Mutually Exclusive CBC-Containing Complexes Contribute to RNA Fate

Mutually exclusive CBC-containing complexes contribute to RNA fate

Simone Giacometti[1,2,3,4,#], Nour El Houda Benbahouche[2,#], Michal Domanski[1,5,#], Marie-Cécile Robert[2], Nicola Meola[1], Michal Lubas[1,6], Jakob Bukenborg[7], Jens S. Andersen[7], Wiebke M. Schulze[8], Celine Verheggen[2], Grzegorz Kudla[3,*], Torben Heick Jensen[1,*] and Edouard Bertrand[2,*]

[1]Centre for mRNP Biogenesis and Metabolism, Department of Molecular Biology and Genetics, Aarhus University, C. F. Møllers Allé 3, Bldg. 1130, DK-8000 Aarhus C, Denmark, [2]Institut de Génétique Moléculaire de Montpellier, Unité Mixte de Recherche 5535, C.N.R.S. and Montpellier University, 34293 Montpellier Cedex 5, France, [3]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK, [4]UCSF School of Medicine, Diabetes Center, 513 Parnassus Avenue, 94143, San Francisco, CA, USA, [5]Present address: Department of Chemistry and Biochemistry, University of Bern, Freiestrasse 3, CH-3012 Bern, Switzerland, [6]Present address: Biotech Research and Innovation Centre, University of Copenhagen, Copenhagen, Denmark, [7]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, [8]European Molecular Biology Laboratory, Grenoble Outstation, 71, Avenue des Martyrs, CS 90181, 38042 Grenoble Cedex 9, France

[#] Equal contribution

*Corresponding authors: Grzegorz Kudla (gkudla@gmail.com), Torben Heick Jensen (thj@mbg.au.dk), Edouard Bertrand (edouard.bertrand@igmm.cnrs.fr)

Running title: Nuclear fate of capped RNAs

Summary

The nuclear cap-binding complex (CBC) stimulates processing reactions of capped RNAs, including their splicing, 3'end formation, degradation and transport. CBC effects are particular for individual RNA families, but how such selectivity is achieved remains elusive. Here, we analyze three main CBC partners known to impact different RNA species. ARS2 stimulates 3'end-formation/transcription termination of several transcript types; ZC3H18 stimulates degradation of a diverse set of RNAs and PHAX functions in pre-sn(o)RNA transport. Surprisingly, these proteins all bind capped RNAs without strong preferences for given transcripts and their steady-state binding correlates poorly with their function. Despite this, PHAX and ZC3H18 compete for CBC-binding and we demonstrate that this competitive binding is functionally relevant. We further show that CBC-containing complexes are short-lived *in vivo*, and therefore suggest that RNA fate involves the transient formation of mutually exclusive CBC complexes, which may only be consequential at particular check-points during RNA biogenesis.

Introduction

All RNA polymerase II (RNAPII) transcripts undergo processing events that are essential for their function. Early during RNA synthesis, an m7-G cap is added to the nascent 5'end by an enzymatic complex that binds the Serine 5 phosphorylated form of the C-terminal domain (CTD) of RNAPII (Bentley, 2014). By protecting the nascent RNA from 5' to 3' degradation, the cap thus represents the hallmark of a successfully initiated RNAPII transcript (Furuichi et al., 1977). Importantly, the cap also serves a key role in many aspects of nuclear RNA biology (Gonatopoulos-Pournatzis and Cowling, 2014). Nuclear cap functions are mediated by the CBP80 and CBP20 proteins (also named NCBP1 and NCBP2), composing the nuclear cap-binding-complex (CBC) that associates co-transcriptionally with the nascent RNA (Glover-Cutter et al., 2008; Görnemann et al., 2005; Narita et al., 2007). CBP20 interacts directly with the m7-G cap through its classical RNA Recognition Motif (RRM), while CBP80 ensures high affinity binding of the full CBC and provides a platform for interactions with other factors (Calero et al., 2002; Izaurralde et al., 1994; Mazza et al., 2001; Mazza et al., 2002).

The CBC is highly specific for guanosine caps modified at position N7 (m7G cap). Cap-adjacent nucleotides may also carry modifications, but it is believed that these nucleotides increase CBC affinity in a rather non-sequence specific manner (Worch et al.). In the following, we therefore refer to 'capped RNA' as transcripts carrying an m7G cap, regardless of the identity or modification of the adjacent nucleotides. The CBC is believed to bind all classes of m7-G-capped RNAs, including precursors and mature forms of mRNAs, stable long non-coding (lnc) RNAs, non-adenylated histone RNAs and precursors of spliceosomal small nuclear RNAs (snRNAs). It also associates with m7-G capped forms of small nucleolar RNAs

(snoRNAs) and labile lncRNAs, such as promoter upstream transcripts (PROMPTs) (Preker et al., 2011; Preker et al., 2008). Through its cap-association, the CBC affects nuclear RNA metabolism in ways that appear specific for different RNA families. In the case of conventional mRNAs, the CBC stimulates the splicing of cap-proximal introns, the processing of RNA 3′ends and the formation of export-competent RNPs (Cheng et al., 2006; Flaherty et al., 1997; Izaurralde et al., 1994; Laubinger et al., 2008). Stimulation of RNA splicing and export has been proposed to involve intercations of the CBC with the U4/U6.U5 tri-snRNP and ALYREF, respectively (Cheng et al., 2006; Pabis et al., 2013). In the case of non-adenylated histone mRNAs, the CBC promotes their 3'end formation in a process involving interactions with the ARS2, NELF-E and SLBP proteins (Gruber et al., 2012; Hallais et al., 2013; Narita et al., 2007). In the case of PROMPTs and other short-lived transcripts, such as products of readthrough transcription, the CBC recruits ARS2, ZC3H18 and the nuclear exosome targeting (NEXT) complex, composed of RBM7, ZCCHC8 and hMTR4. This leads to the formation of the CBC-NEXT (CBCN) complex (Fig. 1A), which promotes RNA degradation via the nuclear RNA exosome (Andersen et al., 2013; Lubas et al., 2015). Finally, in the case of snRNAs, the CBC promotes transcription termination, aided by ARS2, and nuclear export of the resulting precursors (Andersen et al., 2013; Hallais et al., 2013; Ohno et al., 2000; Segref et al., 2001). The latter activity involves the so-called CBC-ARS2-PHAX (CBCAP) complex (Hallais et al., 2013) Fig. 1A), where PHAX acts as an adaptor between the CBC/RNP complex and the nuclear export receptor CRM1 (Ohno et al., 2000). PHAX and the CBC are also involved in the biogenesis of capped snoRNAs, directing the intranuclear transport of nascent snoRNAs to Cajal bodies (Boulon et al., 2004).

Such a broad collection of CBC functions raises the question of how specificity is achieved; i.e. how are different RNA families identified and brought to their proper processing machineries? This question is particularly relevant given the dual RNA-productive and -degradative effects imposed by the CBC on nuclear RNA (Andersen et al., 2013; Hallais et al., 2013). At least part of the answer lies in the different protein partners of the CBC complex (Müller-McNicoll and Neugebauer 2014). As mentioned above, distinct CBC effectors drive different processing reactions, and their recognition of particular RNA families, or even individual transcripts, could potentially provide specificity. This concept is supported by studies of snRNAs and mRNAs in Xenopus oocytes, which indicate that the protein composition of the corresponding capped RNPs is determined by the RNA length and intronic content (Masuyama et al., 2004; Ohno et al., 2002). On the one hand, introns lead to the deposition of the exon junction complex (EJC) onto spliced RNAs (Ideue et al., 2007; Le Hir et al., 2000a), and the EJC communicates with the CBC to recruit the mRNA export adaptor ALY/REF (Cheng et al., 2006). On the other hand, RNA length appears to determine whether PHAX efficiently associates with CBC-bound RNAs or not (Masuyama et al., 2004; Ohno et al., 2002). Indeed, PHAX was suggested to specifically associate with short RNAs due to its active exclusion by hnRNPC tetramers, which bind selectively to RNAs longer than 200 nucleotides (McCloskey et al., 2012). Whether this mechanism applies to all nuclear RNAs is currently unknown. How other CBC-effectors discriminate their transcript targets and how effector-target recognition translates into biological activity are also unanswered questions.

In this study, we employ transcriptome-wide *in vivo* RNA cross-linking methodology, protein-protein interaction assays, factor depletions followed by substrate analysis and fluorescence microscopy to functionally characterize three key

CBC partners, ARS2, PHAX and ZC3H18. Surprisingly, we find that the target specificities of these factors at steady-state are rather broad and therefore unable to explain the RNA family-specific activities of the CBC. In contrast, our data suggest a model where short-lived, mutually exclusive CBC-containing complexes determine RNA fate by reacting to molecular cues imposed at specific time-points during RNA biogenesis.

Results

ARS2, PHAX and ZC3H18 bind mRNA/pre-mRNA in a cap-proximal fashion

To characterize how CBC-interacting factors with different biological activities might achieve RNA family-specific effects, we first performed individual-nucleotide resolution UV cross-linking and immunoprecipitation (iCLIP) with ARS2, PHAX and ZC3H18. These proteins all bind RNA and associate with the CBC, but with distinct outcomes, providing good models to test whether substrate selectivity is accomplished by the specific recognition of RNA by CBC partners. As comparisons, we conducted iCLIP with CBP20, providing a useful baseline on which to compare CBC partners, and included our previous iCLIP analysis of the NEXT component RBM7 (Lubas et al., 2015).

For all proteins except ZC3H18, iCLIP was performed using HeLa Kyoto cell lines expressing, under the control of the respective endogenous gene promoters, 'Localization and Affinity Purification' (LAP)-tagged proteins with an N- or C-terminal GFP moiety ((Poser et al., 2008); Fig. S1A). Since a tagged ZC3H18 HeLa Kyoto cell line could not be obtained, we instead employed a C-terminally 3×FLAG-tagged ZC3H18 cDNA, which was introduced in a single copy into HEK293 Flp-In TREx cells (Andersen et al., 2013). All interrogated factors could be efficiently cross-linked to

RNA in a UV-dependent manner and extensive RNase I treatment of immuno-preciptated (IP'ed) material confirmed that the majority of RNA was attached to the relevant proteins (Fig. S1B). The 'no-tag' control cell lines yielded no detectable PCR products (Fig. S1C), implying a low experimental background. Each IP iCLIP library was produced in duplicate (Table S1) and the distribution of total mapped reads was calculated (Table S2). The replicates were generally similar to each other and different from both cytoplasmic poly(A)$^+$ RNAs and rRNA-depleted total RNAs, revealing both reproducibility and specificity (Fig 1B, Fig. S1D and Table S2).

As expected from their CBC-connections (Andersen et al., 2013; Hallais et al., 2013), ARS2, PHAX, ZC3H18 and RBM7 mainly bound to capped RNAs (Fig. 1B). CBP20 was highly enriched on 'mRNA 1st exons' (Table S2), in line with its direct binding to the cap. ARS2 and PHAX were both enriched on snRNAs and capped snoRNAs, consistent with their functions in snRNA biogenesis. However, all interrogated factors bound mRNA as their primary transcript biotype. For PHAX this was somewhat unexpected, given its reported absence from long capped transcripts in *Xenopus* oocytes (Masuyama et al., 2004; Ohno et al., 2002). Selected iCLIP substrates were however validated by regular IPs followed by RNase protection or RT-qPCR analyses (Fig. S2A, S2B, S2C), as well as by manual CLIP experiments (Fig. S2D).

Visual examination of representative examples of canonical pre-mRNAs demonstrated that CBP20, ARS2, PHAX and ZC3H18 exhibited a cap-proximal cross-linking preference (Fig. 1C and 1D). Although such tendency was also reported for RBM7 (Lubas et al., 2015), this protein associated relatively more with the bodies of the examined transcripts. To more generally assess factor binding, we employed a set of 5,769 well-annotated pre-mRNAs, containing no other annotated transcription

start sites (TSSs) or transcript termination sites (TTSs) in the interrogated regions, and calculated the fraction of iCLIP reads falling within the first 100, 500 or 1000 cap-proximal nucleotides. As expected, the CBP20 CLIP signal was highly enriched at cap-proximal positions (Fig. 1E) and consistent with the individually examined pre-mRNAs, ARS2, PHAX and ZC3H18 displayed more frequent cap-proximal reads than RBM7 or than that observed by the distribution of RNA-Seq reads, using either cytoplasmic poly(A)$^+$ RNAs or rRNA-depleted total RNAs.

To examine the maturation status of mRNAs bound by CBP20, ARS2, PHAX and ZC3H18, we next calculated the fraction of exon-intron (EI) or intron-exon (IE) junction reads in the respective libraries. Whereas RNA-Seq datasets contained mostly spliced reads, iCLIP with CBC and its binding partners recovered many unspliced transcripts, consistent with the nuclear localization of the proteins (Fig. 1F). CBP20 was most strongly enriched on spliced species, closely followed by PHAX, ARS2 and ZC3H18 (Fig. 1F). As expected, RBM7 exhibited a relatively stronger binding to IE junctions, consistent with its accumulation in the 3' ends of introns (Lubas et al., 2015).

Taking these analyses together, we conclude that CBP20, ARS2, PHAX and ZC3H18 associate with both immature and mature mRNAs with a common preference for cap-proximal binding, consistent with previous biochemical experiments (Andersen et al., 2013; Hallais et al., 2013; Izaurralde et al., 1992; Ohno et al., 1987). RBM7, on the other hand, associates with RNA in a less cap proximal fashion. Hence, besides the surprising interaction of PHAX with pre-mRNA/mRNA, we note that the distinct ZC3H18 and RBM7 binding profiles suggest that a stable CBCN complex does not readily form within nuclear pre-mRNP/mRNP.

Targeting of ARS2, PHAX and ZC3H18 to different classes of RNAPII-derived transcripts

To further characterize transcript association of the investigated factors, we first generated metagene profiles of read densities from individual libraries by anchoring sequence tags to pre-mRNA TSSs or TTSs. As expected from our previous analyses, this revealed sharp cap-proximal peaks of CBP20, ARS2, PHAX and ZC3H18, and a more moderate enrichment of RBM7 (Fig. 2A, red coloring). No major differences were observed for these proteins near the RNA 3'ends. Cap-proximal binding profiles for CBP20, ARS2, PHAX and ZC3H18 were also apparent for reverse transcribed PROMPTs (Fig. 2A, blue coloring), which became clearer when CLIP signals were anchored to PROMPT 5'ends (Fig. 2B) as defined by Cap Analysis of Gene Expression (CAGE) data (Andersson et al., 2014; Ntini et al., 2013). As for pre-mRNAs, RBM7 bound PROMPTs with a more moderate cap-proximal tendency. Interrogated proteins also accumulated close to the cap of long intergenic non-coding RNAs (lincRNAs; Fig. S3) and enhancer RNAs (eRNAs; Fig. 2C), although the low-abundant nature of the latter in the utilized exosome-proficient cells only allowed a moderate spatial signal resolution.

