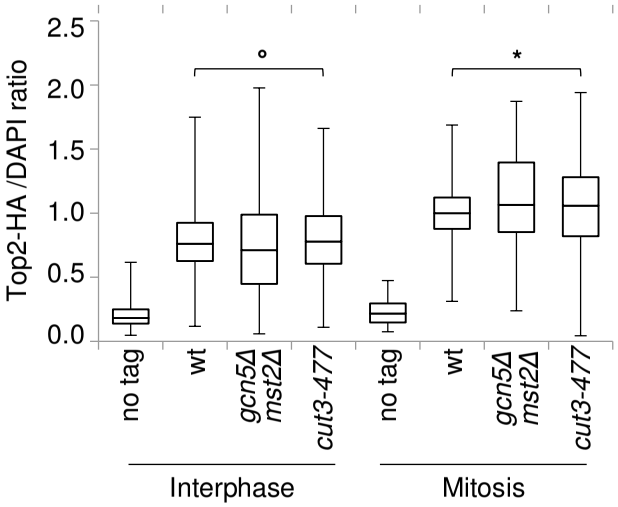

rDNA stretched
rDNA separated
% of anaphases
100
80
60
40
20
0
wt
gcn5-47
cut3-477
wt
gcn5-47
cut3-477
Normal
Defective
Chromosome segregation
(DAPI)


A
top2
cut3
cut14
cnd1
cnd2
cnd3
wt
gcn5-47
gcn5Δ
mRNA levels (ng)
25
20
15
10
5
0
RT + - + - + - + - + - + -
B
no tag
cut3-gfp
cut3-gfp gcn5-47
cut3-gfp gcn5Δ
top2-HA
top2-HA gcn5-47
top2-HA gcn5Δ
top2-HA cut3-477
no tag
GFP
Tub
HA
Tub
C
IP α-GFP
gcn5Δ
wt
gcn5Δ
wt
Cut3-HA - + + + +
Cnd2-GFP - + + - -
Total
α-HA
α-GFP
IP
α-HA
α-GFP
D
cnp3
mRNA level relative to wt
1.2
1.0
0.8
0.6
0.4
0.2
0.0
RT + - + - + -
wt
gcn5-47
gcn5D
E
cnp3
Ratios to act1 normalized to wt
1.2
0.8
0.4
0.0
RT + - + -
wt
gcn5-47 pCNP3
F
25°C
32°C
36°C
gcn5-47
cut3-477
gcn5-47 cut3-477
pCNP3
gcn5-47
cut3-477
gcn5-47 cut3-477

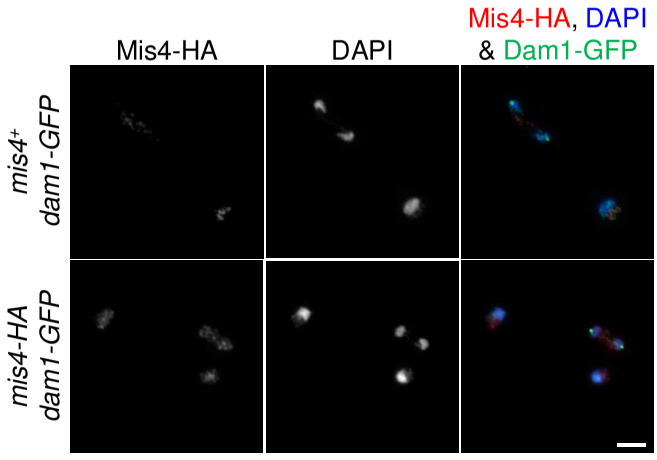

A
ChIP Cnd2-GFP
no tag
cnd2-GFP
cut3-477 cnd2-GFP
% IP
cnt1
dh1
rds1
prl53
exg1
cdc22
snoU14
5S
gly05
B
Mitotic indexes
% of cells
cnd2-GFP
cut3-477 cnd2-GFP
C
no tag (cnd2+)
cnd2-gfp
cnd2-gfp cut3-477
α-GFP
α-Tub
D
ChIP RNA Pol II (Ser2P)
wt
gcn5Δ
gcn5Δ mst2Δ
% Ip (IP/ Total)
CDS
5'IGR
18S
rds1
gas1
prl53
ecm33
B1E7
snoU14
exg1
slp1
cdc22
E
Mitotic Indexes
% of cells
wt
gcn5Δ
gcn5Δ mst2Δ

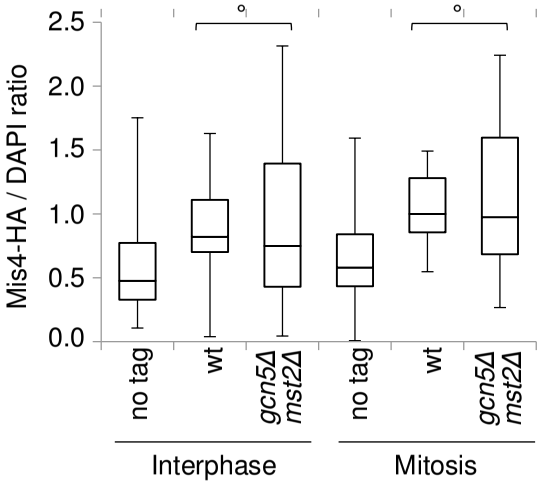
A


DAPI
Dam1-GFP
Top2-HA
Merge
dam1-gfp top2+ (no tag)
dam1-gfp top2-HA

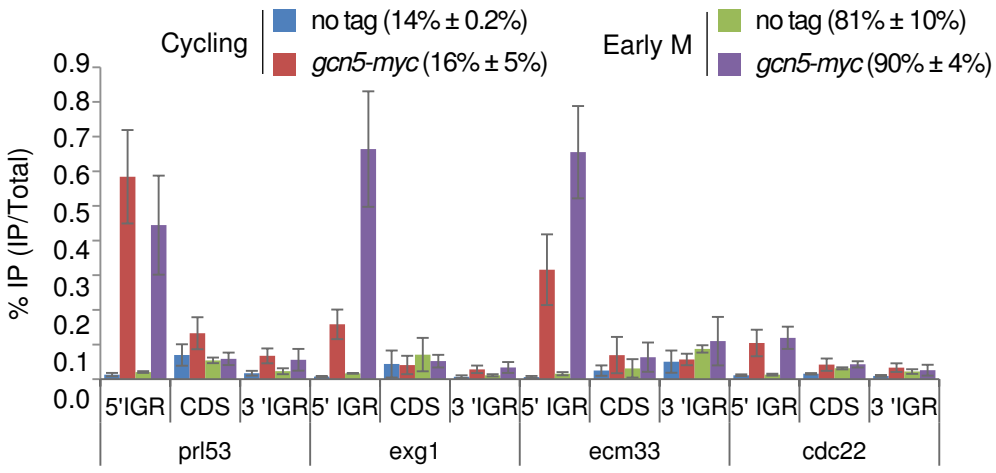

Top2-HA /DAPI ratio
2.5
2.0
1.5
1.0
0.5
0.0
°
*
no tag
wt
gcn5Δ mst2Δ
cut3-477
no tag
wt
gcn5Δ mst2Δ
cut3-477
Interphase
Mitosis

B

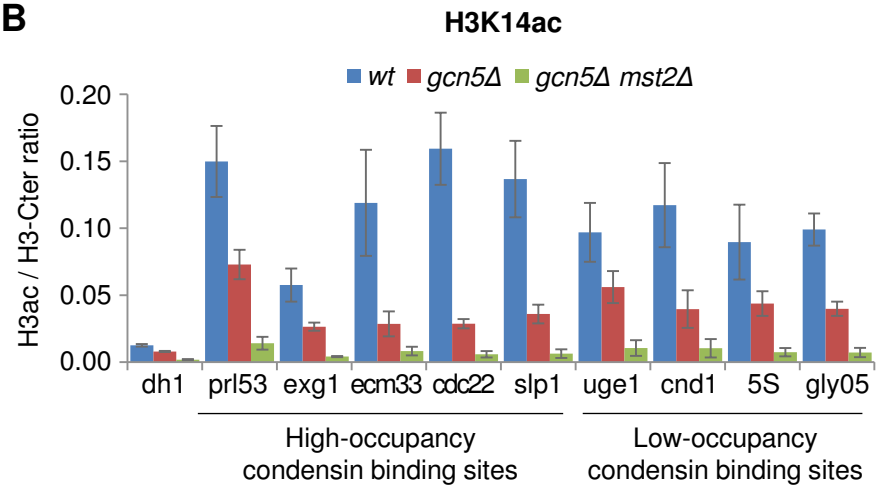

Mis4-HA
DAPI
Mis4-HA, DAPI & Dam1-GFP
mis4+ dam1-GFP
mis4-HA dam1-GFP

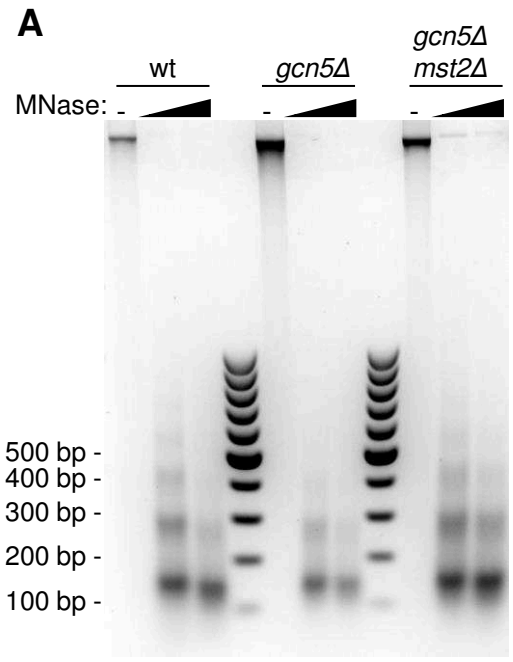

Mis4-HA / DAPI ratio
2.5
2.0
1.5
1.0
0.5
0.0
°
°
no tag
wt
gcn5Δ mst2Δ
no tag
wt
gcn5Δ mst2Δ
Interphase
Mitosis

