universität
wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation/Title of the Doctoral Thesis

„Exploring Intra-splicing and its regulatory potential“

verfasst von / submitted by

Maximilian Radtke

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2017 / Vienna 2017

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 794 685 490
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Molekulare Biologie / Molecular Biology
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Renée Schroeder

TABLE OF CONTENT

ABSTRACT

The diversity of the human proteome is the consequence of a cascade of regulatory network interactions that use the rather limited pool of direct genetically encoded information and push it through extensive transformative processes, allowing for maximum, or rather required, diversity. One of these processes, and arguably the one with the biggest diversification potential is transcription and its coupled RNA processing steps. Many regulatory layers in these processes, touching all aspects of transcription and RNA processing, have already been uncovered, drawing a complex interactive network. This includes the continuous discovery of molecular interactions and interactors and combinatorial events that allow for efficient shaping of the transcriptome in response to external stimuli, internal requirements or cellular diversification attempts. In this thesis, I want to present yet another layer of regulation via the extensively studied splicing process. Intra-splicing, previously hypothesized and later experimentally validated as a mechanism of long intron splicing, recently was accredited with a novel impact on gene expression: recursive exons. These allow for a flexible inclusion of additional genetic information in long introns and presumably facilitate long intron removal. Extending on this approach, I identified intrasplicing events on a genome-wide scale which allows for a more processing-focused approach to gene expression and additionally lead to the discovery of a number of regulatory splicing events that, in the cases studied in detail, lead to gene expression and isoform regulation. Down-regulation is achieved via partial intron retention, and increased gene expression via more efficient intron removal. The impact on isoform selection is hypothesized based on a novel intersection between recursive splicing and exon selection. This thesis provides the theoretical basis for the work performed in this doctorate and the resulting manuscripts attached.

ZUSAMMENFASSUNG

Die Diversität des menschlichen Proteoms ist das Ergebnis einer Kaskade regulativer Interaktionen, die genetisch kodierte Informationen nutzen und diesen durch einen transformativen Prozess führen. Dies erlaubt eine Maximierung der Diversität. Einer dieser Prozesse, und möglicherweise der mit dem größten Diversifikationspotential ist Transkription und die daran gekoppelten RNA Prozessierung. Viele regulative Ebenen dieser Prozesse, von Transkription bis Prozessierung, wurden bereits entdeckt und untersucht. Die führt zu einem bereits heute komplexen Gesamtbild, dieser zellulären Vorgänge. Diese Entdeckungen beinhalten molekulare Interaktionen und Interaktoren, kombinatorische Events, die das Transkriptom effektiv, als Antwort auf externe Stimuli, interne Anforderungen oder zelluläre Diversifikation formen. In der vorliegenden Arbeit präsentiere ich eine weitere regulative Ebene, welche durch den bereits intensiv erforschten Splicingmechanismus agiert. Intra-splicing, ursprünglich ein hypothetischer Mechanismus des langen-Intron-splicings wurde inzwischen experimentell verifiziert. Diesem wurde in jüngeren Studien eine weitere regulative Funktion zugeschrieben: recursive Exons. Diese erlauben eine flexible Einbindung zusätzlicher genetischer Information in langen Introns und ermöglichen außerdem, höchstwahrscheinlich, das Spleißen dieser langen Introns. Aufbauend auf diesem Ansatz habe ich weitere Intra-splicing Events genomweit identifiziert. Dies führte einerseits zu einer neuen Perspektive der RNA-Prozessierung und andererseits zu der Identifikation einiger Intra-splicing spezifischer regulativer Mechanismen. Die im Detail studierten Fälle beeinflussten Genexpression quantitativ und qualitativ. Teilweise Intronretention, beziehungsweise erhöhte RNA-Prozessierungseffizienz regulierten die Abundanz spezifischer Transkripte, während potentielle rekursive splicing Events in einer Art positioniert sind, die alternatives Splicing ermöglichen und, möglicherweise, forcieren.

Die vorliegende Arbeit bildet das theoretische Grundgerüst der während des Doktorats durchgeführten praktischen Arbeiten und den daraus resultierten Publikationen.

PREAMBLE

In a system as complex as a living cell that acts in an adaptive and responsive manner, regulation is key to managing restrained resources and ensure the systems viability. Complex regulatory interactions also allow processing of the available resources according to acute requirements. An instance of such a regulatory network is the transcriptome, one of the most diverse, multi-layered but also flexible domains of a eukaryotic cell. This diversity allows, with the proper regulatory tools, for an increased complexity with a limited gene set. This thesis will touch on many regulatory networks that intersect with the generation of transcriptome responsiveness and diversity. The main focus is RNA processing, especially splicing, as this allows for both, a quantitative and a qualitative regulation of transcriptional output.