We next examined binding of factors to replication-dependent histone (RDH) RNAs, which are 3'end processed by U7 snRNA and therefore not polyadenylated. All the investigated proteins bound to histone mRNAs, with PHAX and ZC3H18 showing the highest fractions of CLIP reads (Fig. 2D). RDH genes also generate 3'extended transcripts that may terminate at cryptic downstream polyadenylation (pA) sites (Gruber et al., 2012). Estimating iCLIP reads mapping to such 3'extensions relative to mature RDH transcript revealed elevated RBM7 binding compared to the other factors (Fig. 2D). A similar tendency was also observed when interrogating

independently transcribed sn(o)RNAs (Fig. 2E, inset). Primary snRNA transcripts are cleaved by the integrator complex to generate pre-snRNAs carrying extensions of less than 20 nucleotides ('short 3'extensions'), which are exported to the cytoplasm by CBC and PHAX to be processed into mature TMG capped snRNAs ((Matera et al., 2007) for review). SnRNA genes also produce transcripts carrying 3'extensions of a few hundred nucleotides ('long 3'extensions') and whose degradation relies on ZC3H18 and NEXT (Andersen et al., 2013). Consistently, RBM7 binding was again elevated on long 3'extensions relative to mature RNAs (Fig. 2E, inset), but somewhat surprisingly this was not the case for ZC3H18 (see below). Finally, binding of factors to snoRNAs deriving from splicing of their host introns were analyzed, and revealed robust RBM7 binding to mature snoRNAs and their 3'extensions (Fig. 2F), consistent with NEXT-mediated decay from intronic 3'ends (Lubas et al., 2015). Interestingly, PHAX bound strongly to mature uncapped snoRNAs, whereas CBP20 and ARS2 did not, suggesting that PHAX may be recruited to these RNAs independently of CBC/ARS2.

Taking the data together, we conclude that the CBC and its partners generally bind the same families of coding and non-coding capped RNAs. However, some differences can be observed. First, RBM7 contacts unprocessed, long 3'extended snRNA and RDH transcripts, which most likely mirrors the NEXT-mediated activity of the RNA exosome on these species. Second, ARS2 and PHAX display a moderate enrichment on snRNAs as compared to e.g. CBP20, which is consistent with their role in snRNA export. This is however contrasted by their quantitatively robust binding to mRNA (Table S2). Such limited specificity of ARS2 and PHAX for snRNAs appears insufficient to faithfully identify these RNAs within the nucleus.

ARS2, PHAX and ZC3H18 display limited specificity within separate RNA families

Although ARS2, PHAX and ZC3H18 bind families of capped RNA without strong selectivity, they might still bind different RNAs within one family. To address this question, we compared iCLIP read counts for individual transcripts between relevant libraries (Fig. 3A). This analysis revealed that all the bound mRNAs (conventional and RDH RNAs), lncRNAs and sn(o)RNAs displayed largely similar binding profiles for CBP20, ARS2, PHAX and ZC3H18. To try identify differently bound RNAs, we focused on PHAX and ZC3H18 that appeared to have the most diverse sets of targets (see Fig1B above). We performed a differential expression (DE)-Seq analysis of their respective iCLIP reads, which demonstrated that out of a total of 11,514 RNAs, 79% were bound indistinguishably by the two proteins, while 7% and 14% were bound preferentially by ZC3H18 and PHAX, respectively (Fig. 3B). Most of the specific PHAX binding events occured on snRNAs, in agreement with previous analyses (Fig. 1B). We then focused on mRNAs and found that 74% of these targets were shared (Fig. 3C). Taken toghether, these analyses thus indicate that even within single RNA families, CBP20, ARS2, PHAX and ZC3H18 bind similar RNAs. This apparent lack of specificity was further confirmed by an analysis of the motifs enriched in the iCLIP reads: in agreement with binding primarily determined by cap proximity, no motifs were clearly identified other than CpG-rich streches, which are generally enriched near transcription start sites (data not shown).

We next analyzed whether transcripts of different lengths would reveal any differential binding. To this end, all analyzed capped RNAs were ranked by their length and the cumulative distribution of reads was computed (Fig. 3D, left panel). This demonstrated a preference of PHAX and ARS2 for short RNAs, while RBM7 bound preferentially longer transcripts in agreement with its enrichment on pre-

mRNAs. We then tested whether this effect was driven by all RNA families and therefore repeated the calculation after removal of snRNAs (3D, middle panel), or both snRNAs and histone mRNAs (3D, right panel). This demonstrated that these two families were largely responsible for the preferential binding of PHAX to small RNAs, leaving only limited size discrimination for the remaining transcripts.

Altogether, we conclude that CBP20, ARS2, PHAX and ZC3H18 bind similar transcripts at steady-state. For the large number of included mRNAs, we fail to detect any strong dependency on length for PHAX binding.

Steady-state RNA binding of PHAX and ZC3H18 correlates poorly with function

The surprise that PHAX and ZC3H18 bind similar RNAs despite having differently reported led us to ask whether the steady-state binding of these proteins correlated with transcript change upon factor depletion. Hence, we depleted PHAX or ZC3H18 by RNAi in HeLa cells and profiled the resulting mRNA contents by RNA-Seq (Fig. 3E). A DE-Seq analysis against a control siRNA revealed that 422 mRNAs were significantly affected by ZC3H18 depletion, while none were significantly affected by PHAX depletion, and this despite similar depletion efficiencies (Log2 ratios of -2.4 and -1.7 for ZC3H18 and PHAX, respectively). This lack of effect of PHAX depletion on mRNAs was consistent with its known function as a pre-snRNA export factor but not with its iCLIP RNA binding profile, which displays robust mRNA binding.

We then considered separately the mRNAs that were preferentially bound by PHAX or by ZC3H18 (see Fig. 3C). However, a similar fraction of mRNA was sensitive to the depletion of ZC3H18 regardless of their binding preference (Fig. 3E), and a similar percentage of mRNA sensitive to ZC3H18 depletion was also identified in the entire mRNA population (Fig. 3E). We conclude that the steady-state RNA

binding profiles of PHAX and ZC3H18 correlate poorly with protein function at the genome-wide level.

ARS2 and ZC3H18 link the CBC to NEXT

A way to rationalize that the interrogated factors largely share RNA targets, yet have different effect, would be that these proteins are part of the same complex. However, while previous analyses showed that the CBCA complex can interact with PHAX (forming CBCAP; (Hallais et al., 2013)), and with ZC3H18 and NEXT (forming CBCN; (Andersen et al., 2013)), no interactions have been reported between PHAX and ZC3H18/NEXT.

Thus, to clarify these physical links further, we first determined protein-protein interactions between factors by performing pair-wise two-hybrid assays of the human proteins in yeast cells (Y2H). As expected, robust interactions were detected between RBM7 and ZCCHC8 as well as between ZC3H18 and ARS2 (Table I). Interactions of the CBC were monitored by co-expressing untagged CBP20 with CBP80 fused to the GAL4 DNA binding domain, together with the various preys fused to the GAL4 activation domain (Hallais et al., 2013). Using this strategy, we detected the expected interactions of the CBC with ARS2, PHAX and NELF-E, a protein previously shown to directly interact with the CBC and used as a positive control (Narita et al., 2007). Interestingly, a weak interaction was also detectable between the CBC and ZC3H18 (Table I). To gather more data, we used human, *Drosophila* and *Arabidopsis* ARS2 as well as human ZC3H18 as baits, and performed Y2H screens of cDNA libraries of matched species. This recapitulated the ARS2-ZC3H18 interaction with *Drosophila* factors and revealed two novel interactions: i) between the *Arabidopsis* homologs of ARS2 and PHAX; and ii)

between human ZC3H18 and ZCCHC8. The latter result was supported by the identification of a fragment located at the end of ZC3H18 (amino acids 746-953), which was sufficient to confer a robust interaction with ZCCHC8 in Y2H assays and co-IP experiments (Table I; Fig. S4A). The detected links of ARS2/ZC3H18 to the CBC, and of ZC3H18 to the NEXT component ZCCHC8 suggested a collective interpretation of the Y2H results as depicted in Fig. 4A. Consistent with previous affinity capture/mass spectrometry (AC/MS) and *in vitro* protein-protein interaction data (Andersen et al., 2013; Hallais et al., 2013; Lubas et al., 2011), the CBC and NEXT complexes constitute separate entities with no apparent direct interaction. Instead, contact between CBC and NEXT appears to be mediated by ZC3H18 and ARS2. Moreover, PHAX, like ZC3H18, is capable of interacting with the CBC and ARS2 (Fig. 4A, (Hallais et al., 2013)).

To substantiate the Y2H interaction results, we conducted a RBM7-LAP co-IP experiment and interrogated the ability of this NEXT component to associate with CBC-related factors in the presence or absence of ARS2, PHAX or ZC3H18. Western blotting analysis of input samples from HeLa RBM7-LAP cells revealed that these three components were downregulated by administration of specific siRNAs, relative to control (CTRL) siRNAs (Fig. 4B, lanes 1-4). RBM7 efficiently co-IP'ed CBP80, ARS2, ZC3H18 and the NEXT component ZCCHC8, whereas PHAX was undetectable (Fig. 4B, lane 5). Consistently, depletion of PHAX did not change the RBM7 interaction pattern (Fig. 4B, lane 7). In contrast, depletion of either ARS2 or ZC3H18 significantly decreased RBM7's interaction with CBP80 (Fig. 4B, compare lanes 5, 6 and 8). Moreover, the ARS2-RBM7 association was lost upon ZC3H18 depletion and the contact between RBM7 and ZC3H18 was moderately affected by ARS2 depletion. None of the RNAi experiments affected the ability of the RBM7-LAP

fusion to be captured by bead-bound GFP antibodies or its precipitation of the NEXT partner ZCCHC8. These results support the protein interactions suggested by the Y2H data and position ARS2 and ZC3H18 as critical factors bridging the CBC with the NEXT complex (Fig. 4B, right panel).

The inability of RBM7 to IP PHAX (Fig. 4B), and the absence of PHAX in IP's of NEXT components and ZC3H18 (Andersen et al., 2013), suggested that the majority of cellular NEXT/ZC3H18 and PHAX might reside in separate protein assemblies. Consistent with this notion, a PHAX-3xFLAG AC/MS experiment efficiently detected ARS2, CBP80 and CBP20, but failed to detect ZC3H18, ZCCHC8 and RBM7 (Fig. 4C, Table S3). Human MTR4 was detected in low, yet significant, yields, which likely reflects its interaction with the exosome, the core subunits of which were detected at similar quantities (Fig. 4D).

PHAX and ZC3H18 compete for the CBC

Given their mutual exclusive presence in IP eluates, we considered that PHAX and ZC3H18 might compete for binding to the CBC. To investigate this possibility, RBM7-LAP interacting proteins were immobilized on GFP antibody-conjugated beads and challenged by increasing amounts of recombinant human PHAX produced in *E. coli*. In vitro, this recombinant protein was able to form a stable complex with the CBC (Fig. S4B). In control experiments without addition of exogenous protein or with 40 μg of added BSA, RBM7-LAP was retained on beads with CBP20, CBP80, ARS2, ZC3H18 and hMTR4 (Fig. 5A, left panel lanes 4 and 6). In contrast, addition of PHAX caused CBP20, CBP80 and ARS2 to be dissociated in a concentration-dependent manner, whereas ZC3H18, and hMTR4 remained bead-bound with RBM7-LAP (Fig. 5A, left panel lanes 5-12). Thus, exogenous PHAX was capable of breaking the link

between ZC3H18/NEXT and the CBC (Fig. 5A, right panel), suggesting a competition between PHAX and ZC3H18 for binding the CBC.

Further support for this idea was obtained by employing the Lumier assay, which yields a quantitative measure of the *in vivo* interaction between two proteins of interest (Fig. 5B, left panel). A construct harboring CBP20 fused at its N-terminus to the Firefly Luciferase (FFL) protein and 3xFLAG (3xFLAG-FFL-CBP20) was transfected into HEK293T cells together with a construct expressing either PHAX (RL-PHAX) or ZC3H18 (RL-ZC3H18) N-terminally fused to the Renilla Luciferase protein. Subsequently, whole cell extracts were subjected to anti-FLAG IPs and luciferase activities were measured in both the input extracts and their IP pellets. As a measure of interaction specificity, RL was first plotted as fold enrichment over control beads with no FLAG antibody, confirming that both RL-PHAX and RL-ZC3H18 exhibited robust interaction with 3xFLAG-FFL-CBP20 (Fig. 5B, right panel). These interactions were then challenged by overexpression of putative competitor proteins (Fig. S4C). Consistent with the proposed CBCN architecture (Fig. 4A), overexpression of NEXT components had no effect on the ZC3H18-CBP20 interaction (Fig. 5C, right panel). A similar result was obtained employing hnRNPC, another proposed CBC binder (McCloskey et al., 2012). However, in agreement with the *in vitro* experiments of Fig. 5A, overexpression of PHAX readily displaced ZC3H18 from CBP20. ARS2 overexpression also decreased the interaction, possibly by titrating ZC3H18 from a CBC/ARS2/ZC3H18 ternary assembly. Challenging the PHAX-CBP20 interaction in a similar manner revealed that overexpression of NEXT components and hnRNPC again had no effects (Fig. 5D), whereas overexpression of ZC3H18 diminished the PHAX-CBC20 contact. Overexpression of ARS2 also

displaced PHAX from CBP20, which again could be due to a titration of PHAX from the CBC-ARS2-PHAX complex.

Based on all of our data, we suggest that NEXT contacts the CBC through Z3CH18 and ARS2, and that the formation of CBC-ARS2-PHAX and CBC-ARS2-ZC3H18 is mutually exclusive.

PHAX and ZC3H18 have opposite effects on RNA levels

Whereas our CLIP data showed that ZC3H18 and PHAX associate with the same set of RNAs, our biochemical experiments demonstrated that these factors cannot simultaneously bind the CBCA complex. This suggests that an RNA bound by CBCA may transition between complexes containing either ZC3H18 or PHAX. If these proteins elicit different functional outcomes, RNA fate might then be dictated by which RNP complex is favored at the time this 'decision' has to be taken. To address the validity of this hypothesis, we first employed a tethering assay to explore the functional consequences of binding PHAX or ZC3H18 to an RNA reporter. Hence, we fused ZC3H18 or PHAX to the MS2 coat protein (MCP), which itself was fused to GFP (MCP-GFP-X), and co-expressed one of these fusion-proteins together with a plasmid expressing an RL RNA reporter carrying two MS2 binding sites in its 3'-UTR as well as a FFL control RNA to adjust for transfection efficiencies (Fig. 6A). Tethering of ZC3H18 decreased RL expression, which was likely due to recruitment of the NEXT complex, since tethering of the ZC3H18$^{746\text{-}953}$ fragment, sufficient for ZCCHC8-interaction (Table 1; Fig. S4A), had a similar effect (Fig. 6B, left panel). In stark contrast, tethering of PHAX induced a robust increase in RL activity. These effects were also reflected at the level of RL RNA (Fig. 6B, right panel).