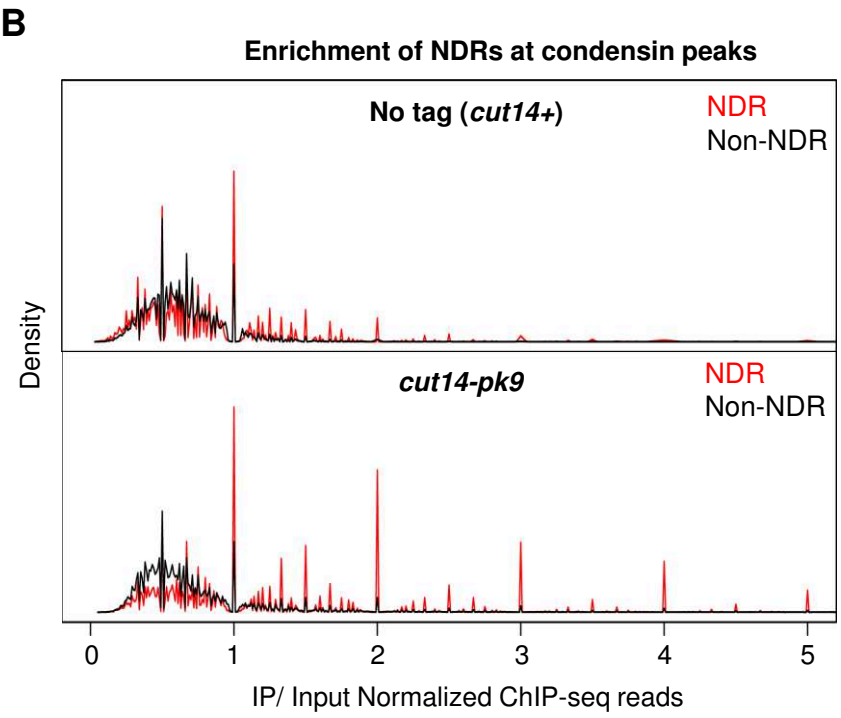
A

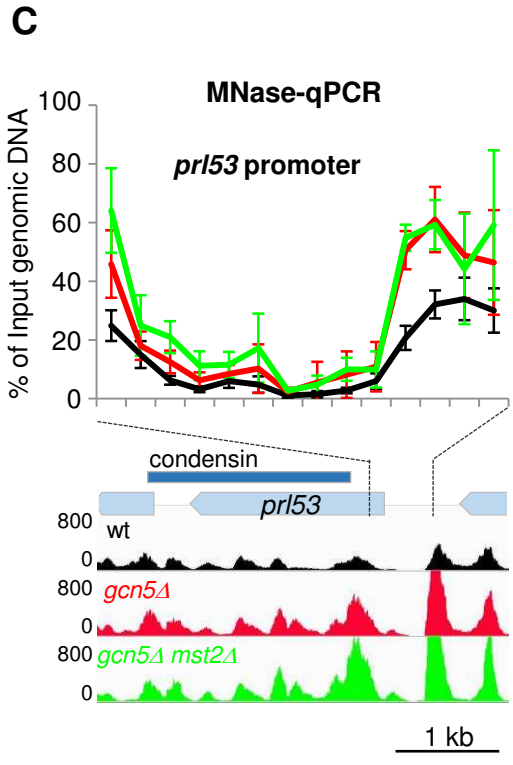

Cycling
no tag (14% ± 0.2%)
gcn5-myc (16% ± 5%)
Early M
no tag (81% ± 10%)
gcn5-myc (90% ± 4%)
% IP (IP/Total)
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
5'IGR
CDS
3 'IGR
5' IGR
CDS
3 'IGR
5' IGR
CDS
3 'IGR
5' IGR
CDS
3 'IGR
prl53
exg1
ecm33
cdc22

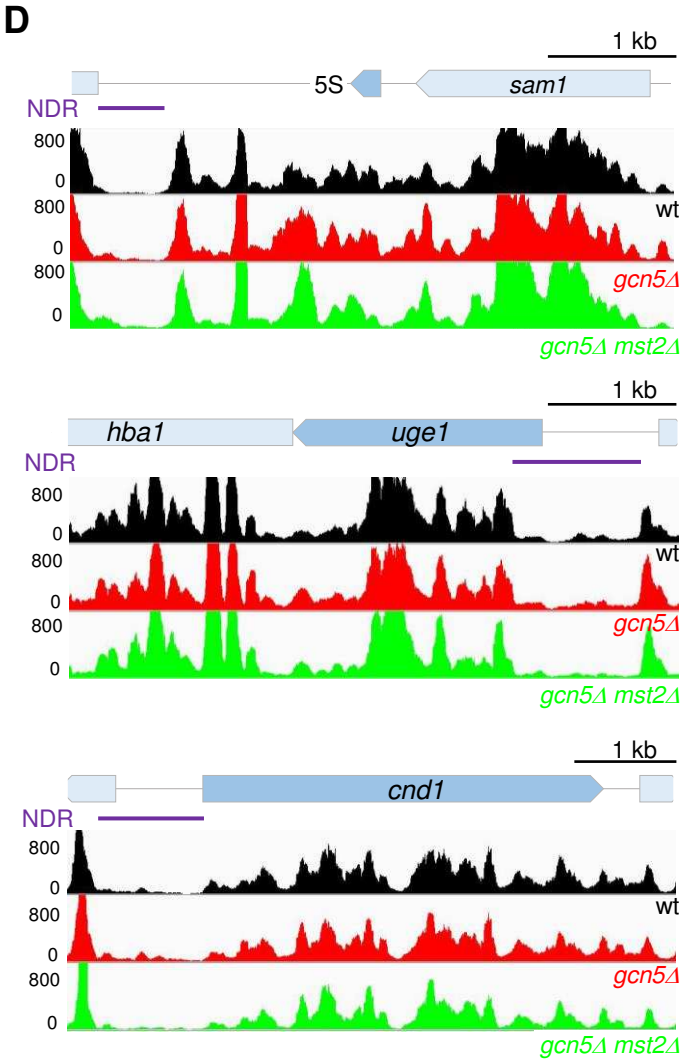
B

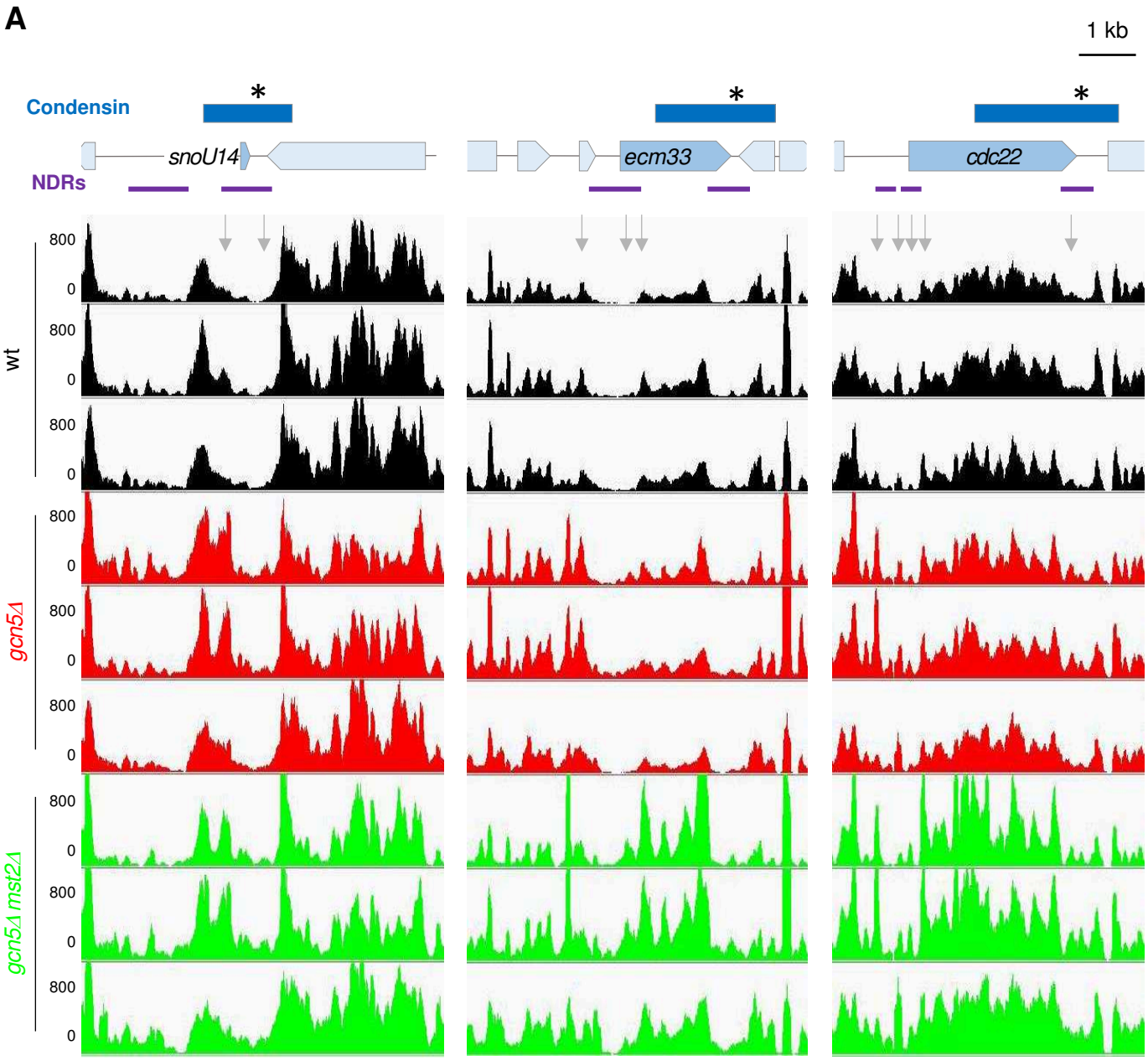

H3K14ac
wt
gcn5Δ
gcn5Δ mst2Δ
H3ac / H3-Cter ratio
0.20
0.15
0.10
0.05
0.00
dh1
prl53
exg1
ecm33
cdc22
slp1
uge1
cnd1
5S
gly05
High-occupancy condensin binding sites
Low-occupancy condensin binding sites

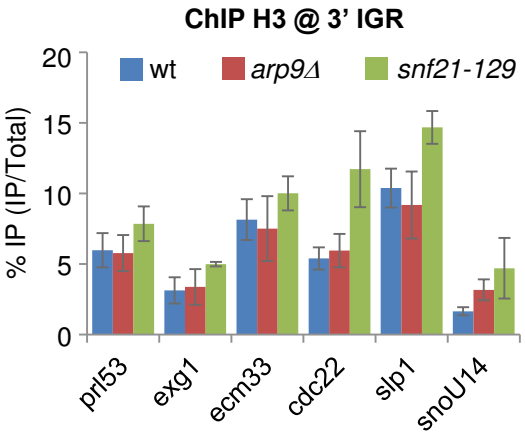

A
wt
gcn5Δ
gcn5Δ mst2Δ
MNase: - - -
500 bp -
400 bp -
300 bp -
200 bp -
100 bp -


B
Enrichment of NDRs at condensin peaks
No tag (cut14+)
NDR
Non-NDR
cut14-pk9
NDR
Non-NDR
Density
0
1
2
3
4
5
IP/ Input Normalized ChIP-seq reads


C
MNase-qPCR
prl53 promoter
% of Input genomic DNA
100
80
60
40
20
0
condensin
prl53
800
wt
0
800
gcn5Δ
0
800
gcn5Δ mst2Δ
0
1 kb


D
1 kb
5S
sam1
NDR
800
0
wt
gcn5Δ
gcn5Δ mst2Δ
hba1
uge1
cnd1


A
1 kb
Condensin
NDRs
snoU14
ecm33
cdc22
wt
gcn5Δ
gcn5Δ mst2Δ
800
0


ChIP H3 @ 3' IGR
wt
arp9Δ
snf21-129
% IP (IP/Total)
20
15
10
5
0
prl53
exg1
ecm33
cdc22
slp1
snoU14

DOI: 10.1016/j.biopen.2016.05.001

The proto-oncogenic protein TAL1 controls TGF-β1 signalling through interaction with SMAD3.

Jean-Michel Terme, Sébastien Lemaire, Didier Auboeuf, Vincent Mocquet and Pierre Jalinot#.

Univ Lyon, ENS de Lyon, Univ Claude Bernard, CNRS UMR 5239, INSERM U1210, Laboratory of Biology and Modelling of the Cell, 46 allée d'Italie Site Jacques Monod, F-69007, Lyon, France.

Short title : TAL1 interaction with SMAD3

Keywords : TAL1; SMAD3; TGF-β1; SMAD7

Corresponding author. Pierre Jalinot. Univ Lyon, ENS de Lyon, Univ Claude Bernard, CNRS UMR 5239, INSERM U1210, Laboratory of Biology and Modelling of the Cell, 46 allée d'Italie Site Jacques Monod, F-69007, Lyon, France.