To test the effects of PHAX and ZC3H18 on endogenous RNAs, we turned to snRNAs, whose long 3'extended species are known to be degraded by the exosome in an ZC3H18/NEXT dependent manner (Andersen et al., 2013), providing useful model substrates. As expected, depleting ZC3H18 generally increased levels of 3'extended RNAs derived from eight different snRNA genes and the capped U3 snoRNA gene (Fig. 6C; see depletion efficacy in Fig. S5). In contrast, levels of the same substrates generally decreased upon PHAX depletion, whereas co-depletion of PHAX and ZC3H18 cancelled the effects of the individual depletions, which was also evident when averaging all snRNA substrates (Fig. 6C, 'All snRNAs'). Interestingly, the effect of co-depletion was not always simply the addition of the individual depletion effects. For instance, depletion of ZC3H18 had little effect on U1.1 3'extended transcripts. Still, it completely cancelled the negative effect of depleting PHAX, suggesting that ZC3H18 had gained access to these RNAs in the absence of PHAX. Thus, the absence of one protein sensitized transcripts to the presence of the other. This is in line with a model where ZC3H18 and PHAX compete for RNA bound by CBCA to yield opposite functional outcomes.

PHAX and ZC3H18 exchange rapidly on the CBC in vivo

The idea that CBCA-bound RNPs might transition between CBCA-PHAX and CBCA-ZC3H18 forms implies that PHAX and ZC3H18 do not simply bind and 'mark' RNPs for different destinies. It also implies that PHAX and ZC3H18 rapidly exchange on and off the CBC. To test this prediction, we employed a LacO/Laci co-recruitment assay (Hallais et al., 2013), to measure the lifetime of these interactions in living U2OS cells. We tethered CBP20 to an array of genomic LacO sites, by fusing it to a red fluorescent version of the Laci protein (mRFP-Laci-CBP20). Transfected cells

displayed a diffuse nuclear mRFP-Laci-CBP20 signal in addition to a concentrated bright spot, corresponding to the location of the LacO array ((Hallais et al., 2013); Fig. S6A). We next tested whether the mRFP-Laci-CBP20 'spot' would recruit its various partners. Indeed, co-transfected GFP-tagged versions of CBP80, ARS2, PHAX and ZC3H18 concentrated in mRFP-Laci-CBP20 spots (Fig. S6A, left and right panels). This recruitment was specific as the proteins were not enriched in a control spot formed by mRFP-Laci-KPNA3 (Fig. S6B). We could also demonstrate that ARS2, PHAX and ZC3H18 interactions were dependent on RNA as a mutant form of CBP20 that does not bind the cap (F83A F85A; (Mazza et al., 2001)), failed to recruit these proteins, and yet did not prevent CBP80 interaction as expected (Fig. S6C). In agreement with these results, we detected polyA+ RNA accumulating in the mRFP-Laci-CBP20 spot (Fig. S7C).

Having established a functional experimental design, we employed 'fluorescent recovery after photobleaching' (FRAP) to measure the dynamics of mRFP-Laci-CBP20 interactions with its GFP-tagged partners. After photobleaching the LacO spot, the mRFP-Laci-CBP20 fluorescence showed very slow recovery over a two minutes time-course, indicating stable binding of the fusion protein to the LacO array (Fig. 7A, right panel). GFP-CBP80 recovered quickly when photobleached in the nucleoplasm, but only slowly (within minutes) in the mRFP-Laci-CBP20 spot, consistent with a stable interaction between these CBC subunits *in vivo*. In contrast, ARS2 and PHAX recovered quickly when photobleached in the Laci-CBP20 spot, with half-times of recovery of only a few seconds (Fig. 7B and 7C). Still, these kinetics were slower than recovery in the nucleoplasm, suggesting that dissociation of ARS2 and PHAX from the CBC is slower than the time it takes these molecules to diffuse through the bleached spot. Because ZC3H18 interacted with itself in the co-

recruitment assay (Fig. S6D), we performed the FRAP assay by tethering mRFP-Laci-ZC3H18 to the LacO array. This ensured that the photobleaching of GFP-CBP80 only measured the interaction between this protein and tethered ZC3H18 (see Experimental Procedures). This revealed a rapid (within seconds) recovery of the GFP-CBP80 signal to the spot formed by mRFP-Laci-ZC3H18 (Fig. 7D).

Modeling of the FRAP data showed that the lifetime of the CBP20-CBP80 interaction was in the order of minutes, whereas the lifetime of CBP20 interactions with ARS2, PHAX or ZC3H18 was much shorter and in the range of 3-13 seconds (Table S4).

Discussion

Eukaryotic cells produce various types of RNA that each follow a certain processing/decay and/or transport pathway. How proper transcript sorting into appropriate pathways occurs is a fundamental but incompletely understood problem. As the CBC promotes the processing of different RNAs, yielding family-specific effects (Gonatopoulos-Pournatzis and Cowling, 2014; Müller-McNicoll and Neugebauer 2014), it provides an interesting model to study the concept of RNA sorting. It has been suggested that such family- or transcript-specificity derives from CBC partners binding only certain RNAs, hereby acting as identity marks (Ohno et al., 2002). Our results do not support this idea, but instead suggest an alternative model where early RNP complexes are constantly remodeled, and determine RNA fate by reacting to external input at specific times during RNA biogenesis.

Binding of some landmark RNA binding proteins (RBPs) is promiscuous and not sufficient to define RNA maturation pathways

Early studies in *Xenopus* oocytes demonstrated that distinct RNA families use non-overlapping nuclear export pathways (Jarmolowski et al., 1994). Consistently, it was found that pre-snRNAs and mRNAs use distinct exportins and export adaptors: PHAX/CRM1 for pre-snRNAs (Ohno et al., 2000), and TAP, in association with ALYREF or other RBPs, for mRNAs (Björk and Wieslander, 2014; Grüter et al., 1998; Segref et al., 1997). Such specificity for a given export pathway appeared to stem from specific binding of key RBPs, such as PHAX or the EJC, to pre-snRNAs and spliced mRNAs, respectively (Ohno et al., 2002). This further suggested the possibility that RNA identity could be determined early on in the nucleus, perhaps even during transcription, and then stably maintained due to specific RNA-coating by certain RBPs. The iCLIP data presented here do not support this hypothesis. This is because we detect binding of PHAX not only to pre-snRNAs as expected, but also to a large range of other capped RNAs, including PROMPTs, eRNAs, lincRNAs, RDH RNAs and polyadenylated mRNAs. In fact, the fraction of total PHAX iCLIP reads mapping to mRNA approaches 40%, and is not restricted to particular mRNA species; not even to short transcripts as would perhaps have been predicted. When compared to CBP20, which expectedly binds to all capped RNAs, PHAX exhibits some preference for pre-snRNAs, but this specificity is moderate. With the notable exception of intronic snoRNAs, it is also important to note that binding of PHAX to RNA is likely to occur mainly through the CBC, which can be appreciated by the largely cap-proximal binding of the protein (see Fig. 1E and Fig. 2). The limited target specificity of PHAX is thus probably not due to promiscuous RNA binding, but rather to its loading onto RNA via cap-bound CBC. Binding of even a key RBP like PHAX is therefore poorly discriminating. It may even be argued that PHAX is a bona fide mRNA binding protein and that it could have a previously unnoticed role in mRNA

biogenesis. However, PHAX depletion revealed little effect on steady-state mRNA levels or splicing patterns in transcriptome-wide experiments . Furthermore, steady-state binding of PHAX and ZC3H18, as determined by iCLIP, correlated poorly with effects on RNA levels upon depletion of these proteins (see Fig. 3E). Using PHAX and ZC3H18 as a paradigm, we therefore suggest that binding specificity per se may generally not be sufficient to identify RNAs and determine their fate. A notable exception may be the EJC, which binds stably to spliced RNA and thus provides a more definitive identity mark (Le Hir et al., 2000a; Le Hir et al., 2000b). However, the EJC is deposited as a result of splicing, and it is thus a stable label for a transient phenomenon, much like the polyA tail is for 3'end processing.

Mutually exclusive formation of CBC complexes at specific maturation checkpoints may determine RNA fate

Live cell imaging of RBPs has demonstrated their transient interaction with RNA, allowing rapid sampling of sequences (Phair and Misteli, 2000). In agreement, our FRAP data show that CBC-containing complexes are quite labile, with a half-life of only a few seconds. With RNAPII elongation rates of about 2 kb/min (Boireau et al., 2007; Jonkers et al., 2014), a medium-sized human gene takes ~50 min to transcribe. Splicing and mRNA export also takes minutes (Audibert et al., 2002; Beyer and Osheim, 1988; Schmidt et al., 2011). This suggests that PHAX and ZC3H18 continuously exchange at the CBC-bound cap during RNA production. Thus, instead of using steady-state binding as a mechanism to identify RNAs and control their fate, many RBPs, including PHAX and ZC3H18, might be part of a 'hit-and-run' mechanism, where transcript fate would originate from ‘locking’ of decisive complexes only at particular checkpoints during pre-mRNA processing. The ability of

RNPs to form mutually exclusive complexes with proteins having opposing activities may reflect the need of the RNP to keep all options open until one outcome would have to be selected out of several possibilities. Indeed, it may simply reflect the fact that RNAPII ‘does not know’ which type of transcription unit it is engaged with until relevant cues are instigated.

We suggest that one such cue, or checkpoint, may occur when a 3'end processing signal emerges from the RNAPII exit channel. Processing signals drive the assembly of specific proteins, which may then synergize with the CBC to lock the proper complex and produce the required outcome. In support of this model, CBCA was shown to stimulate the usage of a range of 3'end processing signals (Hallais et al., 2013). Moreover, NEXT complex components purify with 3'end processing factors (Shi et al., 2009). Thus, a cryptic, cap-proximal 3'end/termination signal might promote an interaction between the CBCA complex at the RNA 5'end with NEXT at the 3'end, via ZC3H18. This would stabilize the CBCN complex, which would serve to exclude PHAX while simultanuously increase the access of NEXT and the exosome to the RNA 3'end. Example substrates for such a scenario would be PROMPTs, whose early termination and degradation rely on promoter proximal polyA sites as well as the CBCA, NEXT and exosome complexes (Andersen et al., 2013; Ntini et al., 2013). In contrast, the 3'end processing signal of an snRNA would recruit the Integrator complex (Baillat et al., 2005), which might bias the competition between PHAX and ZC3H18 toward the formation of the CBCAP complex (Hallais et al., 2013), excluding ZC3H18/NEXT and resulting in productive 3'end formation. If proper 3’end formation is missed, as e.g. in the case of ‘long 3’extended’ sn(o)RNAs, downstream cryptic termination sites might again favor CBCN formation and transcript decay.

In this study, we have focused on RNA transport via PHAX and RNA decay via ZC3H18/NEXT. Because the CBC has many activities, it is however likely that dynamic exchanges of mutually exclusive protein complexes at RNA caps may also interplay with other processing events, such as for instance RNA splicing. We propose that the constant remodeling of CBC-associated complexes allows the dynamic integration of a diverse source of signals, whereas a pre-determined, rigid CBC complex, deposited for instance at the start of transcription, would not allow such regulation.

Experimental Procedures

Cell culture

HeLa, U20S and HEK293 cells were grown in Dulbecco's modified eagle medium (DMEM) containing 10% fetal bovine serum (FBS) and 1% Penicilin/Streptavidin, at 37°C, 5% CO2. For iCLIP experiments, epitope-tagged proteins were expressed in two cell systems: (i) HEK293 Flp-In T-Rex cells stably expressing C-terminally 3XFLAG–tagged ZC3H18 under control of a tetracycline-inducible promoter or (ii) HeLa Kyoto cells stably expressing LAP-tagged proteins (ARS2, CBP20, PHAX) from integrated BACs (Poser et al., 2008). The expression of 3xFLAG-ZC3H18 was induced by replacing cell growth media with fresh media containing Tetracycline. For the RNA IP and manual CLIP experiment, 3x-Flag-ARS2 and 3x-Flag-PHAX were stably expressed in HeLa cells from clones generated by Flp-In recombination (Hallais et al., 2013).

Plasmids

DNA cloning was performed using standard techniques and the Gateway™ system (InVitrogen). The two-hybrid plasmids were based on pACTII, p422 and pAS2dd (Hallais et al., 2013). For the LUMIER competition assay, the bait and prey were expressed on non-replicative plasmids (without the origin of replication of SV40), while the competitor was on a replicative plasmid (pcDNA3.1). The competitors were N-terminally fused to a myc tag. The cDNAs were all of human origin except for PHAX which was a mouse cDNA. For the MS2 tethering experiments, the reporter plasmid was based on PSICHECK-2. For the LacO tethering assay, the proteins of interest were expressed from the mouse L30 promoter, as C-terminal fusions with

GFP or mRFP-Laci (Hallais et al., 2013). Detailled maps and sequences are available upon requests.

siRNAs and RNA-Seq analysis

The utilized siRNAs had the following sequences: 5' CAACAAGCAUGCAGAGAAAdTdT (siARS2); 5' CUUACGCUGAGUACUUCGAdTdT (siControl against firefly luciferase); 5' UAAAUCCUGUGCUAUAUACUCdTdT (siPHAX); 5' GGAAUGAAUUGUAGGUUUAdTdT (siZC3H18). For the RNA-Seq experiments, the PHAX siRNAs were 5′ -UAGUAUCAGCGAGGAACAAAUUA dT dT and 5′-AAGAGUAUAUAGCACAGGAUUUA dT dT. For the LUMIER assays, the control siRNA was 5' CAACAGAAGGAGAGCGAAA dT dT.

Cells were transfected for 3 days using Lipofectamin2000 (20 μl/ml in the transfection mixture, together with 0.4 μM of siRNA), and with at a final siRNA concentration of 20 nM in the cell culture medium. For RNA-Seq experiments, HeLa cells were transfected with PHAX, ZC3H18 or control siRNAs using JetPrime (PolyPlus), and cells were harvested 48h later. RNAs from triplicate experiments were prepared using TRIzol (Thermo Fisher), and rRNA were removed using the RiboMinus kit (Thermo Fisher). RNAs were sequenced using paired-end sequencing, mapped against the human genome, and analyzed using the DE-Seq package in R. Raw sequence data are available in GEO (accession number XXXXX).

Antibodies

Antibodies were from the following sources, and used for western blotting analysis at the indicated dilutions. Anti-ARS2: Abcam (Ab88392; 1:1000); anti-GFP: Santa Cruz (sc-9996; 1:250); anti-hMTR4: Abcam (Ab70551; 1:2500); anti-RBM7: Sigma-Aldrich

(HPA013993); anti-ZCCHC8: Abcam (Ab8739; 1:500; anti-ZC3H18: Sigma-Aldrich (HPA040847; 1:500). Anti-CBP20 and CBP80 were gifts from E. Izaurralde and were used at 1:1000 and 1:10000, respecively. Anti-PHAX was a mouse monoclonal antibody (Hallais et al., 2013), and used at 1:500.

iCLIP

The iCLIP approach was performed as described in (Konig et al., 2011) with the additional modifications of (Lubas et al., 2015), which include differences in sonication and washing buffers. All iCLIP experiments were performed in two biological replicates. Briefly, cells were crosslinked with 300 mJ/cm2 at 254nm UV. Upon cell lysis, RNAs were partially fragmented using low concentrations of RNase I, and protein-RNA complexes were IP'ed with antibodies (single chain GFP-Trap nanobodies for LAP-tagged cells and α-FLAG M2 antibodies for 3×FLAG-ZC3H18 cells) coupled to magnetic beads (Dynabeads M-270 Epoxy, Life Technologies). Beads were washed three times in 50 mM Tris/HCl, pH 8.0, 150 mM NaCl 0.5% (v/v) TritonX100. An additional step of on-beads RNase I treatment was performed, and beads were washed again in 50 mM Tris/HCl, pH 8.0, 2000 mM NaCl, 0.5% (v/v) TritonX100, 2M Urea for GFP-tagged proteins, or 50 mM Tris/HCl, pH 8.0, 1000 mM NaCl, 0.5% (v/v) TritonX100, 1M Urea for 3xFLAG-ZC3H18 CLIPRNAs were then radioactively labelled, subjected to denaturing gel electrophoresis and transferred to a nitrocellulose membrane, to remove RNAs that were not covalently linked to proteins. RNAs were recovered from membranes by proteinase K digestion and reverse transcribed. cDNA molecules were size-purified using denaturing gel electrophoresis, PCR-amplified and subjected to high-throughput sequencing.

Bioinformatics analysis of iCLIP data

High-throughput sequencing of iCLIP cDNA libraries from two replicate experiments, of each interrogated factor, was performed by Illumina HiSeq 2000 sequencing. Sequenced reads were trimmed of the fixed 3'adaptor and filtered by quality and insert length using Flexbar (http://sourceforge.net/projects/flexbar/). The reads were collapsed and demultiplexed based on their 5' adaptor sequences, which contained a 4-nt experiment-specific barcode for sample multiplexing and a random 5 nt region to control for PCR artefacts. Reads were mapped to the hg19 human genome assembly using TopHat (Trapnell et al., 2009) with a known splice junction database (-GTF option) from iGenomes (http://support.illumina.com /sequencing/sequencing_software/igenome.ilmn). Reads mapping to the same position with the same counting barcode were collapsed, and reads mapping to multiple independent genomic locations were assigned randomly to one of these locations, whereas reads mapping across splice junctions were assigned to both sides of the junction. Genomic annotations were assigned based on gene annotations from the UCSC genome browser and from published datasets (Andersson et al., 2014; Cabili et al., 2011; Derrien et al., 2012; Kishore et al., 213; Ntini et al., 2013).

Bedtools (Quinlan and Hall, 2010) were used for analyses of read coverage in genomic feature. For calculation of coverage piecharts, reads were uniquely assigned to RNA biotypes using a hierarchical procedure, starting with the most abundant biotypes. For metagene analyses of coverage around mRNA TSSs and TTSs, we used a filtered subset of 5,769 well-annotated Refseq mRNAs that contained no known rRNA, tRNA, miRNA or sn(o)RNA in the mRNA introns or within 2 kb of the transcript boundaries.

To compare the CLIP data with total RNA abundances, we used representative RNA-Seq datasets downloaded from the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra). We used cytoplasmic poly(A)$^+$-selected data from HeLa (SRR3479116; (Lykke-Andersen et al., 2014)) and HEK293 (SRR1275413), as well as rRNA-depleted total RNA from HeLa (SRR1014903) and HEK293 cells (SRR2096982). It is worth noting that all commonly used protocols for RNA-Seq library preparations have specific biases regarding the length and modification status of RNAs, so that discrepancies between RNA-Seq and iCLIP profiles are expected. RNA-Seq data were analyzed with the same pipeline as iCLIP, except for the demultiplexing step based on the random barcode information, which was only used for iCLIP.

RNA IP and CLIP validation assays

For the IPs of Fig. S2A-S2C, cells were rinsed in PBS, scraped and lysed for 30 min at 4°C in IP buffer (50mM Tris-HCl pH 7.4, 150mM NaCl, 1mM EDTA, 1mM MgCl2, 15% glycerol, 1% NP40, anti-protease cocktail from Roche). Cellular debris was removed by centrifugation (10 min at 9000g), and extracts were incubated with antibody-coated beads (2 to 4 h at 4°C). Beads were then washed five times in IP buffer and resuspended in TRIzol (Invitrogen). The antibodies used were monoclonal α-PHAX antibodies (clones 10E3 and 17E11; (Hallais et al., 2013)) and a monoclonal α-FLAG (M2, Sigma), coupled to protein G sepharose (4 fast flow; GE Healthcare).

For the manual iCLIP experiments of Fig. S2D, cells were UV-crosslinked on ice, using two irradiations of 700 J each. Cells were then lyzed in RBS500 buffer (10mM Tris-Hcl, pH 7.5, 500 mM NaCl, 2.5 mM MgCl2, 0.5% Triton, 1% Empigen-BB, anti-protease cocktail). Empigen-BB is a charged detergent. Extracts were

incubated 3h at 4°C with beads coated with M2 α-FLAG antibody. Subsequently beads were washed 5 times in RBS500, and resuspended in TRIzol (Invitrogen) to extract RNAs. Controls were constituted by the same amount of extract of HeLa cells not expressing 3XFLAG-PHAX.

RNase protection and RT-qPCR assays

RNA was purified using TRIzol according to the manufacturer's instructions. RNAse protection assays were performed with the RNAPIII kit (Ambion), using ^{32}P labeled probes spanning the 3'end of HIST1H4B or RPS28 mRNAs. The RPS28 probe also contained an oligodT stretch to specifically detect polyadenylated RPS28 mRNA.

For RT-qPCR analysis, RNAs were treated with DNase RQ1 (Sigma) for 1 h at 37°C to digest residual genomic DNA. Reverse transcription (RT) was performed using the SuperScript II RT enzyme (Invitrogen) and N6 random priming oligos. qPCR was performed using SYBR Green PCR Master Mix (Roche) or a previously described SYBER Green mix (Lutfalla and Uze, 2006). RT-qPCR reaction were normalized to the mean of the values obtained with 18S and 28S amplicons. The oligo used were the following. Primer_U1.1_F: 5' TTACCTGGCAGGGGAGATAC; Primer_U1.1_R: 5' GCAGTCGAGTTTCCCACATT; U1.1 -RT _F: 5' GTGAAGTCCGCTCAGCTCTT; U1.1 -RT _R: 5' TGGAAGCAGAGGCTGTGTAA; U2.1-F: 5' ATCCGAGGACAATATATTAAATGGA; U2.1-R: 5' CGTTCCTGGAGGTACTGCAA; U2.1 -RT _F: 5' CCTTGAGGTTCCTGATGTGC; U2.1 -RT _R: 5' ATCCTAAGGACCTCCCCAAA; U4.2-F: 5' GCAGTATCGTAGCCAATGAGG; U4.2-R: 5' TGCCAATGCCGACTATATTT; U4.2 -RT_F: 5' GCAGGTTGTGTCTTATGTTTGG; U4.2 -RT_R: 5' AGAACCCCGGACATTCAATC; 18S-F: 5' TGCCCTATCAACTTTCGATG; 18S-R:

5'CTTGGATGTGGTAGCCGTTT; 28S-F: 5' GGGTATAGGGGCGAAAGACT; 28S-R: 5' CGCTTTACCGGATAAAACTGC

AC/MS and IP analyses

For AC/MS analysis, we used HEK293 Flp-In T-Rex cells stably expressing C-terminally 3XFLAG-tagged PHAX under control of a tetracycline-inducible promoter. Cryogenic disruption of cells and 3XFLAG-AC methodology were performed as previously described (Andersen et al., 2013; Domanski et al., 2012).

AC was performed in extraction buffer consisting of 150 mM NaCl, 0.5% Triton X-100, 20mM HEPES pH7.4 and supplemented with protease inhibitor. Experiments were performed label-free and in triplicates. Prior to AC, cell extracts were treated with 100µg/ml RNase A. Elution of bait-captured proteins was performed by mixing in 50µl of 0.5M Acetic Acid for 10 min at RT. Collected eluates were neutralized with 5µl of 5.5M Ammonium Hydroxide. Samples were concentrated to 30µl and processed using the FASP protocol (Wiśniewski et al., 2009). Trypsinized samples were acidified with 0.1% TFA, desalted using C18 stage tips and analyzed by MS using an LTQ Orbitrap Velos instrument (Thermo Scientific). Data acquisition, processing and plotting were performed as described (Hubner and Mann, 2011). The full dataset is accessible in the Table S3.

IP experiments were performed essentially as described (Domanski et al., 2012). For the PHAX competition assay, CBCN assembly was first immobilized on the magnetic beads by co-IP of RBM7-LAP (as above). Next, the beads were divided into 5 tubes, resuspended in 10 µl of the extraction buffer only, extraction buffer containing BSA (40 µg) and extraction buffer containing increasing amount of PHAX (10-40 µg). Samples were incubated for 20 min at RT with mixing (1100 rpm,

thermomixer). After collecting the supernatant (S) beads were washed once with the extraction buffer and remaining protein material was eluted (E) with 1x LDS protein loading buffer. Proteins were separated by SDS-PAGE followed by transfer onto PVDF membrane. To avoid unspecific signal, the membrane was blocked in 5% skimmed milk / PBS-T for 1 hour at RT. Next, the membrane was cut into pieces for incubation with the following antibodies: anti-ZC3H18, -ARS2, -hMTR4, -CBP80, -CBP20 and -GFP. Incubation with the primary and HRP-conjugated secondary antibodies (both in 5% skimmed milk / PBS-T) was performed for 1 hour at RT. The membrane was washed in between 3 x 5 min with PBS-T. After incubation with the ECL substrate, chemiluminescence was detected using X-Ray film.

For the co-IP combined with RNAi, HeLa RBM7-LAP cells were seeded on 150 mm dishes at ~5x10^6 cells in 20 ml of DMEM, transfected with siRNA, and processed after three days as indicated above.

Production of recombinant full-length human PHAX

The human PHAX full-length sequence 1-384 was cloned into pETM11 vector containing a TEV cleavable six-histidine tag. Expression in *E. coli* Rosseta 2 cells was induced with 0.4 mM IPTG and performed for 6h at 25°C. Cells were resuspended in 300 mM NaCl, 50 mM HEPES pH 7.8, 10% (v/v) glycerol, 5 mM β-mercaptoethanol and lysed by sonication. Soluble protein was captured after centrifugation (1h, 10°, 35 000 g) by immobilized metal ion affinity chromatography (*Chelating Sepharose* Fast Flow, GE Healthcare) and eluted after several washes with 300 mM imidazol. After TEV cleavage and dialysis (100 mM NaCl, 20 mM HEPES 7.8, 10 mM β-mercaptoethanol) the protein was further purified by ion exchange chromatography and heparin column (HiPrep Heparin HP, GE Healthcare).

After concentrating (Amicon® Ultra Centrifugal filters, MERCK Millipore) the protein was analyzed by size exclusion chromatography (S200, GE Healthcare).

Yeast two-hybrid assay

Yeast two-hybrid and bridged two-hybrid assays were performed as previously described (Boulon et al., 2008). For bridged assays, pACT-II and p422 (ADE2 multicopy) plasmids were introduced into Y187 strains, while pAS2ΔΔ plasmids were transformed into CG1945 strain. Strains were crossed and diploids were plated on triple and quadruple selective media (−Leu −Trp −Ade or −Leu −Trp −Ade −His). Growth was assessed visually after 3 days at 30°C. A similar protocol was used for regular two-hybrid assays, except that p422 plasmids and adenine selection were omitted.

LUMIER assays

HEK293T cells were grown and transfected on 6-well plates. Two days after transfection, cells were extracted in 450 μl HNTG with anti-protease coktail and RNase A (60 μg/ml), at 4°C for 15 minutes. Cellular debris were removed by centrifugating at 20 000g for 5 minutes. Antibody coated beads were incubated with 180 μl of extracts for 2h at 4°C, and beads were washed three times in HNTG. Beads were resuspended in PBL buffer (Promega), and luciferase activity were measured in the IP and pellets using the dual-luciferase assay (Promega). HNTG is 20 mM HEPES, pH 7.9, 150 mM NaCl, 1% Triton, 10% glycerol, 1 mM MgCl2, 1 mM EGTA, and protease inhibitors (Roche).

MS2 Tethering assay

HEK293 cells were cotransfected with the luciferase reporter plasmid containing two MS2 stem-loops in its 3' UTR and with plasmids expressing MCP-GFP fused to the protein of interest. Two days later, cells were lysed in PBL buffer (Promega) and firefly and renilla luciferase activities were measured using the dual luciferase kit as recommended by the manufacturer (Promega).

Microscopy and LacO FRAP assay

U2OS cells carrying a LacO array (Marzec et al., 2015) were plated on coverslips and co-transfected using JetPrime (PolyPlus) with plasmids expressing the GFP fusion of interest together with the mRFP-Laci fusion of interest. Two days later, cells were either fixed and visualized by wide-field microscopy, or imaged live using a Zeiss Lsm780 microscope. FRAP was performed on a spot with a radius of 1.5 μm using 10 iterations at full laser power, and images were collected every 96 ms. The mean fluorescence intensities of a bleached and of a non-bleached area were calculated for each time point (I_{spot} and I_{cell}). The background signal was measured outside the cell (I_{bkg}). The bleaching and background corrected fluorescence intensity was then calculated at each time point $I = (I_{spot}-I_{bkg}) / (I_{cell}-I_{bkg})$. This value was then normalized to 1 by dividing it with the value of I computed with the averaged pre-bleach time points.

CBP20 was tethered to the LacO array while its partners were expressed as soluble GFP fusions, exept for ZC3H18. Indeed, ZC3H18 interacts with itself and by tethering CBP20 to LacO and expressing GFP-ZC3H18, complexes of the following type could form: LacO-CBP20-ZC3H18-ZC3H18-CBP20-LacO. In such complexes, ZC3H18 could dissociate from CBP20 and nevertheless could remain attached to the array by virtue of its interaction with the second molecule of ZC3H18. This difficulty is

avoided if ZC3H18 is tethered to LacO, because the soluble GFP-CBC complex can only make a single interaction with ZC3H18.

Acknowledgements

We thank I. Poser and A. Hyman for the gift of the BAC-LAP cell lines and E. Izaurralde for anti-CBP20 and -CBP80 antibodies. Work in the T.H.J. laboratory was supported by the ERC (grant 339953) and the Danish National Research Council as well as the Danish National Research- (grant DNRF58), the Lundbeck- and the Novo Nordisk-Foundations. S.G was partly supported by an Eiffel PhD fellowship. Work in the E.B. laboratory was supported by a grant from the Ligue Nationale Contre le Cancer. G.K. was supported by the Wellcome Trust grant 097383 and by the U.K. Medical Research Council. N.B was supported by a fellowship from the Algerian ministry of higher education and the Ligue Nationale Contre le Cancer.

References

Andersen, P., Domanski, M., Kristiansen, M., Storvall, H., Ntini, E., Verheggen, C., Schein, A., Bunkenborg, J., Poser, I., Hallais, M.*, et al.* (2013). The human cap-binding complex is functionally connected to the nuclear RNA exosome. Nat Struct Mol Biol. *20*, 1367-1376.

Andersson, R., Andersen, R., Valen, E., Core, L., Bornholdt, J., Boyd, M., Jensen, T., and Sandelin, A. (2014). Nuclear stability and transcriptional directionality separate functionally distinct RNA species. Nat Commun. *5*, 5336.

Audibert, A., Weil, D., and Dautry, F. (2002). In vivo kinetics of mRNA splicing and transport in mammalian cells. Mol Cell Biol. *22*, 6706-6718.

Baillat, D., Hakimi, M., Näär, A., Shilatifard, A., Cooch, N., and Shiekhattar, R.

(2005). Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. Cell. *123*, 265-276.