Tel.: (33)472728563. Fax: (33)472728080. e. mail: pjalinot@ens-lyon.fr.

Abstract:

TGF-β1 is involved in many aspects of tissue development and homeostasis including haematopoiesis. The TAL1 transcription factor is also an important player of this latter process and is expressed very early in the myeloid and erythroid lineages. We previously established a link between TGF-β1 signalling and TAL1 by showing that the cytokine was able to induce its proteolytic degradation by the ubiquitin proteasome pathway. In this manuscript we show that TAL1 interacts with SMAD3 that acts in the pathway downstream of TGF-β1 association with its receptor. TAL1 expression strengthens the positive or negative effect of SMAD3 on various genes. Both transcription factors activate the inhibitory SMAD7 factor through the E box motif present in its transcriptional promoter. DNA precipitation assays showed that TAL1 present in Jurkat or K562 cells binds to this SMAD binding element in a SMAD3 dependent manner. SMAD3 and TAL1 also inhibit several genes including ID1, hTERT and TGF-β1 itself. In this latter case TAL1 and SMAD3 can impair the positive effect exerted by E47. Our results indicate that TAL1 expression can modulate TGF-β1 signalling by interacting with SMAD3 and by increasing its transcriptional properties. They also suggest the existence of a negative feedback loop between TAL1 expression and TGF-β1 signalling.

1. Introduction:

TAL1 (T-cell acute lymphoblastic leukaemia protein 1), also known as SCL (stem cell leukaemia) is a tissue-specific transcription factor belonging to the basic helix-loop-helix (bHLH) family which dimerizes with other bHLH E factors and binds E-box consensus motifs. By associating with DNA directly or indirectly through interactions with other transcription factors, TAL1 can act either positively or negatively on transcription of specific genes. This has been well-illustrated by systematic approaches characterizing the various TAL1 targets [1,2,3]. In the case of a positive action it has been shown that TAL1 can intervene in multiprotein complexes by associating with the LIM-only proteins LMO1 and LMO2, as well as the GATA factors [4,5,6,7,8]. TAL1 has also been shown to interact with the p300 coactivator through its basic domain [9]. These molecular interactions, along with others as with SP1 [10], can lead to transcriptional activation by TAL1. However TAL1 can also act negatively [1,2,3] and it has been reported that it binds to the mSin3A transcriptional repressor, also through its basic domain [11]. Experiments in mice have shown that TAL1 represses the E47/HEB activity in thymocytes by inducing recruitment of mSin3A [12]. The expression of TAL1 is restricted to specific cells and development stages [13]. In particular it is expressed in early haematopoiesis and plays an important role in the generation of the erythroid and myeloid lineages. In the T-cell lineage expression of TAL1 is normally lost early in the differentiation process, but its abnormal maintenance as a consequence of chromosomal rearrangements or epigenetic activation is commonly associated with T-cell acute lymphoblastic leukaemia (T-ALL) [14,15,16]. TAL1 expression in this context is likely to play an important role as its suppression by RNA interference in Jurkat cells has been shown to lead to proliferation stop [2]. We have previously shown that TAL1 can be regulated

by extracellular cytokines as TGF-β1 which induces its degradation through the ubiquitin proteasome pathway [17]. Association of TGF-β1 with its receptors leads to phosphorylation of the receptor-regulated SMADs SMAD2 and SMAD3 [18,19]. These factors can then associate with SMAD4 and enter the nucleus. These complexes by associating with specific sequence elements regulate either positively or negatively many genes [20,21]. TGF-β1 is important in many aspects of tissue homeostasis and in particular plays an important role in haematopoiesis as TAL1 does [22]. It has a regulatory role on haematopoietic stem cells (HSCs) and it is also important for differentiation of the myeloid and endothelial lineage. These effects involve SMAD3 as shown by gene disruption experiments which have established a role of this transcription factor in haematopoiesis [23].

In this manuscript we show that TAL1 can specifically bind SMAD3 and potentiate its positive or negative effect on specific genes. By activating negative regulator of the pathway as SMAD7 and inhibiting TGF-β1 expression itself, TAL1 and SMAD3 counteracts the TGF-β1 pathway functioning. These data together with previous observations are in favour of negative feedback loops between TGF-β1 and TAL1.

2. Materials and Methods:

2.1. Constructs:

Plasmids used in this study have been previously described: pSGF-TAL1 [24], pSG5-MYC-TAL1 [25] and pSG-HA-Ub [26]. pCMV-Flag-E47 was kindly provided by C. Gallego [27]. pCDNA3-FLAG-SMAD2, -SMAD3 and –SMAD4 were kindly provided by A. Favre-Bonvin.

2.2. Cell culture and transfection:

Cell culture and transfection were performed as previously described [24]. When indicated, cells were treated with human platelet TGF-β1 (100 pM). To achieve inhibition of the proteasome, MG132 (Sigma-Aldrich, StLouis, USA) was added at 10 μM for 4 h. Cellular extracts were normalized with respect to protein concentrations which were quantified with the DC protein assay kit (BioRad, Hercules, USA). HeLa cells were stably-transfected with the pCEP-FLAG-TAL or the control pCEP-FLAG-GFP constructs, as previously reported [24].

2.3. Immunoprecipitation and Immunoblot:

Transfected cells were lysed in Nonidet P-40-desoxycholate buffer (50 mM Tris pH 7.4, 150 mM NaCl, 1% Nonidet P-40, 0.5% Na desoxycholate, 0.5 mM Tris(2-carboxyethyl)phosphine, and 0.5 mM Pefabloc). Immunoprecipitation were carried out by addition of antibodies diluted 1:250 and incubation at 4°C for 1h30. Protein A sepharose beads were added and incubated with the mix for 1h. Beads were collected by centrifugation and washed three times with Nonidet P-40-desoxycholate buffer. For experience in supplementary figure 1 protein G magnetic beads were used. Proteins were eluted in 2x SDS sample buffer at 80°C for 10 min. After protein gel electrophoresis and transfer to PVDF membrane, detection was carried out by incubation with primary antibodies diluted 1:1,000 or as indicated by the manufacturer and revelation was performed by chemiluminescence using ECL, ECL+ or ECL Prime kits. Following antibodies were purchased: FLAG (clone M2, Sigma), MYC (clone 9E10, Sigma), HA (clone 12CA5, Roche), SMAD3 (NB600-1258), SMAD2/3 (566412, Calbiochem) and TAL1 (BTL73, Millipore).

2.4. Luciferase assays:

HeLa and 293T cells (1.4 10^5 cells) were transfected with 1.5µg of constructs including the firefly luciferase coding sequence under the control of various promoters: TGF-β1 [28], 3Tplux [29], CAGA boxes [30], hTERT [31], ID1 [32], SMAD7-WT and SMAD7-mE [33] together with 15 ng of thymidine kinase Renilla-luciferase (pRL-TK) by calcium phosphate precipitation. Reporter gene analysis was performed 48 h after transfection by using the Dual-Luciferase™ Reporter Assay System (Promega, Madison, USA). The luciferase activity associated with each construct was normalized on the basis of pRL-TK activity. The values are those obtained in triplicate, from one representative experiment (out of 3 experiments) and are represented with standard deviation. In figures one, two and three stars indicate Student's T test (two tail, unpaired) p-value less than 0.05, 0.01 and 0.001, respectively.

2.5. Real-time quantitative RT-PCR:

Total RNAs were extracted from frozen cells using the Rneasy mini-kit (Qiagen, Hilden, Germany). One step RT-PCR reactions were performed using the QuantiTect™ SYBR Green RT-PCR kit (Qiagen) and the LightCycler apparatus (Roche). Sequences of sense and antisense primers were described previously [24].

2.6. Nuclear extracts:

Jurkat and K562 cells were grown in RPMI-1640 with 5% foetal calf serum and treated with 781 pM TGF-β1 for 1h. Cells were collected by centrifugation, washed three times in PBS 1x and incubated for 30 min on ice in 3 pellet volumes of hypotonic buffer (10 mM Hepes pH 7.9, 10 mM KCl, 1.5 mM MgCl2, 0.1 mM EDTA pH 8.0, 0.1 mM EGTA, 1 mM DTT, 1x cOmplete™ mini EDTA-free protease inhibitor cocktail, 0.5 mM sodium vanadate).

Cells were lysed using Dounce B pestle and nuclei were collected by a 10 min centrifugation at 900 g. These nuclei were incubated for 15 min on ice in extraction buffer (20 mM Hepes pH 7.9, 360 mM NaCl, 1.5 mM MgCl2, 20% glycerol, 0.1 mM EDTA pH 8.0, 10 μMZnCl2, 1 mM DTT, 1x cOmplete™ mini EDTA-free protease inhibitor cocktail, 0.5 mM sodium vanadate). Nuclear extracts corresponded to the supernatants after a 15 min centrifugation at 12,000 g.

2.7. DNA precipitation assays:

The probe corresponding to the SMAD binding element (SBE) of the human SMAD7 gene was obtained by hybridizing a 5' biotinylated oligonucleotide (5'-AGCGACAGGGTGTCTAGACGGCCACGTGA -3') with another corresponding to the complementary strand. Probe (0.25 μM) was incubated in 160 μl with 100 μg of nuclear extracts, polydIdC and 15 mM Tris pH 7.9, 3 mM MgCl2 and 0.2 % Triton X-100 (final concentration) for 15 min on ice. Streptavidin magnetic beads were added and the mix was further incubated for 15 min on ice. The magnetic beads were washed three times in wash buffer (20 mM Tris pH 7.9, 50 mM KCl, 1.5 mM MgCl2, 10% glycerol, 0.1 mM EDTA pH 8.0, 10 μM ZnCl2, 1 mM DTT, 1x cOmplete™ mini EDTA-free protease inhibitor cocktail, 0.25 % Triton X-100). For competition experiments the mix was preincubated for 15 min on ice with double stranded oligonucleotides corresponding to the SMAD7 SBE wild type (5'-CAGGGTGTCTAGACGGCCAC - 3') or mutated (5'- CAGGGTCATAGCGTGGCCAC - 3'). Proteins associated with DNA were eluted by heating the beads for 10 min at 80°C in protein loading buffer and analysed by immunoblot after separation in 9% polyacrylamide SDS gel.

3. Results :

3.1. TAL1 interacts with SMAD3:

As we have previously shown that the intracellular amount of TAL1 was regulated by TGF-β1 [17], we investigated further whether this transcription factor was able to interfere with the TGF-β1 pathway through association with the SMAD proteins which are tightly regulated by association of the cytokine with its receptor. To test a possible direct binding of TAL1 to these factors, 293 T cells were transfected with vectors expressing SMAD2, SMAD3 and SMAD4 tagged with the FLAG epitope, together with a construct producing TAL1 fused to the MYC epitope. These cells, as HeLa cells do not express endogenous TAL1. These three SMADs were expressed at similar levels as evaluated by an immunoblot analysis of the cellular extracts (Figure 1A, bottom panel). Coexpression of the SMADs did not modify the level of MYC-tagged TAL1 in the extracts (Figure 1A, middle panel). Immunoprecipitation using the FLAG antibody showed a clear coprecipitation of TAL1 with SMAD3 (Figure 1A, top panel, lane 3), but no association with SMAD2 or SMAD4 (Figure 1A, top panel, lanes 2 and 4). The SMAD3 signal was absent when cells were transfected with the TAL1 or SMAD3 vector alone as controls. To verify this interaction, the experiment was also performed in the reverse way by precipitating an HA-tagged form of TAL1 and detecting SMAD3. Similarly a clear and specific binding of TAL1 to SMAD3 was observed (Supplementary Figure 1). These observations indicated that TAL1 can bind SMAD3.