Bentley, D. (2014). Coupling mRNA processing with transcription in time and space. Nat Rev Genet. *15*, 163-175.

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

Björk, P., and Wieslander, L. (2014). Mechanisms of mRNA export. Semin Cell Dev Biol. *32*, 47-54638-54650.

Boireau, S., Maiuri, P., Basyuk, E., de la Mata, M., Knezevich, A., Pradet-Balade, B., Bäcker, V., Kornblihtt, A., Marcello, A., and Bertrand, E. (2007). The transcriptional cycle of HIV-1 in real-time and live cells. J Cell Biol. *179*, 291-304.

Boulon, S., Marmier-Gourrier, N., Pradet-Balade, B., Wurth, L., Verheggen, C., Jády, B., Rothé, B., Pescia, C., Robert, M., Kiss, T.*, et al.* (2008). The Hsp90 chaperone controls the biogenesis of L7Ae RNPs through conserved machinery. J Cell Biol. *180*, 579-595.

Boulon, S., Verheggen, C., Jady, B.E., Girard, C., Pescia, C., Paul, C., Ospina, J.K., Kiss, T., Matera, A.G., Bordonné, R.*, et al.* (2004). PHAX and CRM1 are required sequentially to transport U3 snoRNA to nucleoli. Mol. Cell *16*, 777-787.

Cabili, M., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915-1927.

Calero, G., Wilson, K., Ly, T., Rios-Steiner, J., Clardy, J., and Cerione, R. (2002). Structural basis of m7GpppG binding to the nuclear cap-binding protein complex. Nat Struct Biol. *9*, 912-917.

Chen, Y., Pai, A., Herudek, J., Lubas, M., Meola, N., Järvelin, A., Andersson, R.,

Pelechano, V., Steinmetz, L., Jensen, T.*, et al.* (2016). Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters. Nat Genet. *in press*.

Cheng, H., Dufu, K., Lee, C., Hsu, J., Dias, A., and Reed, R. (2006). Human mRNA export machinery recruited to the 5' end of mRNA. Cell *127*, 1389-1400.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.*, et al.* (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775-1789.

Domanski, M., Molloy, K., Jiang, H., Chait, B., Rout, M., Jensen, T., and LaCava, J. (2012). Improved methodology for the affinity isolation of human protein complexes expressed at near endogenous levels. Biotechniques. *0*, 1-6.

Flaherty, S., Fortes, P., Izaurralde, E., Mattaj, I.W., and Gilmartin, G. (1997). Participation of the nuclear cap binding complex in pre-mRNA 3' processing. Proc Natl Acad Sci U S A. *94*, 11893-11898.

Furuichi, Y., LaFiandra, A., and Shatkin, A. (1977). 5'-Terminal structure and mRNA stability. Nature *266*, 235-239.

Glover-Cutter, K., Kim, S., Espinosa, J., and Bentley, D. (2008). RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. Nat Struct Mol Biol. *15*, 71-78.

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

Görnemann, J., Kotovic, K., Hujer, K., and Neugebauer, K. (2005). Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex. Mol Cell. *19*, 53-63.

Gruber, J., Olejniczak, S., Yong, J., La Rocca, G., Dreyfuss, G., and Thompson, C. (2012). Ars2 promotes proper replication-dependent histone mRNA 3' end formation. Mol Cell. *45*, 87-98.

Grüter, P., Tabernero, C., von Kobbe, C., Schmitt, C., Saavedra, C., Bachi, A., Wilm, M., Felber, B., and Izaurralde, E. (1998). TAP, the human homolog of Mex67p, mediates CTE-dependent RNA export from the nucleus. Mol Cell. *1*, 649-659.

Hallais, M., Pontvianne, F., Andersen, P., Clerici, M., Lener, D., Benbahouche, N.H., Gostan, T., Vandermoere, F., Robert, M., Cusack, S.*, et al.* (2013). CBC-ARS2 stimulates 3'-end maturation of multiple RNA families and favors cap-proximal processing. Nat Struct Mol Biol. *20*, 1358-1366.

Hubner, N., and Mann, M. (2011). Extracting gene function from protein-protein interactions using Quantitative BAC InteraCtomics (QUBIC). Methods. *53*, 453-459.

Ideue, T., Sasaki, Y., Hagiwara, M., and Hirose, T. (2007). Introns play an essential role in splicing-dependent formation of the exon junction complex. Genes Dev. *21*, 1993-1998.

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

Izaurralde, E., Stepinski, J., Darzynkiewicz, E., and Mattaj, I. (1992). A cap binding protein that may mediate nuclear export of RNA polymerase II-transcribed RNAs. J Cell Biol. *118*, 1287-1295.

Jarmolowski, A., Boelens, W., Izaurralde, E., and Mattaj, I. (1994). Nuclear export of different classes of RNA is mediated by specific factors. *J Cell Biol.* , 627-635.

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

e02407.

Kishore, S., Gruber, A., Jedlinski, D., Syed, A., Jorjani, H., and Zavolan, M. (213). Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. *Genome Biol.* , R45.

Konig, J., K., Z., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D., Luscombe, N., and Ule, J. (2011). iCLIP--transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. J Vis Exp. *50*, 2638.

Laubinger, S., Sachsenberg, T., Zeller, G., Busch, W., Lohmann, J., Rätsch, G., and Weigel, D. (2008). Dual roles of the nuclear cap-binding complex and SERRATE in pre-mRNA splicing and microRNA processing in Arabidopsis thaliana. Proc Natl Acad Sci U S A. *105*, 8795-8800.

Le Hir, H., Izaurralde, E., Maquat, L., and Moore, M. (2000a). The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. EMBO J. *19*, 6860-6869.

Le Hir, H., Moore, M., and Maquat, L. (2000b). Pre-mRNA splicing alters mRNP composition: evidence for stable association of proteins at exon-exon junctions. Genes Dev. *14*, 1098-1108.

Lubas, M., Andersen, P., Schein, A., Dziembowski, A., Kudla, G., and Jensen, T. (2015). The human nuclear exosome targeting complex is loaded onto newly synthesized RNA to direct early ribonucleolysis. Cell Rep. *10*, 178-192.

Lubas, M., Christensen, M., Kristiansen, M., Domanski, M., Falkenby, L., Lykke-Andersen, S., Andersen, J., Dziembowski, A., and Jensen, T. (2011). Interaction profiling identifies the human nuclear exosome targeting complex. Mol Cell. *43*, 624-637.

Lutfalla, G., and Uze, G. (2006). Performing quantitative reverse-transcribed

polymerase chain reaction experiments. Methods Enzymol. *386*, 386-400.

Lykke-Andersen, S., Chen, Y., Ardal, B., Lilje, B., Waage, J., Sandelin, A., and Jensen, T. (2014). Human nonsense-mediated RNA decay initiates widely by endonucleolysis and targets snoRNA host genes. Genes Dev. *28*, 2498-2517.

Marzec, P., Armenise, C., Pérot, G., Roumelioti, F., Basyuk, E., Gagos, S., Chibon, F., and Déjardin, J. (2015). Nuclear-receptor-mediated telomere insertion leads to genome instability in ALT cancers. Cell. *160*, 913-927.

Masuyama, K., Taniguchi, I., Kataoka, N., and Ohno, M. (2004). RNA length defines RNA export pathway. Genes Dev. *18*, 2074-2085.

Matera, A.G., Terns, R.M., and Terns, M.P. (2007). Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. Nat Rev Mol Cell Biol *8*, 209-220.

Mazza, C., Ohno, M., Segref, A., Mattaj, I., and Cusack, S. (2001). Crystal structure of the human nuclear cap binding complex. Mol Cell. *8*, 383-396.

Mazza, C., Segref, A., Mattaj, I., and Cusack, S. (2002). Large-scale induced fit recognition of an m(7)GpppG cap analogue by the human nuclear cap-binding complex. EMBO J. *21*, 5548-5557.

McCloskey, A., Taniguchi, I., Shinmyozu, K., and Ohno, M. (2012). hnRNP C tetramer measures RNA length to classify RNA polymerase II transcripts for export. Science., 1643-1646.

Müller-McNicoll, M., and Neugebauer , K. (2014). Good cap/bad cap: how the cap-binding complex determines RNA fate. Nat Struct Mol Biol. *21*, 9-12.

Narita, T., Yung, T.M.C., Yamamoto, J., Tsuboi, Y., Tanabe, H., Tanaka, K., Yamaguchi, Y., and Handa, H. (2007). NELF interacts with CBC and participates in 3' end processing of replication-dependent histone mRNAs. Mol. Cell *26*, 349-365.

Ntini, E., Järvelin, A., Bornholdt, J., Chen, Y., Boyd, M., Jørgensen, M., Andersson,

R., Hoof, I., Schein, A., Andersen, P.*, et al.* (2013). Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nat Struct Mol Biol. *20*, 923-928.

Ohno, M., Sakamoto, H., and Shimura, Y. (1987). Preferential excision of the 5' proximal intron from mRNA precursors with two introns as mediated by the cap structure. Proc Natl Acad Sci U S A. *84*, 5187-5191.

Ohno, M., Segref, A., Bachi, A., Wilm, M., and Mattaj, I.W. (2000). PHAX, a mediator of U snRNA nuclear export whose activity is regulated by phosphorylation. Cell 187-198.

Ohno, M., Segref, A., Kuersten, S., and Mattaj, I.W. (2002). Identity Elements Used in Export of mRNAs. Molecular Cell *9*, 659-671.

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

Phair, R., and Misteli, T. (2000). High mobility of proteins in the mammalian cell nucleus. Nature. *404*, 604-609.

Poser, I., Sarov, M., Hutchins, J., Hériché, J., Toyoda, Y., Pozniakovsky, A., Weigl, D., Nitzsche, A., Hegemann, B., Bird, A.*, et al.* (2008). BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. Nat. Methods *5*, 409-415.

Preker, P., Almvig, K., Christensen, M., Valen, E., Mapendano, C., Sandelin, A., and Jensen, T. (2011). PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. Nucleic Acids Res. *39*, 7179-7193.

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.,

Mapendano, C., Schierup, M., and Jensen, T.H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. Science *322*, 1851-1854.

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

Schmidt, U., Basyuk, E., Robert, M., Yoshida, M., Villemin, J., Auboeuf, D., Aitken, S., and Bertrand, E. (2011). Real-time imaging of cotranscriptional splicing reveals a kinetic model that reduces noise: implications for alternative splicing regulation. J Cell Biol. *193*, 819-829.

Segref, A., Mattaj, I.W., and Ohno, M. (2001). The evolutionarily conserved region of the U snRNA export mediator PHAX is a novel RNA-binding domain that is essential for U snRNA export. . RNA *7*, 351-360.

Segref, A., Sharma, K., Doye, V., Hellwig, A., Huber, J., Lührmann, R., and Hurt, E. (1997). Mex67p, a novel factor for nuclear mRNA export, binds to both poly(A)+ RNA and nuclear pores. EMBO J. *16*, 3256-3271.

Shi, Y., Di Giammartino, D., Taylor, D., Sarkeshik, A., Rice, W., Yates, J.r., Frank, J., and Manley, J. (2009). Molecular architecture of the human pre-mRNA 3' processing complex. Mol Cell. *33*, 365-376.

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

Wiśniewski, J., Zougman, A., and Mann, M. (2009). Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. J Proteome Res. *8*, 5674-5678.

Worch, R., Niedzwiecka, A., Stepinski, J., Mazza, C., Jankowska-Anyszka, M., Darzynkiewicz, E., Cusack, S., and Stolarski, R. Specificity of recognition of mRNA 5' cap by human nuclear cap-binding complex. RNA *11*, 1355-1363.

Table I. Two-hybrid interactions between CBC-related proteins

	ARS2[1]	PHAX	ZC3H18	RBM7	ZCCHC8	NELF-E	Alix[2]
ARS2[1]	nd	-	**	-	-	-	-
PHAX	-	nd	-	-	-	-	-
RBM7	-	-	-	nd	**	-	-
ZCCHC8	na	na	na	na	na	na	na
CBC[3]	**	**	*	-	-	**	-
ZC3H18	**	-	nd	-	-	-	-
ZC3H18 (746-953)[4]	-	nd	-	nd	**	nd	-
ARS2 screen[5]	nf	nf	nf	nf	nf	nf	nf
dARS2 screen[5]	nf	nf	7 (380-639)	nf	nf	nf	nf
atARS2 screen[5]	nf	1 (18-154)	nf	nf	nf	nf	nf
ZC3H18 screen[5]	nf	nf	nf	nf	1 (265-694)	nf	nf

[1]Colomns and rows indicate the relevant cDNA fused to the GAL4 activation- and DNA-binding domains, respectively. Interaction strength is indicated by the number of stars. nd: not done; na: not applicable due to constitutive activation of the reporter; nf: not found in the library screen. -: no interaction.
[2]Human protein used as a negative control.
[3]CBC subunits were co-expressed in yeast cells. CBP80 was fused to the GAL4 DNA-binding domain and CBP20 was non-tagged.
[4]In this row, ZC3H18$^{746-953}$ was tested against the indicated preys except for ZCCHC8, for which the fragment used was ZCCHC8$^{265-694}$. This is the ZCCHC8 fragment found in the ZC3H18 Y2H screen.
[5]Screens were performed with the full-length proteins fused to the DNA-binding domain of LexA. Numbers before parentheses denote how many independent clones were found. Numbers within parentheses denote the smallest interacting fragment of the prey (in amino acids).

Figure Legends

Figure 1. Cap-proximal mRNA binding by ARS2, PHAX and ZC3H18

A. Schematic overview of the different protein complexes relevant for this study. CBCAP is shown in yellow, NEXT in purple, and CBCN is circled in green. See text for details.

B. Fractions of iCLIP reads, from replicate libraries, mapping to the indicated classes of capped or uncapped RNA expressed as proportions of total library reads. Reads marked as "others" could not be unambiguously assigned to any of the above categories. For comparison, we show cytoplasmic poly(A)$^+$-selected RNA-Seq data from HEK293 cells (SRR1275413) and HeLa cells (SRR3479116), as well as RNA-Seq data from rRNA-depleted total RNA from HEK293 cells (SRR2096982) and HeLa cells (SRR1014903).

C, D. Genome browser views of representative protein-coding genes *PPIA* (B) and *RPS16* (C), showing iCLIP reads from replicate CBP20, ARS2, PHAX, ZC3H18 and RBM7 samples. Reads mapping to the *PPIA* and *RPS16* RNAs are shown as mapped Reads Per Million (RPM) library reads (see scale bar to the right of the image). Purple color implies that displayed reads exceed the used scale.

E. Fractions of iCLIP or RNA-Seq reads mapping within cap-proximal regions of 100, 500 or 1000 nts of 5,769 well-annotated pre-mRNA genes. The iCLIP results represent averages of replicate experiments.

F. Fractions of exon-intron (EI) and intron-exon (IE) junction reads, averaged between replicate experiments, mapping over Refseq pre-mRNAs. Fractions were calculated as EI/(EI+IE+EE) and IE/(EI+IE+EE) as indicated. EE: exon-exon junction reads. Note that EI- are higher than IE-fractions for CBP20 libraries in agreement

with the cap-binding nature of this protein. Conversely, IE- are higher than EI-fractions for RBM7 libraries as previously reported (Lubas et al., 2015).

Figure 2. ARS2, PHAX and ZC3H18 are targeted to common RNA families

A. Density profiles of reads from the indicated iCLIP libraries displayed as Reads Per Million (RPM) library reads, around -/+2kb regions of transcription start sites (TSSs, left part) and Transcript Termination Sites (TTSs, right part) of the protein-coding genes from Fig. 1D. Transcription directions are indicated by arrows as forward ('mRNA direction') and reverse ('PROMPT direction'). Red and blue reads map to forward and reverse strands, respectively. Signal corresponding to 1 RPM is indicated. Note that CBP20 and ZC3H18 mRNA profiles were disrupted to ease visual inspection.

B. Density profiles as in (A) but only showing reverse read densities in -/+2kb regions anchored around PROMPT TSSs as defined by CAGE summits (Chen et al., 2016). Signal corresponding to 1 RPM is indicated.

C. Density profiles as in (A) but showing forward and reverse read densities in -/+2kb regions anchored around eRNA TSSs as defined by CAGE summits (Chen et al., 2016). Signal corresponding to 0.05 RPM is indicated.

D. Proportion of reads from the indicated replicate libraries mapping to mature (white columns) and 3'extended regions (light green columns) of RDH RNAs. '3' extensions' denote 1-500 nt downstream of the annotated mature RDH RNA 3'end. Note disruption of the y-axis to ease visual inspection of all data.

E. Proportion of reads mapping to mature (white columns), short- (light green) and long- (dark green) 3'extended regions of independently transcribed sn(o)RNAs. 'short 3'extensions' and 'long 3'extensions' denote 1-20 nts and 50-500 nts, respectively,

downstream of the annotated mature sn(o)RNA 3'ends. Inset shows the ratio of reads mapping to long 3'extensions relative to mature RNA.

F. Proportion of reads mapping to 5'extended- (blue columns), mature- (white columns) and 3'extension- (light green columns) regions of uncapped snoRNAs located in introns. 5'- and 3'-extension denote regions from the mature snoRNA 5'- and 3'-ends to the respective intronic 5'- and 3'-ends, respectively.

Figure 3. ARS2, PHAX and ZC3H18 are targeted to common transcripts

A. Scatter plots showing RPKM values of iCLIP-tags from one indicated library versus another. Each RNA species is a dot. Grey: pre-mRNAs; violet: histone mRNAs; light blue: lncRNAs; red: sn(o)RNAs.

B. Scatter plot showing the log2 fold changes in PHAX vs. ZC3H18 binding, as a function of normalized read counts for all RNAs identified in the iCLIP experiments. RNAs binding similarly to PHAX and ZC3H18 (grey dots) or significantly more to one protein (red dots) were determined by the DE-Seq package.

C. Venn diagram displaying mRNAs bound by PHAX (yellow) and/or ZC3H18 (green), as determined by DE-Seq analysis of the iCLIP data.

D. Cumulative distribution of iCLIP reads from the indicated replicate libraries ranked as a function of RNA size (x-axis). Left: all capped RNAs; middle: all capped RNAs except snRNAs; right: all capped RNAs except snRNAs and histone mRNAs.

E. Bar plots displaying fractions of mRNA affected by ZC3H18 depletion (red) in the entire mRNA population (left) or in the mRNAs preferentially bound by PHAX or ZC3H18 (middle and left, respectively). The mean change in expression levels upon depletion of ZC3H18 is shown in blue, for the same RNA population. PHAX- and

ZC3H18-bound mRNAs are shown in Fig. 3C. The differences between the three populations are not statistically significant.

Figure 4. Molecular organization of CBC-related complexes

A. Schematic overview of Y2H data acquired from pair-wise tests and cDNA library screens (see Table I). The interaction of hMTR4 and the core exosome with RBM7/ZCCHC8 is indicated. 'Previously demonstrated direct physical interaction' is from (Andersen et al., 2013; Hallais et al., 2013; Lubas et al., 2011; Ohno et al., 2000).

B. Left panel: Western blotting analysis of RBM7-LAP co-IP experiments conducted from extracts of HeLa cells depleted of factors using siRNAs as indicated. 'CTRL' denotes control siRNA targeting firefly luciferase mRNA. Input samples used for IP are shown to the left (lanes 1-4) and eluate samples from the IP are shown to the right (lanes 5-8). Schematics to the right depict the interpretation of the conducted co-IPs.

C. Volcano plot displaying the result of triplicate PHAX-3XFLAG AC/MS experiments. The log2 fold change of peptide MS intensities between bait and reference ('bait-less' cell line) eluate samples (x-axis) were plotted against the negative log10 p-values (y-axis) calculated across the triplicate data (t-test). A dashed red curve separates specific PHAX-interacting proteins (upper right part of plot) from enriched proteins from the reference cell line (upper left part of plot). Some PHAX-interacting protein groups are color-coded as indicated in legend, and protein names relevant for this study are denoted. The full data set of specific co-precipitants is given in Table S3.

D. Column chart displaying abundance of selected proteins from PHAX-3XFLAG AC eluates. Peptide intensities divided by protein MW were normalized to results for the

bait protein. In this analysis, reference values were not subtracted from bait values as the reference procedure yielded more background material binding to unshielded antibody epitopes sometimes obscuring analysis (data not shown). Note disruptions of the y-axis to reveal intensities of all plotted factors.

Figure 5. PHAX and ZC3H18 make mutually exclusive interactions with the CBC *in vitro* and *in vivo*

A. Left panel: Western blotting analysis of RBM7-LAP co-IP experiments challenged with increasing amounts of exogenously added PHAX (lanes 7-12) or BSA (40 μg) (lanes 5 and 6) as a negative control. CTRL denotes that no exogenous protein was added. PHAX or BSA was added to bead-bound RBM7-LAP complexes. IN: Input; FT: flow-through; S: bead supernatant upon addition of the indicated protein; E: SDS-eluate of the materials left on the beads following addition of the indicated protein. Antibodies used for the analysis are shown to the left. Right panel: Schematic interpretation of the experimental result.

B. LUMIER assay showing interaction of 3xFLAG-FFL-CBP20 with RL-PHAX and RL-ZC3H18. Left panel: Schematic representation of the assay. Right Panel: Graph depicting efficiency of RL-PHAX and RL-ZC3H18 interactions with 3xFLAG-FFL-CBP20. Values are enrichment fold of RL-ZC3H18/RL-PHAX in the FLAG IP over a control IP performed with empty beads. Extracts were prepared from HEK293T cells transiently transfected with the corresponding plasmids.

C. LUMIER assay testing effect of overexpression of MYC-tagged competitor proteins on RL-ZC3H18 binding to 3x-FLAG-FFL-CBP20. Left: schematic of the assay. Right: graph depicting efficiency of RL-ZC3H18 interaction with 3xFLAG-FFL-

CBP20. Values are enrichment fold of RL-ZC3H18 (IP/Input), normalized by the 3xFLAG-FFL-CBP20 values (IP/Input).

D. LUMIER assay as in (C) but testing the effect of overexpression of MYC-tagged competitor proteins on RL-PHAX binding to 3x-FLAG-FFL-CBP20.

Figure 6. PHAX and ZC3H18 exhibit antagonistic effects on RNA levels

A. Schematic representation of the employed tethering assay. An RL reporter RNA containing two MS2 binding sites in its 3'UTRs was contained on a plasmid also harboring an FL reporter to control for transfection efficiencies. This plasmid was co-transfected with a plasmid expressing candidate polypeptides fused to MS2-GFP (MCP-GFP-X) or with a plasmid expressing MS2-GFP alone.

B. Effects on RL reporter activity of tethering MCP-GFP-X fusions. Left panel: RL/FFL activity ratios obtained with the MCP-GFP-X fusion, and normalized to the same ratio derived from the corresponding MCP-GFP control sample. Right panel: RL/FFL RNA ratios measured by RT-qPCR and expressed as Log_2 fold ratios between the MCP-GFP-X protein and the control MCP-GFP fusion. Bars represent standard deviations from >5 experiments.

C. Effects of PHAX and ZC3H18 single- and double-depletions on levels of snRNA species carrying long 3'extension. Levels of the indicated transcripts were measured by RT-qPCR on RNA extracted from HeLa cells treated with the indicated siRNAs (color-coded as displayed on the right). Values are displayed as Log_2 fold changes relative to samples treated with a control FFL siRNA. Bars represent standard deviations from >3 independent transfection experiments. Stars indicate significatively different values ($p<0.02$ with a t-test).

Figure 7. PHAX, ARS2 and ZC3H18 exchange rapidly on the CBC *in vivo*

A-C. Two left panels: confocal images of U2OS cells carrying a LacO array and co-transfected with plasmids expressing the indicated proteins (fields of view are 30x30 microns; left: GFP; right: mRFP). Middle rectangular panels: confocal images of a FRAP experiment of the GFP-tagged protein (fields of view are 10x48 microns). Right panels: fluorescent recovery curves of the indicated proteins. The FRAP experiments in the green and red channels were performed independently. Dark green: The indicated GFP-tagged protein in the nucleoplasm; light green: The indicated tagged protein in the LacO spot; red: the mRFP-Laci-CBP20 fusion in the LacO spot. Y-axes denote fluorescence intensities corrected for photobleaching and normalized to pre-bleach intensities. X-axes denote time in seconds. Grey bars represent standard deviations calculated from >10 different cells.

D. As in A-C, except that ZC3H18 was fused to Laci and tethered to the LacO spot in place of CBP20.

Figure 1

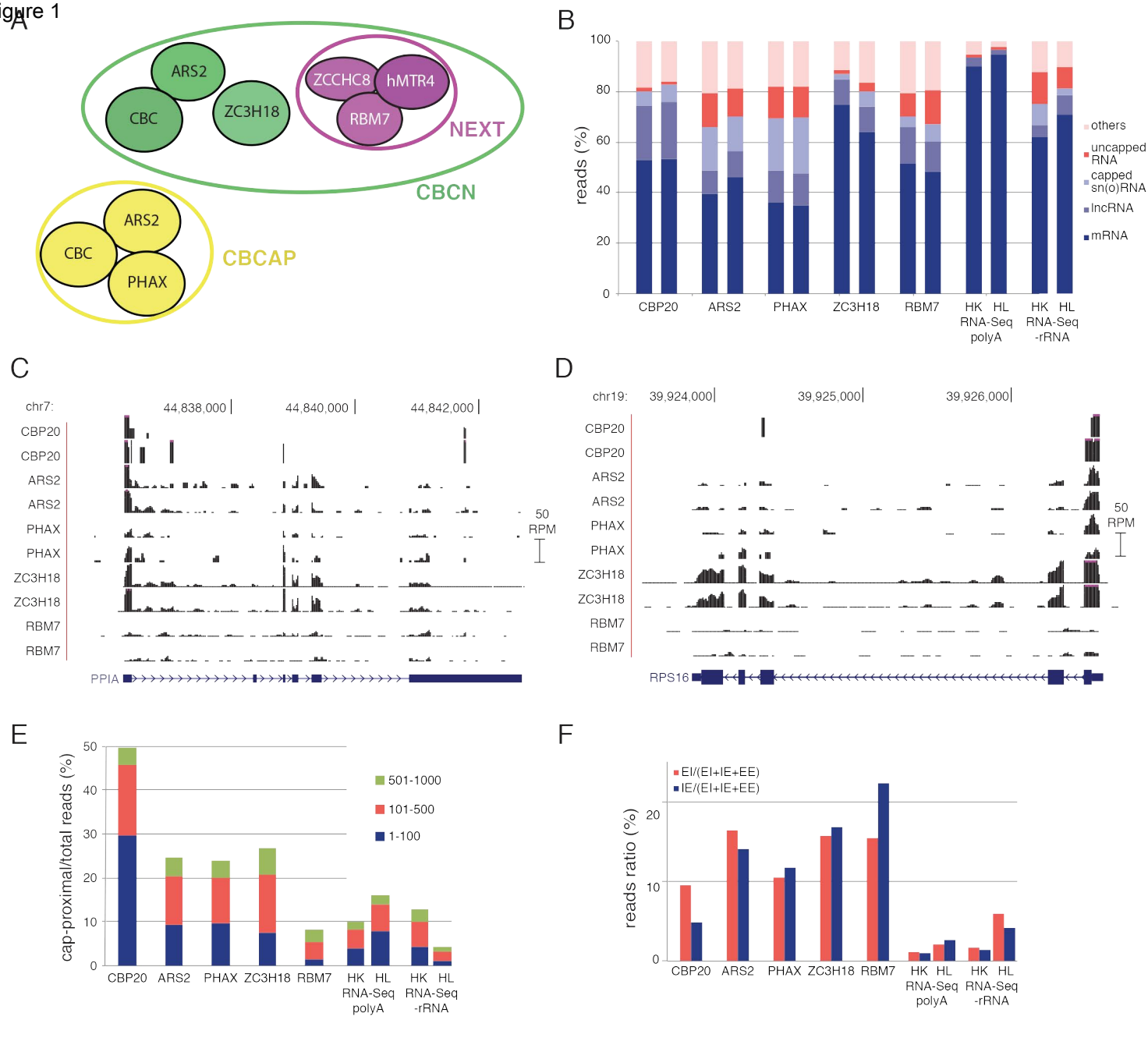

A
ARS2
CBC
ZC3H18
ZCCHC8
hMTR4
RBM7
NEXT
CBCN
ARS2
CBC
PHAX
CBCAP
B
reads (%)
100
80
60
40
20
0
CBP20
ARS2
PHAX
ZC3H18
RBM7
HK RNA-Seq polyA
HL RNA-Seq polyA
HK RNA-Seq -rRNA
HL RNA-Seq -rRNA
others
uncapped RNA
capped sn(o)RNA
lncRNA
mRNA
C
chr7:
44,838,000
44,840,000
44,842,000
CBP20
CBP20
ARS2
ARS2
PHAX
PHAX
ZC3H18
ZC3H18
RBM7
RBM7
50 RPM
PPIA
D
chr19:
39,924,000
39,925,000
39,926,000
CBP20
CBP20
ARS2
ARS2
PHAX
PHAX
ZC3H18
ZC3H18
RBM7
RBM7
50 RPM
RPS16
E
cap-proximal/total reads (%)
50
40
30
20
10
0
501-1000
101-500
1-100
CBP20
ARS2
PHAX
ZC3H18
RBM7
HK RNA-Seq polyA
HL RNA-Seq polyA
HK RNA-Seq -rRNA
HL RNA-Seq -rRNA
F
reads ratio (%)
EI/(EI+IE+EE)
IE/(EI+IE+EE)
20
10
0
CBP20
ARS2
PHAX
ZC3H18
RBM7
HK RNA-Seq polyA
HL RNA-Seq polyA
HK RNA-Seq -rRNA
HL RNA-Seq -rRNA