As we have previously shown that TGF-β1 was able to induce polyubiquinylation and proteasome degradation of TAL1 [17] we tested if coexpression of SMAD2 or SMAD3 affects this process. By treating cells transfected with the TAL1 expression vector with TGF-β1, a decrease in the level of TAL1 protein was observed in agreement with our previous observations (Figure 1B, top panel, compare lanes 1 and 4). This effect was not modified by

coexpression of either SMAD2 or SMAD3 (Figure 1B, top panel, lanes 2, 3, 5 and 6). The SMAD3 protein is also modified post-translationnally by polyubiquitinylation [34]. This can be observed by transfecting cells with a SMAD3 expression vector and a construct producing ubiquitin tagged with the HA epitope. Immunoblot analysis using the antibody to HA of immunoprecipitated SMAD3 reveals a smear of polyubiquitinylated forms (Figure 1C, lane 6). This modification of SMAD3 is not affected by coexpression of TAL1 (Figure 1C, lanes 6 and 7).

These experiments show an interaction of TAL1 with SMAD3 and that this binding does not interfere with the modification by polyubiquitinylation of both proteins.

3.2. TAL1 and SMAD3 cooperate in activation of TGF-β1-induced genes:

To investigate the functional relevance of this interaction between TAL1 and SMAD3 we tested whether it affected the activity of these transcriptional regulators. To this end cells were transfected with constructs placing the luciferase reporter gene under the control of sequence motifs of the human plasminogen activator inhibitor-1 gene mediating transcriptional activation by TGF-β1. These motifs corresponded to a fragment of the PAI-1 gene (3TPLux) or to CAGA boxes in tandem. Overexpression of SMAD3 transactivated both constructs, as well as TAL1 but in this case to a much lesser extent (Figure 2A and B compare lanes 1, 2 and 3). Combination of SMAD3 and TAL1 led to a higher transactivation showing that both proteins were able to act synergistically (Figure 2A and B, lanes 4). To confirm that both proteins were able to cooperatively activate TGF-β1-responsive genes we tested a construct associating luciferase and the SMAD7 transcriptional promoter. SMAD7 expression is known to be induced by TGF-β1 and to downregulate signalling by this cytokine by causing degradation of TGF-β receptor type I (TGFBR1). Both TAL1 and SMAD3 transactivate the SMAD7 promoter and association of both proteins led to a further increase (Figure 3A compare lanes 1, 2, 3 and 4). Mutation of the E box of the SMAD7 promoter led to a

complete loss of the transactivation by the TAL1 SMAD3 combination (Figure 3A, lane 5). Considering the sensitivity of the effect to a mutation of the SMAD binding element in the SMAD7 transcriptional promoter we tested whether association of both proteins on this DNA sequence could be observed using endogenous proteins. This was done by performing DNA precipitation assays using a biotinylated probe corresponding to the SMAD7 SBE. Nuclear extracts were prepared from TGF-β1 treated Jurkat cells which express endogenous TAL1. In these extracts expression of both SMAD3 and TAL1 as a 34 kD protein can be detected by immunoblot analysis (Figure 3B, lane 1). Both proteins were also detected in the proteins eluted from the DNA probe (Figure 3B, lane 4), whereas no signal was obtained when this probe was omitted in the incubation, indicating that these two proteins did not bind non specifically to the magnetic beads under these conditions (Figure 3B, lane 3). These experiments were also performed with an antibody recognizing both SMAD3 and SMAD2. This was done with nuclear extracts of Jurkat and K562 erythroid cells, these latter also expressing endogenous TAL1. In these extracts the signals corresponding to SMAD2 and SMAD3 were clearly detected, the relative abundance of SMAD2 being higher in Jurkat (Figure 3C, lanes 1 and 5). Interestingly after DNA precipitation only the SMAD3 signal was detected with extracts of both cell types (Figure 3C, lanes 4 and 6, upper panels). As with Jurkat, the TAL1 protein of K562 cells bound to the SBE probe together with SMAD3 (Figure 3C, lanes 4 and 6, lower panels). To check that the TAL1 binding was specific of SMAD3 we performed competition experiment with a 4x and 20x excess of double stranded oligonucleotide corresponding to the SBE. This was done with nuclear extracts of K562 cells and sequences corresponding to wild type or mutated SBE. At 20x excess a clear reduction of the SMAD3 signal was seen with the wild type SBE oligonucleotide but not with the mutated one (Figure 3D, compare lanes 5 and 6, upper panel). As compared to the mutated one, the wild type SBE oligonucleotide also reduced the TAL1 signal (Figure 3D, lanes 5 and 6, lower

panel). This shows that impairment of SMAD3 binding also affects TAL1 detection indicating that a SMAD3-TAL1 complex binds and activates the SMAD7 SMAD responsive promoter.

3.3. TAL1 and SMAD3 can also cooperate in down-regulation of cellular genes:

TAL1 has been described as a transcriptional activator, but also as a transcriptional repressor. Such a negative effect was observed in particular by testing a limited series of genes in a cell line constitutively expressing TAL1 from an episomic EBV derived vector. Using this tool two cell lines were generated expressing either GFP, as a control, or TAL1 as previously reported [24]. In these constructs both proteins were tagged with the FLAG epitope. Quantitative real time PCR analysis of several genes expression level between these two cell lines showed that TAL1 was able to repress expression of IκBα and also of the ID1, ID2 and ID3 genes (Figure 4A) which are known to be R-SMAD sensitive. By contrast expression of the β-actin, cyclin B1, lamin, E47, ATM, ATF, p16, p21 and HPI-α genes which have not been described as responding to TGF-β was not affected by the presence of TAL1. Immunoblot analysis using the antibody to FLAG showed that both proteins were expressed at a similar level (Figure 4B, compare lanes 1 and 2). The TAL1 negative effect was also observed in transient expression studies using a construct including the luciferase reporter gene under the control of the ID1 transcriptional promoter (Figure 4C, compare lanes 1 and 2). Interestingly SMAD3 was also able to repress the ID1 promoter and coexpression of both TAL1 and SMAD3 led to a stronger effect (Figure 4C, lanes 3 and 4). We previously reported that TAL1 was able to negatively regulate the promoter of the protein subunit of the telomerase enzyme [25]. By testing the effect of SMAD3 on the hTERT promoter we also observed a dose dependent negative effect (Figure 5, compare lanes 1, 2 and 3). Coexpression of TAL1 also reinforced this negative effect of SMAD3 (Figure 5, lanes 5, 6).

Finally as SMAD3 is known to be a negative regulator of the TGF-β1 promoter we also investigated the TAL1 effect on this transcriptional regulatory element. TAL1 was observed to down regulate the TGF-β1 promoter (Figure 6A, compare lanes 1, 2 and 3) and also to counteract its activation by the type I bHLH protein E47 (Figure 6B). Indeed overexpression of E47 markedly increased the activity of this promoter, but coexpression of TAL1 impeded this effect in a dose dependent manner (Figure 6B, lanes 4, 5 and 6). By expressing a low amount of TAL1 this negative effect on E47 transactivation was not observed (Figure 6C, lanes 2 and 3) but coexpression of SMAD3 led to a clear impairment of the E47 effect (Figure 6C, lane 5). This clearly showed that the TAL1 SMAD3 complex can negatively downregulate several genes, including TGF-β1 itself.

By combining these different observations it appears that TAL1 and SMAD3 might play a clear role in a negative feedback loop downregulating TGF-β1 signalling. Indeed by activating SMAD7 and inhibiting TGF-β1 they are likely to act negatively on this pathway. In addition we have previously shown that TGF-β1 was able to induce TAL1 degradation through AKT activation. This shows that TAL1 in combination with SMAD3 could act negatively on key factors of the TGF-β1 pathway (Figure 7).

4. Discussion:

Both TAL1 and TGF-β1 have been established as important regulators of embryonic haematopoiesis. This is in particular illustrated by knock-out experiments in mice. TAL1-deficient mice indeed die around E9.5, with defects of haematopoietic and endothelial cell lineages [35,36]. Targeted disruption of the TGF-β1 gene also results in lethality at E10.5 in about 50% of mice with defective haematopoiesis and yolk sac vasculogenesis, transient survival of other mice being interpreted as a consequence of maternal transfer of the cytokine

[22,37]. Unexpectedly our results indicate a link between the events downstream of TGF-β1 binding to its receptor and TAL1.

Indeed analysis of the interaction of TAL1 with the SMADs mediating the TGF-β1 effect showed a specific binding of TAL1 to SMAD3. Both TAL1 and SMAD3 have been described to be regulated by polyubiquitinylation and proteasome degradation [17,34]. From our observations this post translational modification does not seem to be affected by interaction between both proteins. In particular the overexpression of SMAD3 does not modify the TGF-β1-induced degradation of TAL1 that we previously reported [17]. From a functional point of view, association between TAL1 and SMAD3 results in an increase of their effect on specific transcriptional targets, both in the case of a positive or a negative action. Analysis of the sequences mediating TGF-β1-activation of the human PAI-1 gene has shown the importance of three CAGA boxes present in the promoter [38]. SMAD3 overexpression leads to activation of reporter constructs bearing a PAI-1 TGF-β1 responsive element including this motif [29] or multiple repeats of these CAGAs boxes [30] and coexpression of TAL1 clearly potentiates this effect. Conversely SMAD3 can downregulate several genes, as hTERT or ID1, and TAL1 reinforces this effect. Such an inhibitory effect of SMAD3 on ID1 has already been reported [39], but appears to be cell-type dependent [40]. From the limited number of genes tested in this study TAL1 and SMAD3 seem to act in the same way and to potentiate their mutual action on specific targets. Future systematic studies performed in specific cell types should help to test more extensively this overlap between TAL1 and SMAD3 targets. As a first clue to this question we looked at the available chromatin immunoprecipitation sequencing (ChIP-seq) data. If many experiments have been performed with TAL1 and SMAD3 we did not find ChIP-seq results performed with both TAL1 and SMAD3 in the same cells in the presence of TGF-β1. However Micrococcal Nuclease (MNase) ChIP-seq data with both TAL1 and SMAD2/3 are reported in human embryonic stem cells [41]. Hence

we analysed these data and observed that several TAL1 and SMAD2/3 peaks overlap in a number of genes (Supplementary Figure 2). This can be observed in the promoter region and also in the transcribed sequence. These sequences represent potential regulatory elements involving both TAL1 and SMAD3. The list of the genes showing combined TAL1-SMAD2/3 peaks is given in supplementary Table 1. Interestingly one the gene exhibiting a combined TAL1/SMAD3 peak in its promoter, SKIL, codes for a component of the SMAD pathway which binds to promoters of TGF-β responsive genes and recruits a repressor complex [42]. Two other genes belonging to this category code for proteins also related to the SMAD pathway, ZCCHC14 is indeed reported to interact with SMAD3 and ZZEF1 with SKIL [43, 44]. These data support the notion that TAL1 and SMAD3 can co-regulate various genes. However this ChIP-seq approach will have to be carried out in various conditions of TGF-β1 treatment to increase the number of SMAD3 binding peaks to promoters which is relatively low in these HUES64 cells. Depending on the cellular context as well as on the presence of other partner factors in relation with the effect of TGF-β1, the regulatory effects reported in this manuscript might vary. However, as also supported by the MNase ChIP-seq data mentioned, our experiments show that co-binding of both TAL1 and SMAD3 to promoter sequences can be clearly observed with the endogenous proteins of both erythroid and leukemic lymphoid cells. Future detailed analyses will help to determine the exact contribution of the SMAD3-TAL1 protein-protein interaction and of DNA binding of both transcription factors to promoter sequences. Interestingly Dogan et al. have recently reported that binding of both TAL1 and SMAD1 faithfully predicts the existence of an enhancer element, in a better way that the presence of specific epigenetic marks. Hence co-binding of both TAL1 and SMAD3 is likely to be a good predictor of an active regulatory element [45].

Our analysis of the transcriptional targets of SMAD3 and TAL1 interestingly showed that both factors are likely to retroact negatively on TGF-β1 signalling. SMAD7 is indeed well

established as a negative regulator of the TGF-β1 signalling by impairing proper activation of the receptor-regulated SMAD2 and SMAD3 [46] or by triggering degradation of activated TGFBR-I [47]. TAL1 and SMAD3 synergistically activate expression of this inhibitory SMAD. Also analysis of the TGF-β1 promoter itself indicates a negative effect of both factors. In agreement with previous description of the negative effect of TAL1 on E47 [12], expression of the former was observed to impede the strong activation exerted by the latter on this promoter and SMAD3 strengthened this effect. These observations indicate that both factors are likely to facilitate termination of TGF-β1 signalling through these specific actions. As we previously showed that TGF-β1 causes degradation of TAL1 by the ubiquitin-proteasome pathway through AKT phosphorylation it appears that a negative feedback loop is likely to exist between TGF-β1 and TAL1.

Hence in TAL1 expressing cells, in particular the precursors of the haematopoietic and endothelial lineages, the action of TGF-β1 is probably restricted by the presence of this tissue-specific bHLH factor. In this line we have observed the SMAD3-TAL1 interaction on the SMAD7 SBE was observed with K562 cells extracts. Accordingly the well-established negative effect of TGF-β1 on proliferation of HSC [22] might be related to a downregulation of TAL1. It is interesting to mention that SMAD3 gene targeted disruption mimics at least to some extent the TGF-β1 effects on haematopoiesis [23], showing a specific effect of this particular SMAD in these cells. TGF-β1 restricts HSC proliferation, but is also important for differentiation of these cells [22]. Our observations indicate that this process might involve a fine tuning of the TGF-β1 signalling by the level of TAL1 expression. It will be also interesting to study this connection between TGF-β1 signalling and TAL1 in the case of T-ALL as we have also observed its binding to SMAD3 in such cells. TAL1 is an important factor in many pediatric and adult T-ALL and mutations in the TGFBR1 receptor have been

reported in some cases of T-ALL [48]. Hence it will be interesting to decipher how both TAL1 and TGF-β1 intervene in various types of T-ALL cells.

In conclusion our observations establish a new and unanticipated link between TAL1 and TGF-β1 through SMAD3. These data along with the current knowledge on these proteins incite to develop future studies to assess the importance of this connection in haematopoiesis, as well as in T-cell leukaemia onset.

5. Acknowledgements:

We are very grateful to J. Campisi, Z. Chang, X. Hua, S.J. Kim and A. Seth for generously providing us with expression vectors. We also wish to thank Armelle Roisin for help with cell culture. This work was supported by "Association pour la Recherche sur le Cancer" (grant and VM fellowship) and by the "Fondation pour la Recherche Médicale" (J-M T fellowship).

6. References:

1. Kassouf MT, Hughes JR, Taylor S, McGowan SJ, Soneji S, et al. (2010) Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. Genome Res 20: 1064-1083.

2. Palomero T, Odom DT, O'Neil J, Ferrando AA, Margolin A, et al. (2006) Transcriptional regulatory networks downstream of TAL1/SCL in T-cell acute lymphoblastic leukemia. Blood 108: 986-992.

3. Wilson NK, Miranda-Saavedra D, Kinston S, Bonadies N, Foster SD, et al. (2009) The transcriptional program controlled by the stem cell leukemia gene Scl/Tal1 during early embryonic hematopoietic development. Blood 113: 5456-5465.

4. Deleuze V, Chalhoub E, El-Hajj R, Dohet C, Le Clech M, et al. (2007) TAL-1/SCL and its partners E47 and LMO2 up-regulate VE-cadherin expression in endothelial cells. Mol Cell Biol 27: 2687-2697.

5. Lahlil R, Lecuyer E, Herblot S, Hoang T (2004) SCL assembles a multifactorial complex that determines glycophorin A expression. Mol Cell Biol 24: 1439-1452.

6. Osada H, Grutz G, Axelson H, Forster A, Rabbitts TH (1995) Association of erythroid transcription factors: complexes involving the LIM protein RBTN2 and the zinc-finger protein GATA1. Proc Natl Acad Sci U S A 92: 9585-9589.

7. Valge-Archer VE, Osada H, Warren AJ, Forster A, Li J, et al. (1994) The LIM protein RBTN2 and the basic helix-loop-helix protein TAL1 are present in a complex in erythroid cells. Proc Natl Acad Sci U S A 91: 8617-8621.

8. Yu M, Riva L, Xie H, Schindler Y, Moran TB, et al. (2009) Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. Mol Cell 36: 682-695.

9. Huang S, Qiu Y, Stein RW, Brandt SJ (1999) p300 functions as a transcriptional coactivator for the TAL1/SCL oncoprotein. Oncogene 18: 4958-4967.

10. Lecuyer E, Herblot S, Saint-Denis M, Martin R, Begley CG, et al. (2002) The SCL complex regulates c-kit expression in hematopoietic cells through functional interaction with Sp1. Blood 100: 2430-2440.

11. Huang S, Brandt SJ (2000) mSin3A regulates murine erythroleukemia cell differentiation through association with the TAL1 (or SCL) transcription factor. Mol Cell Biol 20: 2248-2259.

12. O'Neil J, Shank J, Cusson N, Murre C, Kelliher M (2004) TAL1/SCL induces leukemia by inhibiting the transcriptional activity of E47/HEB. Cancer Cell 5: 587-596.

13. Pulford K, Lecointe N, Leroy-Viard K, Jones M, Mathieu-Mahul D, et al. (1995) Expression of TAL-1 proteins in human tissues. Blood 85: 675-684.

14. Aplan PD, Lombardi DP, Reaman GH, Sather HN, Hammond GD, et al. (1992) Involvement of the putative hematopoietic transcription factor SCL in T-cell acute lymphoblastic leukemia. Blood 79: 1327-1333.

15. Lecuyer E, Hoang T (2004) SCL: from the origin of hematopoiesis to stem cells and leukemia. Exp Hematol 32: 11-24.

16. Macintyre EA, Smit L, Ritz J, Kirsch IR, Strominger JL (1992) Disruption of the SCL locus in T-lymphoid malignancies correlates with commitment to the T-cell receptor alpha beta lineage. Blood 80: 1511-1520.

17. Terme JM, Lhermitte L, Asnafi V, Jalinot P (2009) TGF-beta induces degradation of TAL1/SCL by the ubiquitin-proteasome pathway through AKT-mediated phosphorylation. Blood 113: 6695-6698.

18. Massague J, Seoane J, Wotton D (2005) Smad transcription factors. Genes Dev 19: 2783-2810.

19. Shi Y, Massague J (2003) Mechanisms of TGF-beta signalling from cell membrane to the nucleus. Cell 113: 685-700.

20. Koinuma D, Tsutsumi S, Kamimura N, Imamura T, Aburatani H, et al. (2009) Promoter-wide analysis of Smad4 binding sites in human epithelial cells. Cancer Sci 100: 2133-2142.

21. Qin H, Chan MW, Liyanarachchi S, Balch C, Potter D, et al. (2009) An integrative ChIP-chip and gene expression profiling to model SMAD regulatory modules. BMC Syst Biol 3: 73.

22. Soderberg SS, Karlsson G, Karlsson S (2009) Complex and context dependent regulation of hematopoiesis by TGF-beta superfamily signalling. Ann N Y Acad Sci 1176: 55-69.

23. Epperly MW, Cao S, Goff J, Shields D, Zhou S, et al. (2005) Increased longevity of hematopoiesis in continuous bone marrow cultures and adipocytogenesis in marrow stromal cells derived from Smad3(-/-) mice. Exp Hematol 33: 353-362.

24. Terme JM, Wencker M, Favre-Bonvin A, Bex F, Gazzolo L, et al. (2008) Crosstalk between expression of the HTLV-1 Tax transactivator and the oncogenic bHLH transcription factor TAL1. J Virol 82: 7913-7922.

25. Terme JM, Mocquet V, Kuhlmann AS, Zane L, Mortreux F, et al. (2009) Inhibition of the hTERT promoter by the proto-oncogenic protein TAL1. Leukemia 23: 2081-2089.

26. Buchsbaum S, Morris C, Bochard V, Jalinot P (2007) Human INT6 interacts with MCM7 and regulates its stability during S phase of the cell cycle. Oncogene 26: 5132-5144.

27. Liu Y, Encinas M, Comella JX, Aldea M, Gallego C (2004) Basic helix-loop-helix proteins bind to TrkB and p21(Cip1) promoters linking differentiation and cell cycle arrest in neuroblastoma cells. Mol Cell Biol 24: 2662-2672.

28. Lee SJ, Yang EK, Kim SG (2006) Peroxisome proliferator-activated receptor-gamma and retinoic acid X receptor alpha represses the TGFbeta1 gene via PTEN-mediated p70 ribosomal S6 kinase-1 inhibition: role for Zf9 dephosphorylation. Mol Pharmacol 70: 415-425.

29. Li H, Seth A (2004) An RNF11: Smurf2 complex mediates ubiquitination of the AMSH protein. Oncogene 23: 1801-1808.

30. Xin H, Xu X, Li L, Ning H, Rong Y, et al. (2005) CHIP controls the sensitivity of transforming growth factor-beta signalling by modulating the basal level of Smad3 through ubiquitin-mediated degradation. J Biol Chem 280: 20842-20850.

31. Kyo S, Takakura M, Taira T, Kanaya T, Itoh H, et al. (2000) Sp1 cooperates with c-Myc to activate transcription of the human telomerase reverse transcriptase gene (hTERT). Nucleic Acids Res 28: 669-677.

32. Nehlin JO, Hara E, Kuo WL, Collins C, Campisi J (1997) Genomic organization, sequence, and chromosomal localization of the human helix-loop-helix Id1 gene. Biochem Biophys Res Commun 231: 628-634.

33. Hua X, Miller ZA, Benchabane H, Wrana JL, Lodish HF (2000) Synergism between transcription factors TFE3 and Smad3 in transforming growth factor-beta-induced transcription of the Smad7 gene. J Biol Chem 275: 33205-33208.

34. Gao S, Alarcon C, Sapkota G, Rahman S, Chen PY, et al. (2009) Ubiquitin ligase Nedd4L targets activated Smad2/3 to limit TGF-beta signalling. Mol Cell 36: 457-468.

35. Robb L, Lyons I, Li R, Hartley L, Kontgen F, et al. (1995) Absence of yolk sac hematopoiesis from mice with a targeted disruption of the scl gene. Proc Natl Acad Sci U S A 92: 7075-7079.

36. Shivdasani RA, Mayer EL, Orkin SH (1995) Absence of blood formation in mice lacking the T-cell leukaemia oncoprotein tal-1/SCL. Nature 373: 432-434.

37. Dickson MC, Martin JS, Cousins FM, Kulkarni AB, Karlsson S, et al. (1995) Defective haematopoiesis and vasculogenesis in transforming growth factor-beta 1 knock out mice. Development 121: 1845-1854.