Figure 2

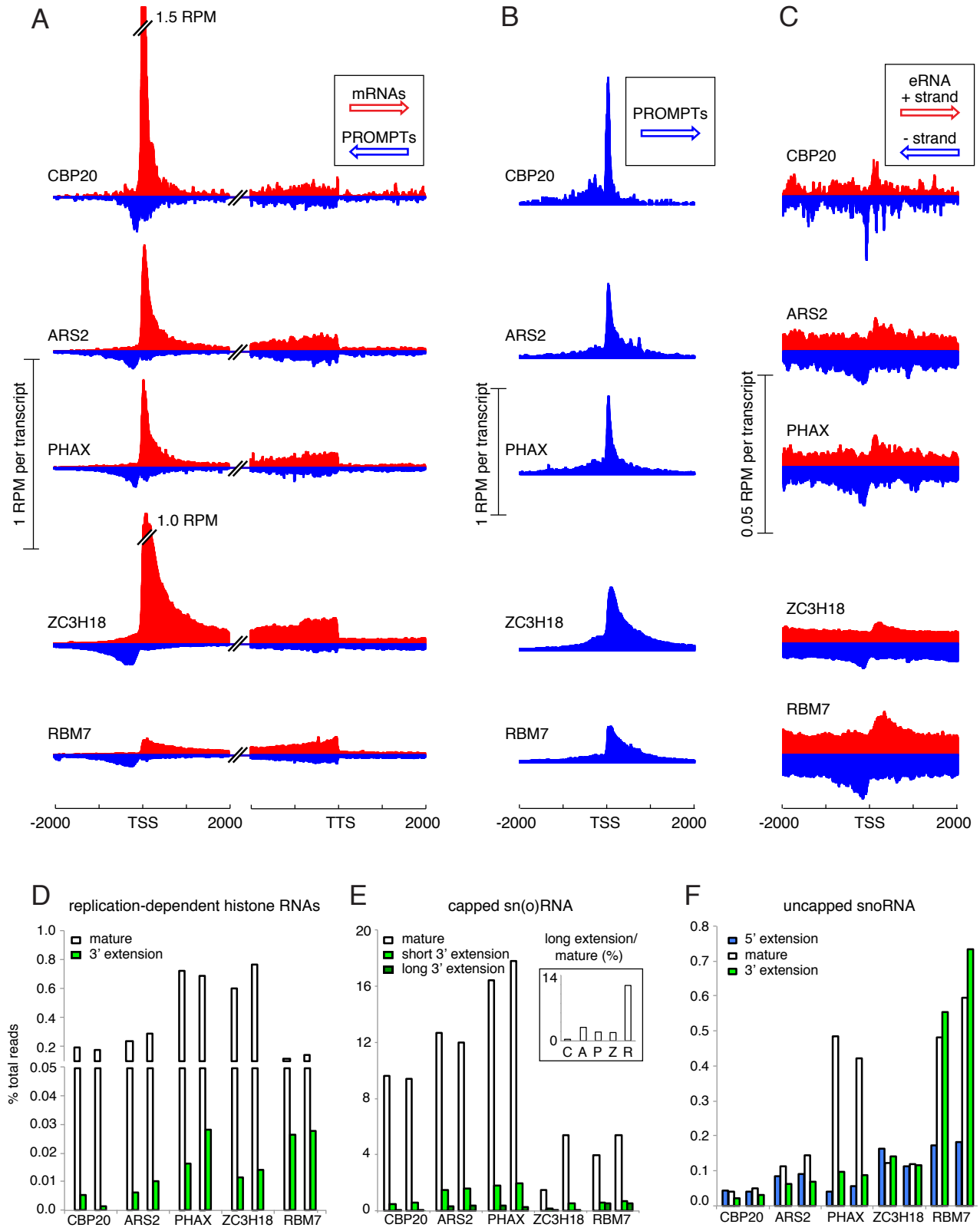
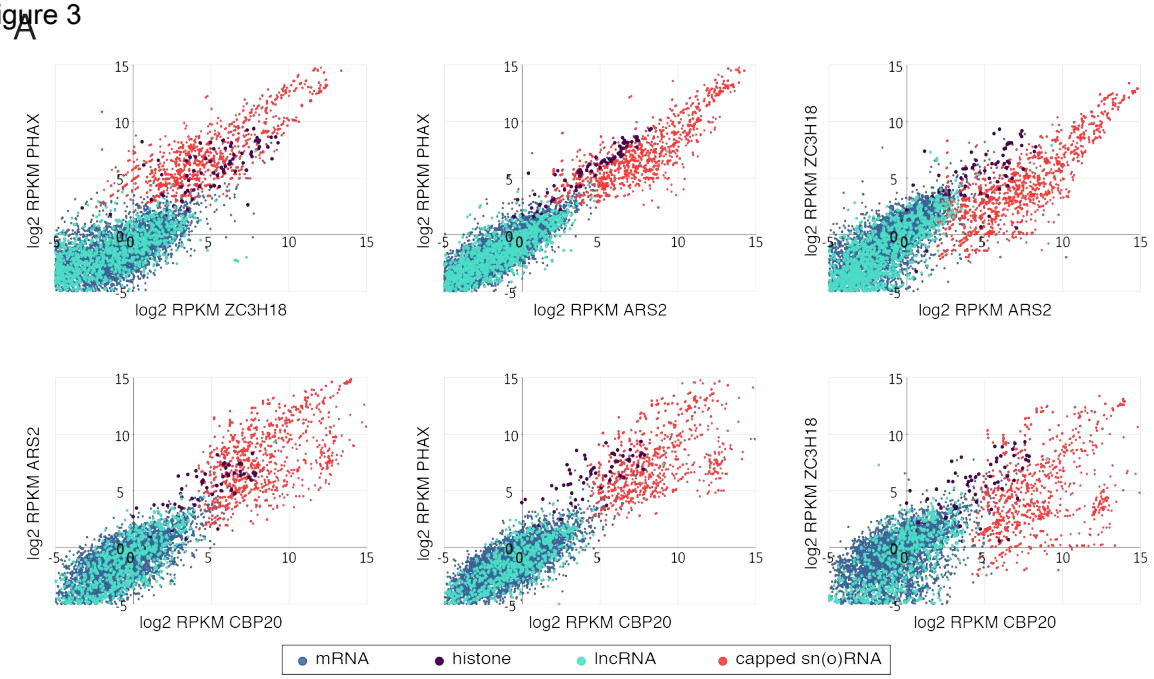
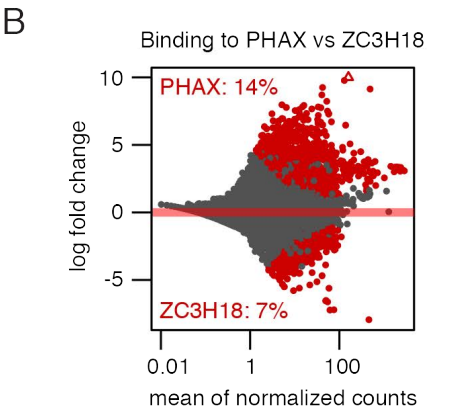
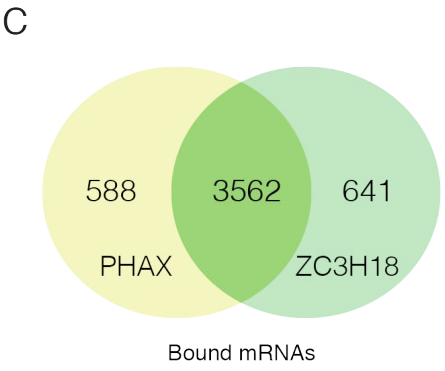


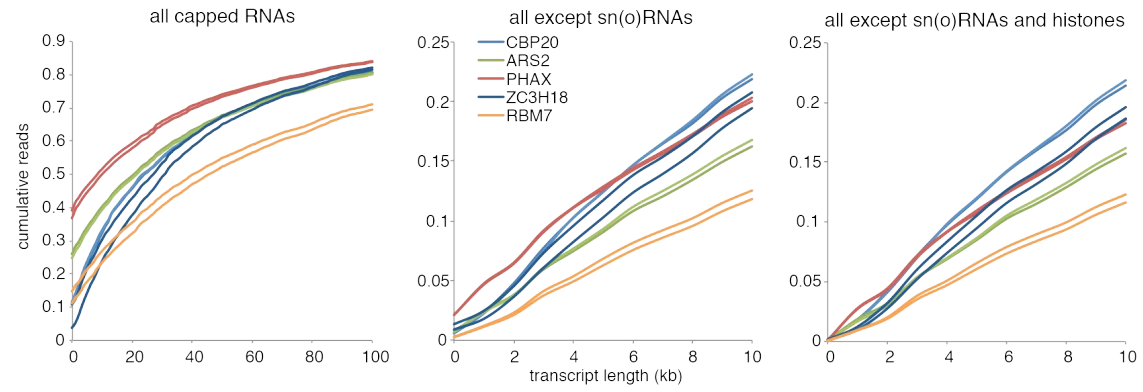
Figure 2

Figure 3









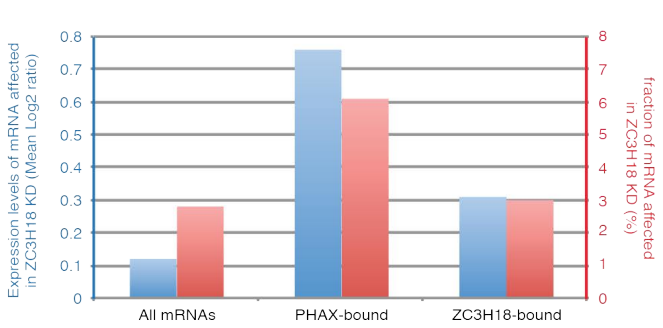


Figure 3

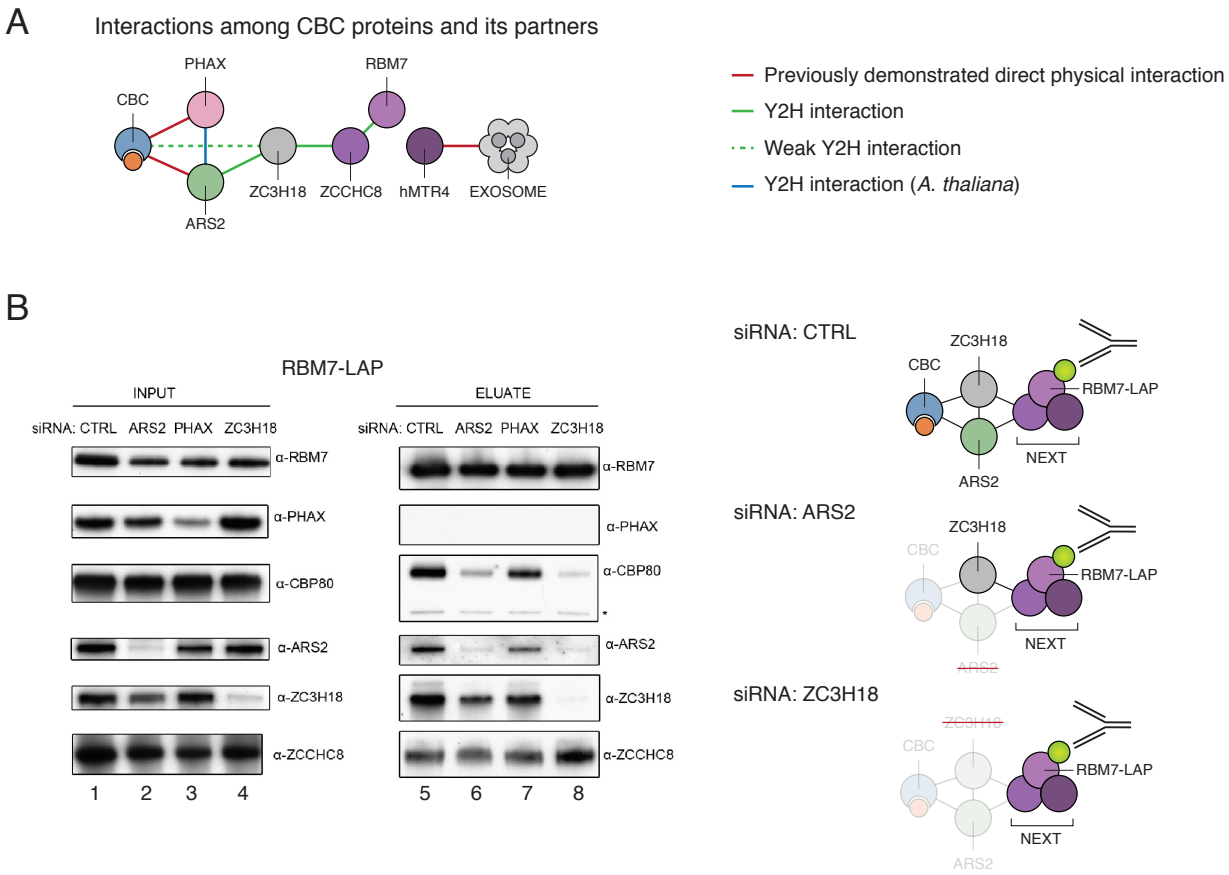
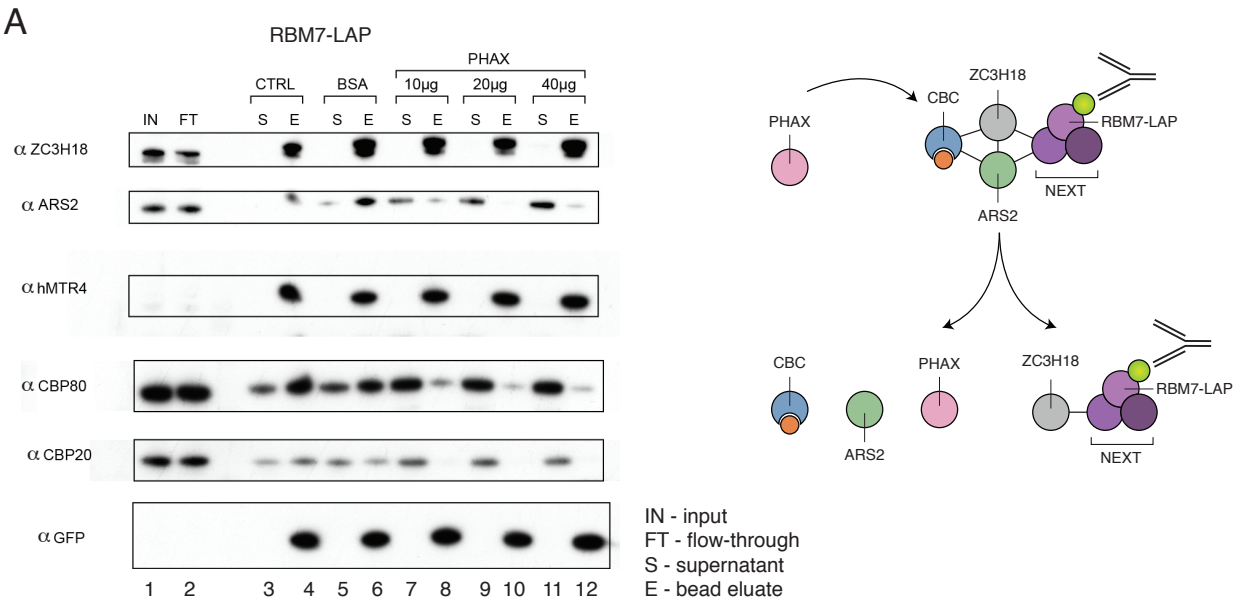
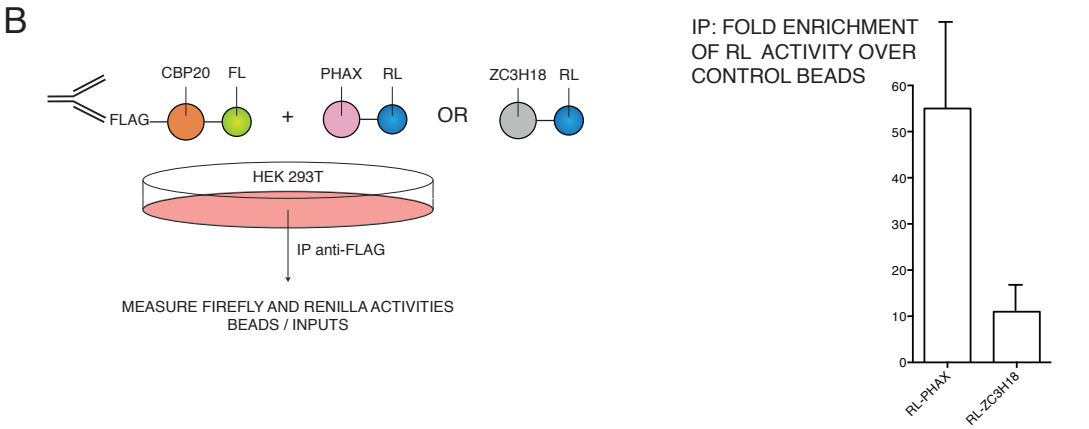
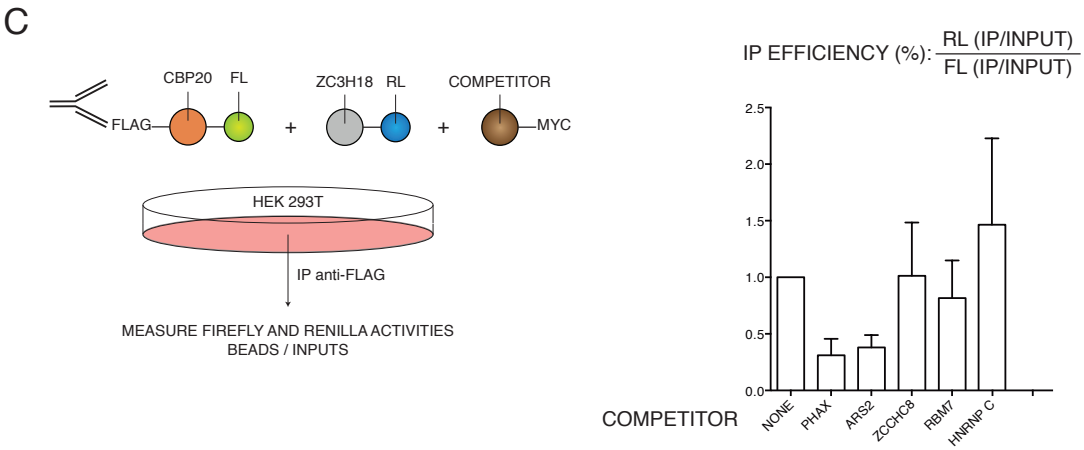