38. Dennler S, Itoh S, Vivien D, ten Dijke P, Huet S, et al. (1998) Direct binding of Smad3 and Smad4 to critical TGF beta-inducible elements in the promoter of human plasminogen activator inhibitor-type 1 gene. Embo J 17: 3091-3100.

39. Song H, Guo B, Zhang J, Song C (2010) Transforming growth factor-beta suppressed Id-1 Expression in a smad3-dependent manner in LoVo cells. Anat Rec (Hoboken) 293: 42-47.

40. Liang YY, Brunicardi FC, Lin X (2009) Smad3 mediates immediate early induction of Id1 by TGF-beta. Cell Res 19: 140-148.

41. Tsankov AM, Gu H, Akopian V, Ziller MJ, Donaghey J, et al. (2015) Transcription factor binding dynamics during human ES cell differentiation. Nature 518:344-349.

42. Stroschein SL1, Wang W, Zhou S, Zhou Q, Luo K. (1999) Negative feedback regulation of TGF-beta signaling by the SnoN oncoprotein. Science 286:771-774.

43. Wang J, Huo K, Ma L, Tang L, Li D, et al. (2011) Toward an understanding of the protein interaction network of the human liver. Mol Syst Biol. 7:536.

44. Colland F, Jacq X, Trouplin V, Mougin C, Groizeleau C, et al. (2004) Functional proteomics mapping of a human signaling pathway. Genome Res. 14:1324-1332.

45. Dogan N, Wu W, Morrissey CS, Chen KB, Stonestrom A, et al. (2015) Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. Epigenetics Chromatin 8:16.

46. Hayashi H, Abdollah S, Qiu Y, Cai J, Xu YY, et al. (1997) The MAD-related protein Smad7 associates with the TGFbeta receptor and functions as an antagonist of TGFbeta signalling. Cell 89: 1165-1173.

47. Yan X, Liu Z, Chen Y (2009) Regulation of TGF-beta signalling by Smad7. Acta Biochim Biophys Sin (Shanghai) 41: 263-272.

48. Kim SJ, Letterio J (2003) Transforming growth factor-beta signalling in normal and malignant hematopoiesis. Leukemia 17: 1731-1737.

7. Figure legends:

7.1. Figure 1: Interaction between TAL1 and SMAD3: (A) 293T cells were transfected with 2µg of pSG5-MYC-TAL1 and either pCDNA3-FLAG-SMAD2, -SMAD3 or -SMAD4 expression vectors as indicated. To avoid possible TAL1 degradation cells were treated with

the MG132 proteasome inhibitor. Cell lysates were used for immunoprecipitation with an antibody to FLAG. Immunoprecipitated proteins were analyzed by immunoblot using antibodies to MYC (top panels) and to FLAG (bottom panel). (B) 293T cells were transfected with pSG5-MYC-TAL1 either alone or in combination with pCDNA3-FLAG-SMAD2 or –SMAD3 and treated with 100 pM of TGF-β1 during 9 h. Cellular extracts were normalized with respect to protein concentration and analyzed by immunoblot using antibodies to MYC (top panel) or to β-actin as control (bottom panel). (C) 293T cells were transfected with pSG5-MYC-TAL1, pCDNA3-FLAG-SMAD3 and pSG-HA-Ub as indicated. Cell lysates were used for immunoprecipitation with an antibody to FLAG. Immunoblot analysis was done using the antibody to HA (top panel) or to MYC (bottom panel).

7.2. Figure 2: TAL1 and SMAD3 transactivate TGF-β responsive promoter sequence elements: HeLa cells were transfected with 300 ng of pSGF or pSGF-TAL1, and 100 ng of pCDNA3 or pCDNA3-SMAD3, together with 1.5 µg of 3TPLux (A) and CAGA (B). Cell extracts were prepared 48 h after transfection and analysed for luciferase activity. The graph represents the relative luciferase activity (mean of three points) and error bar corresponds to standard deviation. P-values of a Student's T test are represented as described in materials and methods.

7.3. Figure 3: TAL1 cooperates with SMAD3 to induce the SMAD7 promoter: (A) HeLa cells were transfected with 300 ng of pSGF or pSGF-TAL1, and 100 ng of pCDNA3 or pCDNA3-SMAD3, together with 1.5 µg of pSMAD7-WT (WT) or pSMAD7-mE which includes a mutated E-box motif. Analysis of luciferase activity was performed as described in legend to Figure 2. (B). DNA precipitation assay was carried out using the SMAD7 SBE probe and a nuclear extract of Jurkat cells treated for 1 h with TGF-β1. Lane 1 corresponds to the direct loading of 10% of the extract and lane 2 to loading of a protein molecular weight marker (MWM). Proteins recovered after DNA precipitation were loaded in lanes 3 and 4 but

in lane 3 the DNA probe was omitted as a control of non-specific binding to the beads. The upper and lower panels correspond to immunoblot analyses with the antibodies directed against SMAD3 and TAL1, respectively. The star on the left indicates a non-specific signal. The depicted images correspond to the superposition of colorimetric and chemiluminescence acquisitions with a ChemiDoc Touch Imaging System apparatus (BioRad, Hercules, USA). The molecular weights of the marker bands are indicated on the left. (C). The DNA precipitation experiment was carried out as in B with nuclear extracts of K562 (lanes 1 to 4) or Jurkat cells (lanes 5 to 6) and SMAD revelation was done using an antibody recognizing both SMAD2 (upper band) and SMAD3 (lower band) as indicated. The immunoblots were also revealed with the monoclonal antibody to TAL1 (lower panels). Lanes 1 and 5 correspond to direct loading of the extract. In lane 3 the DNA probe was omitted as control. (D) The DNA precipitation assay was performed with a nuclear extract of TGF-β1-treated K562 cells and the results are shown as in B. Lane 1 corresponds to 10% of this extract. In lanes 3 to 6 the nuclear extract was preincubated with a 4x or 20x excess of double stranded oligonucleotides corresponding to the SBE wild type (SBE) or mutated (SBEm) as indicated. The molecular weight marker was loaded in lane 7 and the molecular weights of the bands are indicated on the right.

7.4. Figure 4: ID1 promoter repression by TAL1 and SMAD3: (A) HeLa cells were stably-transfected with the pCEP-FLAG-TAL1 or the control pCEP-FLAG-GFP constructs. After selection of stably-transfected cells, total RNAs were prepared and analyzed by real-time quantitative RT-PCR for expression of various genes as indicated. The amount of mRNAs are expressed as a ratio FLAG-TAL1 cells versus FLAG-GFP cells. (B) Immunoblot analyses of cells transfected with pCEP-FLAG-GFP (lane 1) or pCEP-FLAG-TAL1 (lane 2) using antibodies to FLAG. (C) HeLa cells were transfected with 150 ng of pSGF (lanes 1 and 3) or pSGF-TAL1 (lanes 2 and 4), and 100 ng of pCDNA3 (lanes 1 and 2) or pCDNA3-

SMAD3 (lanes 3 and 4), together with 1.5 µg of pID1-Luc. Analysis of luciferase activity was performed as described in legend to Figure 2.

7.5. Figure 5: Repression of the hTERT promoter by TAL1 and SMAD3: HeLa cells were transfected with 300 ng of pSGF (lanes 1 to 3) or pSGF-TAL1 (lanes 4 to 6), 100 and 50 ng of pCDNA3 (lanes 1, 4 and 2, 5) and 50 or 100 ng of pCDNA3-SMAD3 (lanes 2, 5 and 3, 6), together with 1.5 µg of pGL3-hTERT-3300. Analysis of luciferase activity was performed as described in legend to Figure 2.

7.6. Figure 6: TAL1 represses TGF-β1 promoter: (A) HeLa cells were transfected with pSGF (lanes 1) or 300 ng and 600 ng of pSGF-TAL1 (lanes 2 and 3), together with 1.5 µg of pTGFβ1-1132-luc (B) HeLa cells were transfected with pSGF (lanes 1 and 4) or 300 ng and 600 ng of pSGF-TAL1 (lanes 2 and 3, 5 and 6), together with 150 ng of pCMV (lanes 1 to 3) or pCMV-E47 (lanes 4 to 6) and 1.5 µg of pTGFβ1-1132-luc. (C) HeLa cells were transfected with 50 ng of pSGF or pSGF-TAL1, 50 ng of pCMV or pCMV-E47, 75 ng of pCDNA3 or pCDNA3-SMAD3 as indicated, together with 1.5 µg of pTGFβ1-1132-luc. For all three panels analyses of luciferase activity were performed as described in legend to Figure 2.

7.7. Figure 7: Crosstalk between TGF-β signalling pathways and TAL1: This scheme recapitulates how TAL1 and SMAD3 act positively or negatively on TGF-β1 target genes, this leading to a downregulation of TGF-β1 signalling through SMAD7 and reduction of the cytokine expression. Conversely TGF-β1 can lead to TAL1 degradation through AKT activation.

Figure 1

Click here to download high resolution image

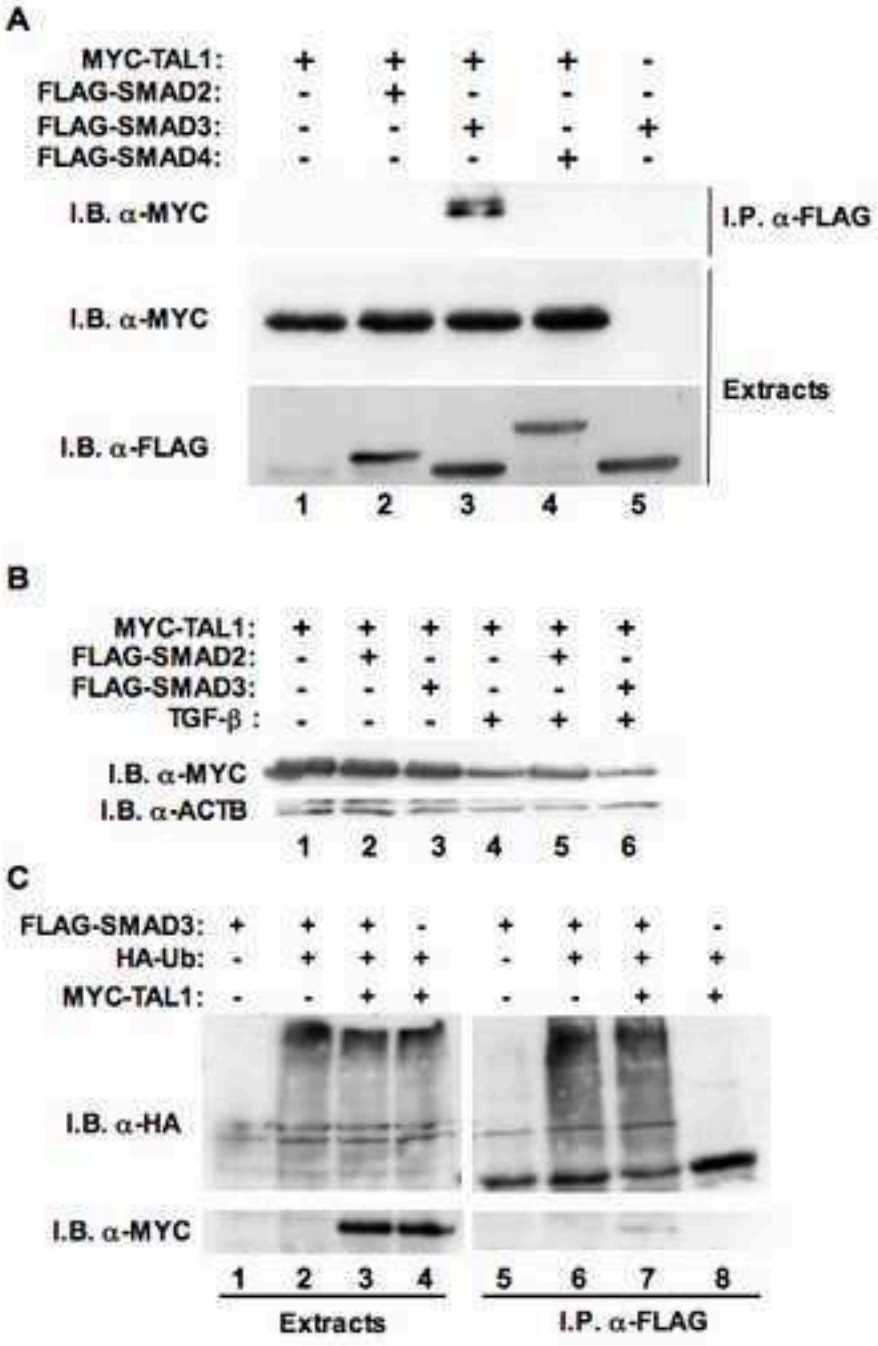


Figure 2

Click here to download high resolution image

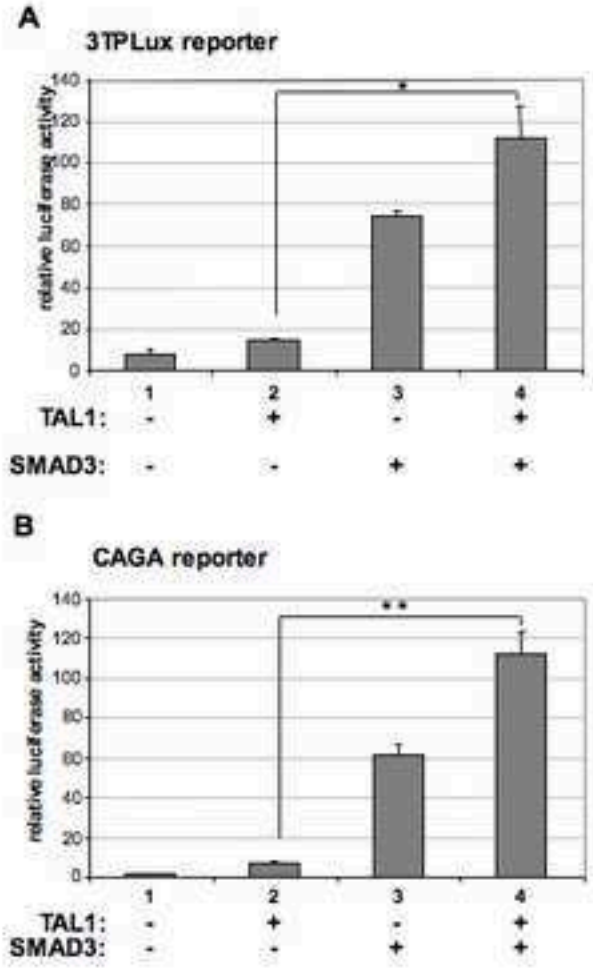


Figure 3

Click here to download high resolution image

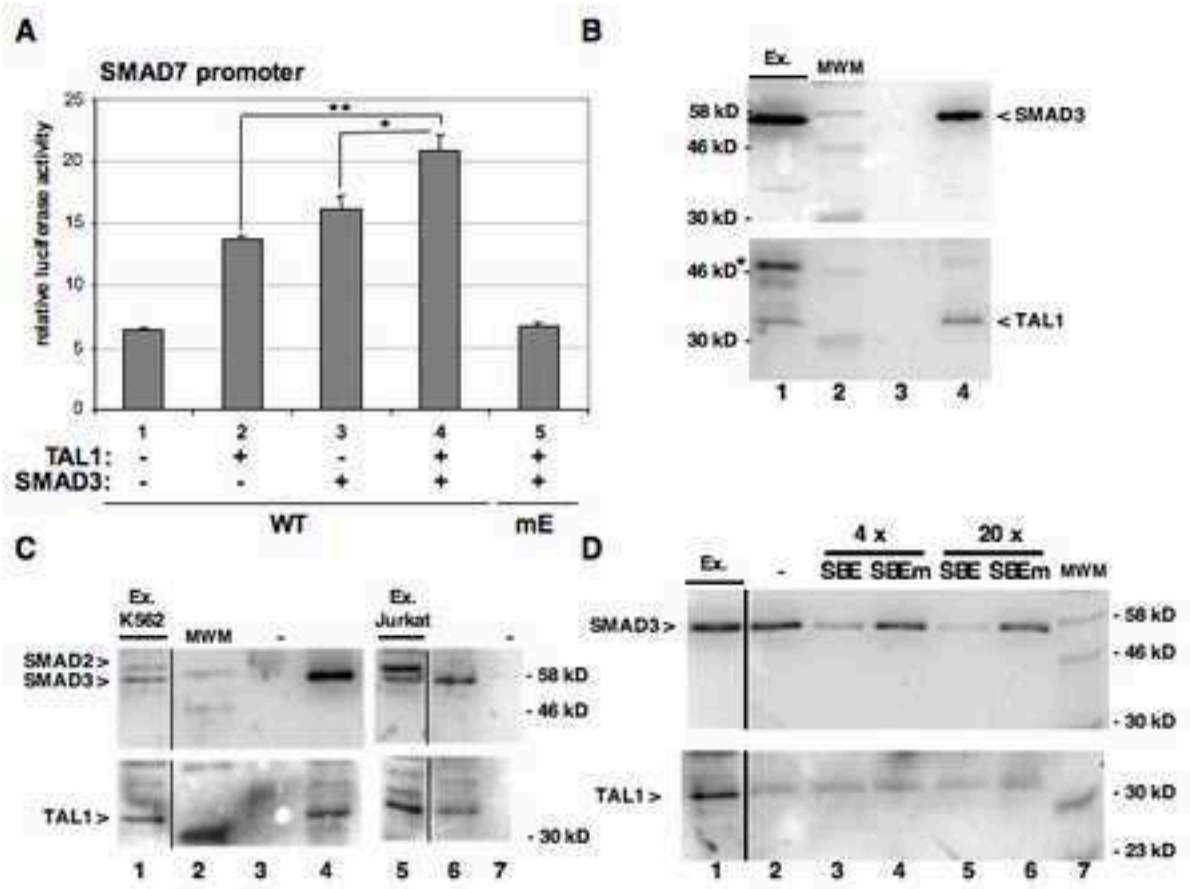


Figure 4

Click here to download high resolution image

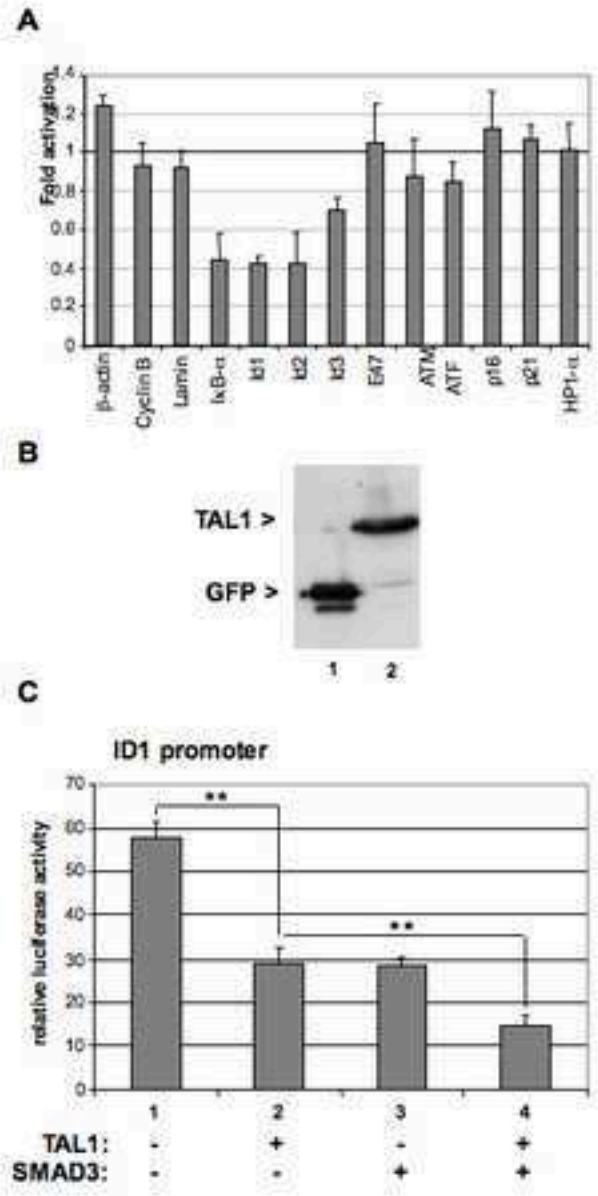


Figure 5

Click here to download high resolution image

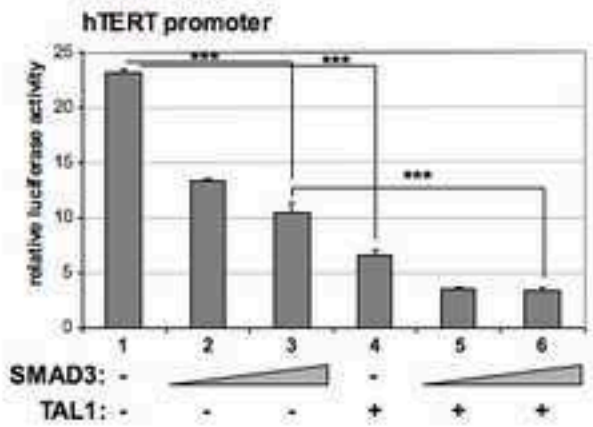


Figure 6

Click here to download high resolution image

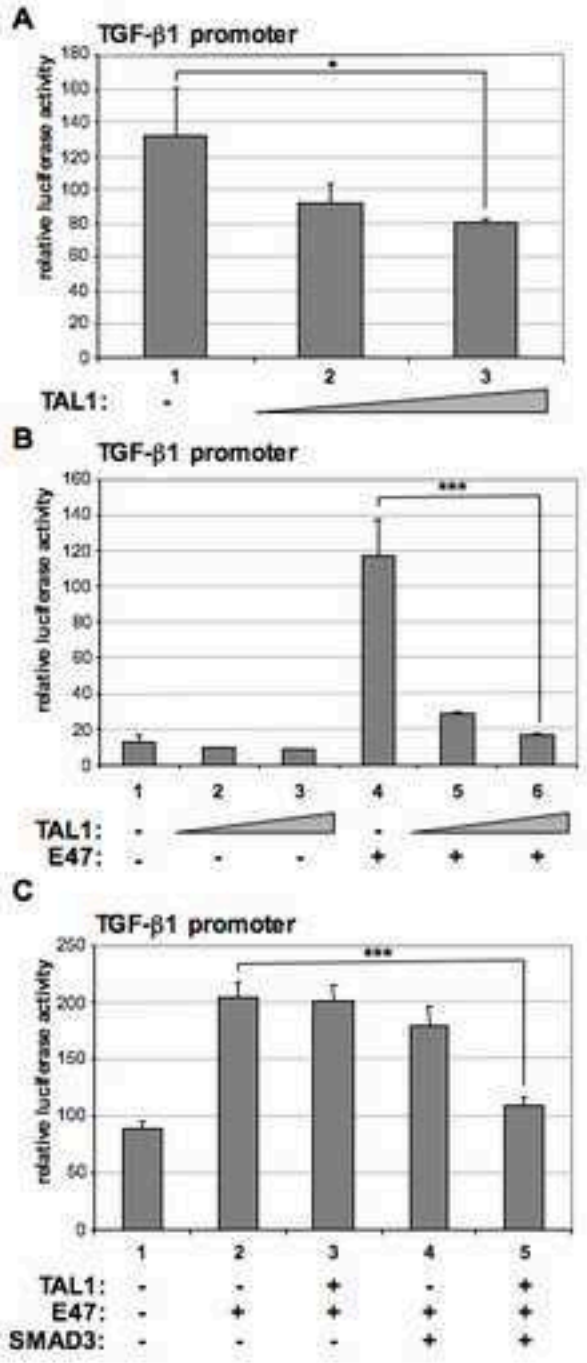
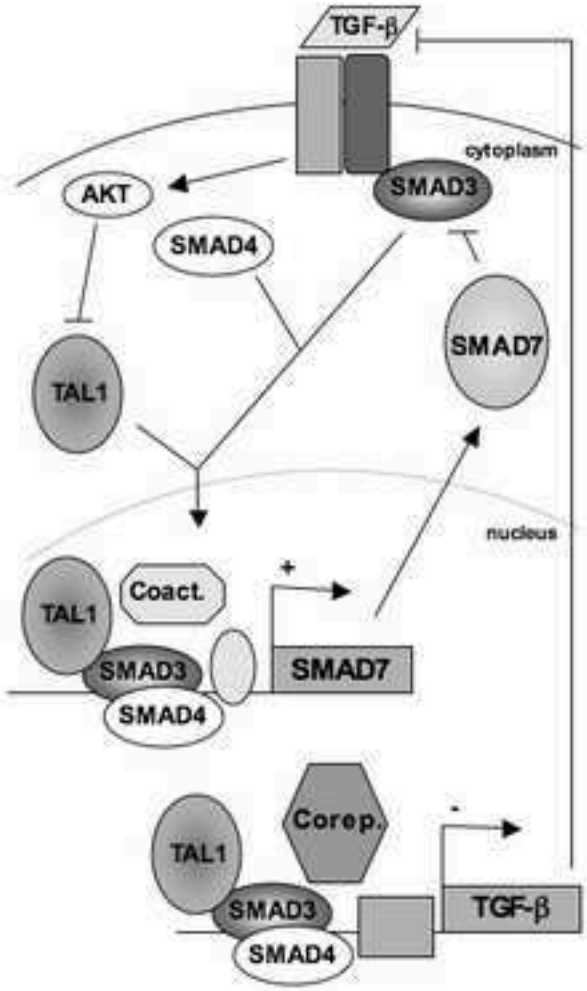


Figure 7

Click here to download high resolution image

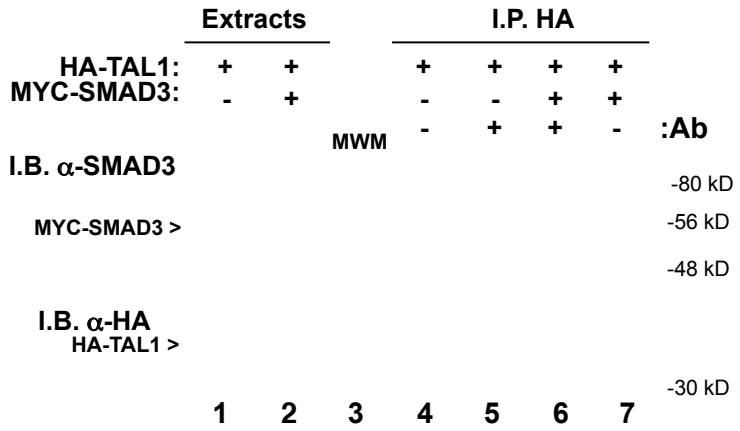


***Referee suggestions**

Please see original submission.

Supplementary Figure 1

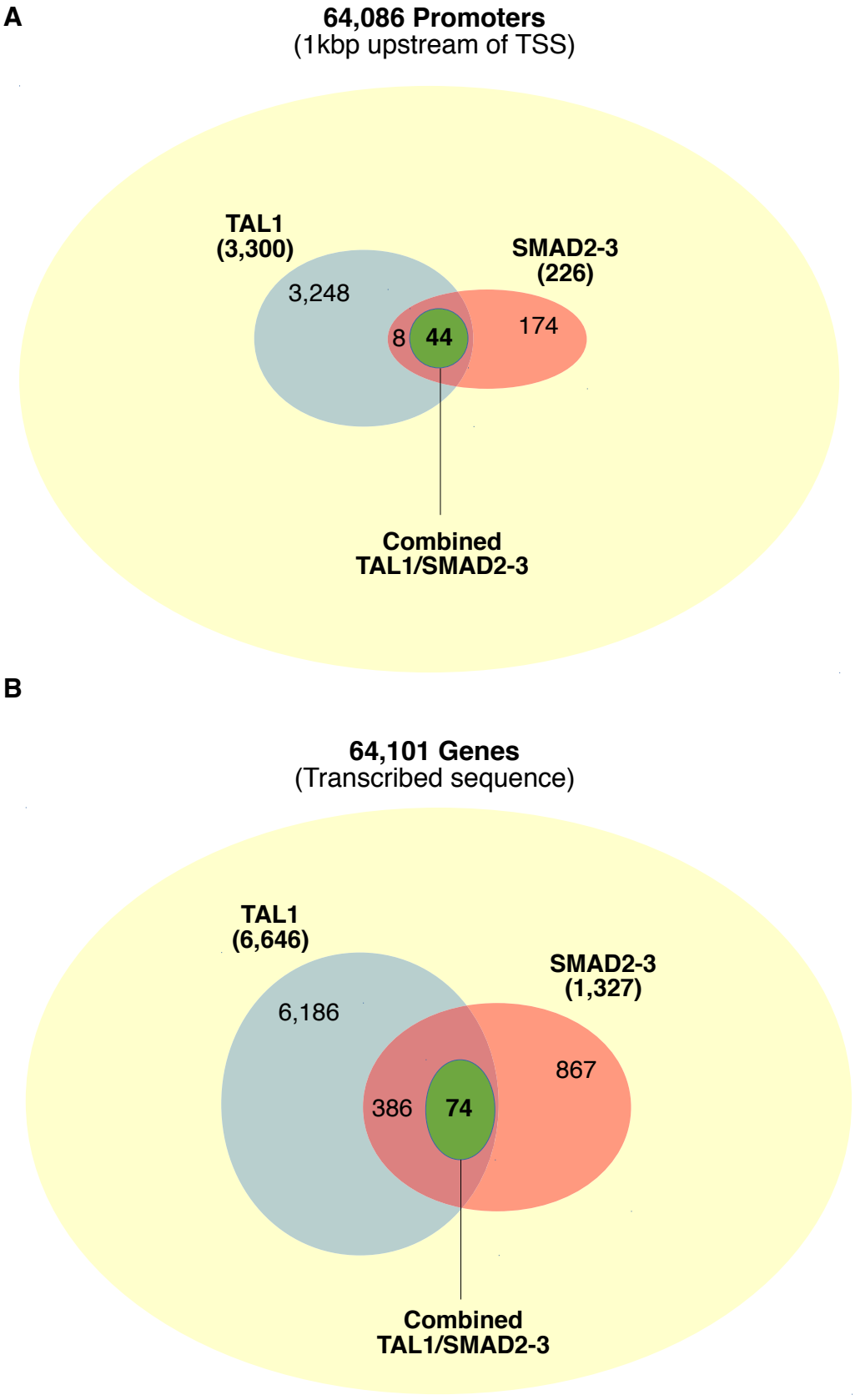
Click here to download Supplementary Material: SupplementaryFigure1TS.pdf



Supplementary Figure 1 : Immunoprecipitation of TAL1 coprecipitates SMAD3. HeLa cells were transfected with vectors expressing HA-tagged TAL1 and MYC-tagged SMAD3 as indicated. Immunoprecipitation was carried out using a monoclonal antibody to HA which was omitted in lanes 4 and 7 as control. Immublot was first analyzed with an antibody to SMAD3 (upper panel) and secondly with the antibody to HA (lower panel). Lanes 1 and 2 correspond to direct loading of an extract aliquot and a protein molecular weight marker (MWM) was loaded in lane 3. The positions of MWM bands, MYC-SMAD3 and HA-TAL1 are indicated.

Supplementary Figure 2

Click here to download Supplementary Material: SupplementaryFigure2TS.pdf