A
Interactions among CBC proteins and its partners
PHAX
RBM7
CBC
ZC3H18
ZCCHC8
hMTR4
EXOSOME
ARS2
Previously demonstrated direct physical interaction
Y2H interaction
Weak Y2H interaction
Y2H interaction (A. thaliana)
B
RBM7-LAP
INPUT
ELUATE
siRNA: CTRL ARS2 PHAX ZC3H18
siRNA: CTRL ARS2 PHAX ZC3H18
α-RBM7
α-PHAX
α-CBP80
α-ARS2
α-ZC3H18
α-ZCCHC8
1 2 3 4
5 6 7 8
siRNA: CTRL
siRNA: ARS2
siRNA: ZC3H18
NEXT
C
PHAX-3xFLAG
P−value (T-Test)
log2 (bait/reference)
CBP80
ARS2
CBP20
ZC3H18
hMTR4
ZCCHC8
RBM7
PHAX
CBCA
PHAX
NEXT
Exosome
Import and Export factors
snoRNPs
RBPs, Helicases
D
PHAX-3XFLAG
Intensity / MW (normalized to bait)
PHAX
CBCA
EXOSOME
PHAX CBP80 ARS2 CBP20 hRRP6 hRRP43 hMTR3 MPP6 hRRP41 hRRP4 hRRP40 hMTR4 hRRP42 hCSL4 hRRP46 hRRP45

Figure 4

Figure 5







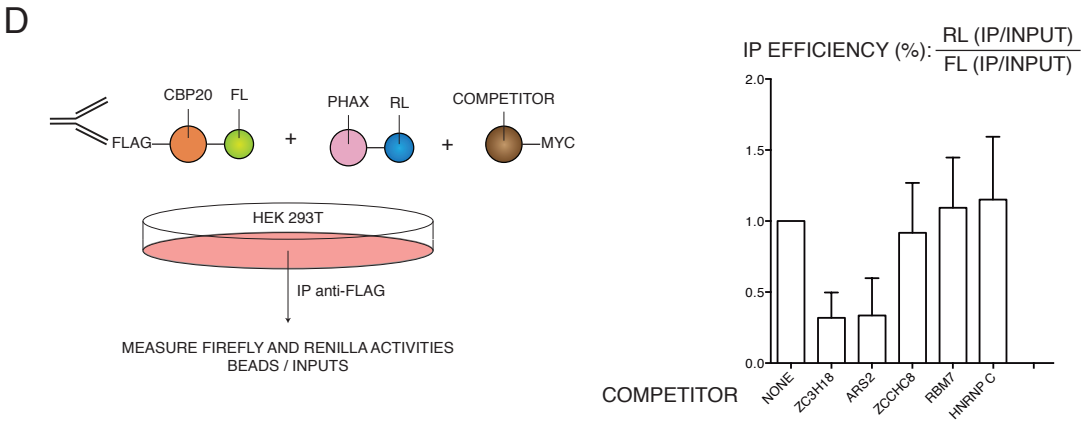


Figure 5

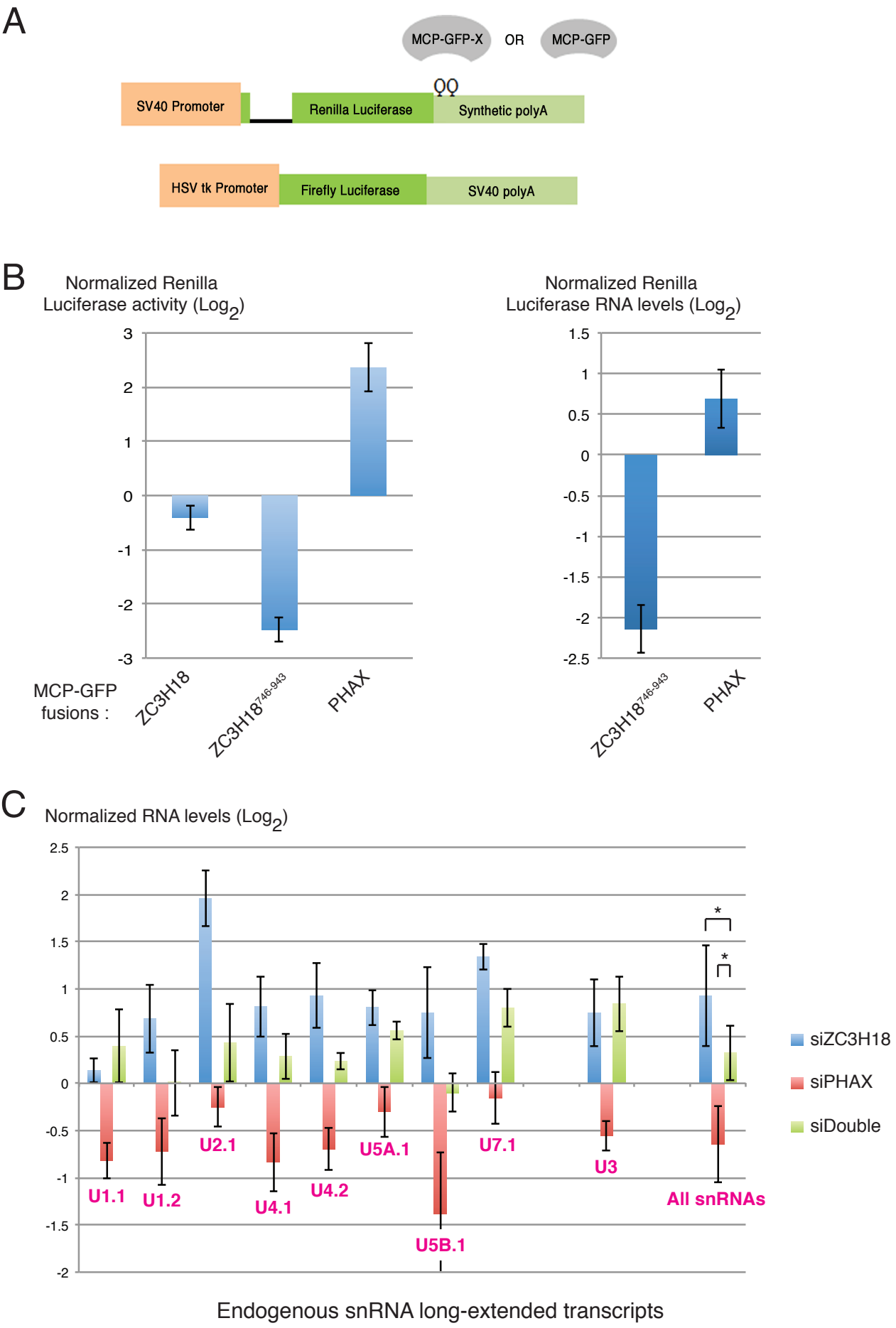

A
MCP-GFP-X
OR
MCP-GFP
SV40 Promoter
Renilla Luciferase
Synthetic polyA
HSV tk Promoter
Firefly Luciferase
SV40 polyA
B
Normalized Renilla Luciferase activity (Log2)
Normalized Renilla Luciferase RNA levels (Log2)
MCP-GFP fusions :
ZC3H18
ZC3H18746-943
PHAX
C
Normalized RNA levels (Log2)
U1.1
U1.2
U2.1
U4.1
U4.2
U5A.1
U5B.1
U7.1
U3
All snRNAs
siZC3H18
siPHAX
siDouble
Endogenous snRNA long-extended transcripts

Figure 6

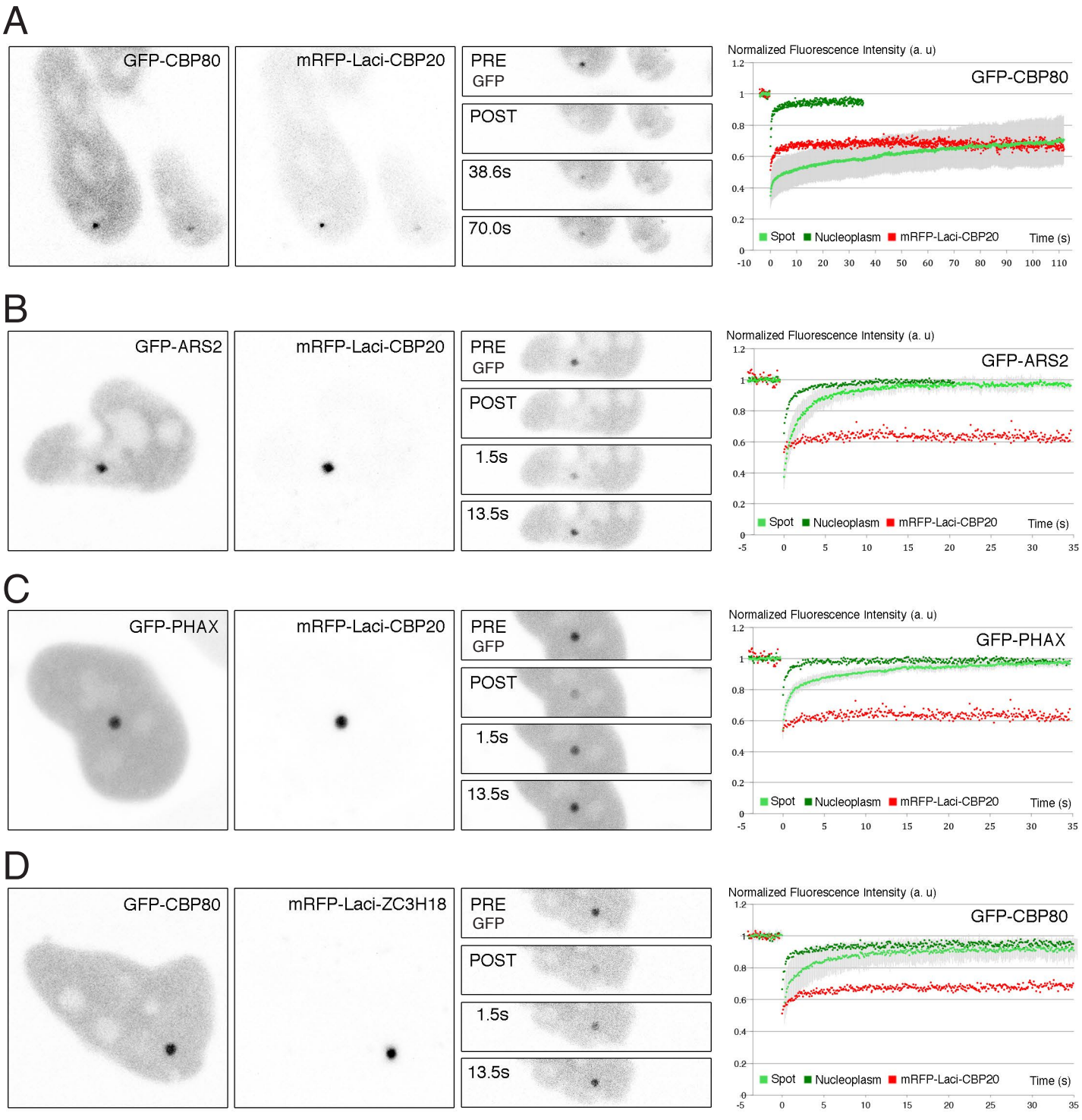

A
GFP-CBP80
mRFP-Laci-CBP20
PRE
GFP
POST
38.6s
70.0s
Normalized Fluorescence Intensity (a. u)
GFP-CBP80
Spot
Nucleoplasm
mRFP-Laci-CBP20
Time (s)
B
GFP-ARS2
mRFP-Laci-CBP20
PRE
GFP
POST
1.5s
13.5s
Normalized Fluorescence Intensity (a. u)
GFP-ARS2
Spot
Nucleoplasm
mRFP-Laci-CBP20
Time (s)
C
GFP-PHAX
mRFP-Laci-CBP20
PRE
GFP
POST
1.5s
13.5s
Normalized Fluorescence Intensity (a. u)
GFP-PHAX
Spot
Nucleoplasm
mRFP-Laci-CBP20
Time (s)
D
GFP-CBP80
mRFP-Laci-ZC3H18
PRE
GFP
POST
1.5s
13.5s
Normalized Fluorescence Intensity (a. u)
GFP-CBP80
Spot
Nucleoplasm
mRFP-Laci-CBP20
Time (s)

Figure 7