Supplementary Figure 2 : Venn diagrams of the TAL1 and SMAD2-3 ChIP-seq peaks in promoters (taken as 1,000 bp upstream of the transcription start site, **A**) and transcribed sequences (**B**) of human genes. The data were taken from the analysis performed by Tsankov et al. in HUES64 cells (GSM1505785 (TAL1) and GSM1505747 (SMAD2/3)). The green circle represents overlapping TAL1 and SMAD2-3 peaks.

Supplementary Table 1

Click here to download Supplementary Material: Supplementary Table 1.xlsx

Chromosome	associated gene name	Ensembl ID	Strand
chr1	SLC35E2	ENSG00000215790	-1
chr1	NOTCH2NL	ENSG00000213240	1
chr1	RP11-458D21.5	ENSG00000255168	1
chr1	RNVU1-18	ENSG00000206737	-1
chr1	RNA5S1	ENSG00000199352	-1
chr1	RNA5S2	ENSG00000201588	-1
chr1	RNA5S4	ENSG00000200381	-1
chr1	RNA5S5	ENSG00000199396	-1
chr1	RNA5S7	ENSG00000202521	-1
chr1	RNA5S10	ENSG00000199910	-1
chr1	RNA5S11	ENSG00000199334	-1
chr1	RNA5S13	ENSG00000202526	-1
chr1	RNA5S14	ENSG00000201355	-1
chr1	RNA5S15	ENSG00000201925	-1
chr1	RNA5S16	ENSG00000202257	-1
chr2	UCN	ENSG00000163794	-1
chr2	PELI1	ENSG00000197329	-1
chr2	SOWAHC	ENSG00000198142	1
chr2	Sep-10	ENSG00000186522	-1
chr2	MARCH7	ENSG00000136536	1
chr3	RP11-148G20.1	ENSG00000228350	1
chr3	CELSR3	ENSG00000008300	-1
chr3	SKIL	ENSG00000136603	1
chr5	ZSWIM6	ENSG00000130449	1
chr6	PHIP	ENSG00000146247	-1
chr7	AP4M1	ENSG00000221838	1
chr7	POLR2J3	ENSG00000168255	-1
chr8	REXO1L11P	ENSG00000223524	1
chr9	HABP4	ENSG00000130956	1
chr12	RP11-214K3.18	ENSG00000270095	-1
chr14	KTN1-AS1	ENSG00000186615	-1
chr14	MNAT1	ENSG00000020426	1
chr16	TCEB2	ENSG00000103363	-1
chr16	ZCCHC14	ENSG00000140948	-1
chr17	CYB5D2	ENSG00000167740	1
chr17	ZZEF1	ENSG00000074755	-1
chr17	RP5-1050D4.3	ENSG00000262429	1
chr17	ZSWIM7	ENSG00000214941	-1
chr17	RP11-744K17.9	ENSG00000266795	1
chr17	NLE1	ENSG00000073536	-1