When the human genome project set out to uncover the human genomic sequence and its embedded genes, common estimates based on the observed complexity of the proteome led to the assumption of orchestrated activity of hundreds of thousands of protein coding genes. However, merely 22.000 are in fact discovered to date[1] contrasting strongly with the observed diversity of the proteome. The bridge over this gap between genome and proteome must be found in the transcriptome, as shown by Nilsen[2], Pan[3] and Wang[4] shortly after deciphering the human genome. 95% of genes were shown to produce more than one isoform, by means of alternative promoter selection, alternative cleavage and polyadenylation or diverse shuffling of intra- and intergenic exons, i.e. alternative splicing. These processes have been shown to be tightly regulated and shape, in concert with gene expression induction and posttranslational modification, the transcriptome, the proteome and ultimately, the cell type specific regulatory network.

Novel findings now reveal that there are multiple regulatory domains influenced by splicing. This arises from the modification of the RNA itself as well as RNA interactors that are deposited in a splicing-dependant manner. This thesis aims to clarify the interactions of regulatory networks that frame splicing as well as present three studies that uncovered novel regulatory mechanisms, closely coupled to RNA transcription and splicing.

INTRODUCTION

TRANSCRIPTION

On a pure sequence basis, the eukaryotic genome does not allow for a strong regulation of transcription initiation. Sequence requirements are mainly restrained to a TATA-box and an initiator region close to the transcriptional start site (TSS)[5]. Additionally, certain gene-specific sequence elements allow for recruitment of specific transcription factors to modulate initiation efficiency (reviewed by Coombes & Boeke (2005)[6]). To achieve specificity in eukaryotes, transcription initiation needs to be a combinatorial process. The regulatory potential of transcription initiation unfolds in the interplay of these basic sequence elements with transcription initiation factors, genomic enhancers and their protein or RNA interactors, epigenetic modifications on histone or DNA level[7], histone occupancy, promoter binding proteins and finally the 12 subunits of polymerase II (polII). Thus, not the genomic sequence alone governs regulation but a complex interplay of many factors is involved. Their individual availability, proximity and state compose the bigger regulatory network which is involved at almost every aspect of gene expression.

In this introduction on transcription, I want to focus on the key elements that are important to understand the studies included in this thesis. These include regulatory RNA elements and R-Loops.

TRANSCRIPTION REGULATION BY RNA

The mouse small RNA B2 was one of the first RNAs described to interfere with trancription by directly and reversibly binding to RNA polII[8]. This trans-action acts on the preinitiation complex, preventing transcriptional initiation and plays a role in regulating transcription globally during heat-shock responses.

Recent work in our lab has utilized the SELEX approach[9] to identify further RNA polyemerase interacting RNAs on a genome-wide scale. This study was carried using the *S. cerevisiae*, *E. coli* and human in-vitro transcribed genomes and respective polymerases, leading to a wealth of genomically encoded RNA aptamers. The action of these aptamers (RAPs for RNA aptamers of polymerase), the impact on the respective RNA polymerase and their general feature study is subject to a number of manuscripts, which are currently in the publication process (yeast RAPs: Klopf et al., (under revision); E. coli: Sedlyarova et al., (Mol Cell 2017, June 19, patent filed); human: Boots et al., (under revision and attached (manuscript 3)); Matylla-Kulinska et al., (manuscript in preparation and attached (manuscript 2))).

The activity of these RNA aptamers is in all tested cases in-cis, making them potential regulators of their hosting gene expression. Among the strongest polII binding RNAs was a sequence of nucleotides found in the human repeat element ACRO1 satellites. Similar to previously identified polymerase RNA aptamers[8,10], this interaction leads to transcriptional senescence, arguably a mechanism of blocking repeat element transcription. Another repeat element with functional RNA aptamers are α-satellites, a set of centromeric repeats, spanning up to multiple megabases. Interaction of these RAPs with polII does not lead to transcriptional interference, as the above described cases, but leads to RNA-dependant RNA polymerase activity of polII, resulting in a reverse complement transcript to the original α-satellite RNA. Besides this RAP-activity, the attached study sheds light on α-satellite transcription, turnover, structure and localization.

R-LOOPS

A newly synthesized RNA has by its nature, complementary sequence to the adjacent DNA strand. It occurs that this RNA, after leaving polII reintegrates itself into the DNA double helix, forming a DNA:RNA hybrid and a displaced ssDNA strand. This structure is termed R-loop and has a number of implications on transcription and genome integrity. It has been determined that G-rich sequences induce R-loop formation[11], and thus this process has emerged as a potential mechanism for polII transcription termination, as polyA sites are often associated with G-rich sequence elements. The occurrence of an R-loop leads to the recruitement of a RNA:DNA helicase (Sen1 in yeast, Senataxin in human), unwinding the hybrid and inducing transcription termination. It must be desired by the cell to solve R-loop hybrids quickly as they, as just mentioned, might serve in transcription termination, but secondly have been shown to induce genomic instability. More specifically, the discplaced ssDNA strand shows higher mutagenesis susceptability than a stable dsDNA helix. Additionally, the R-loop RNA can serve a as primer for mutageneic DNA polymerase elongation[12]. The celullar actions for R-loops solving are not yet fully understood. Nrl1, a protein identified by Aronica et al., has been recently shown to play a role in this process (manuscript (4) attached, [13]) and furthermore links R-loop formation to DNA damage response and splicing.

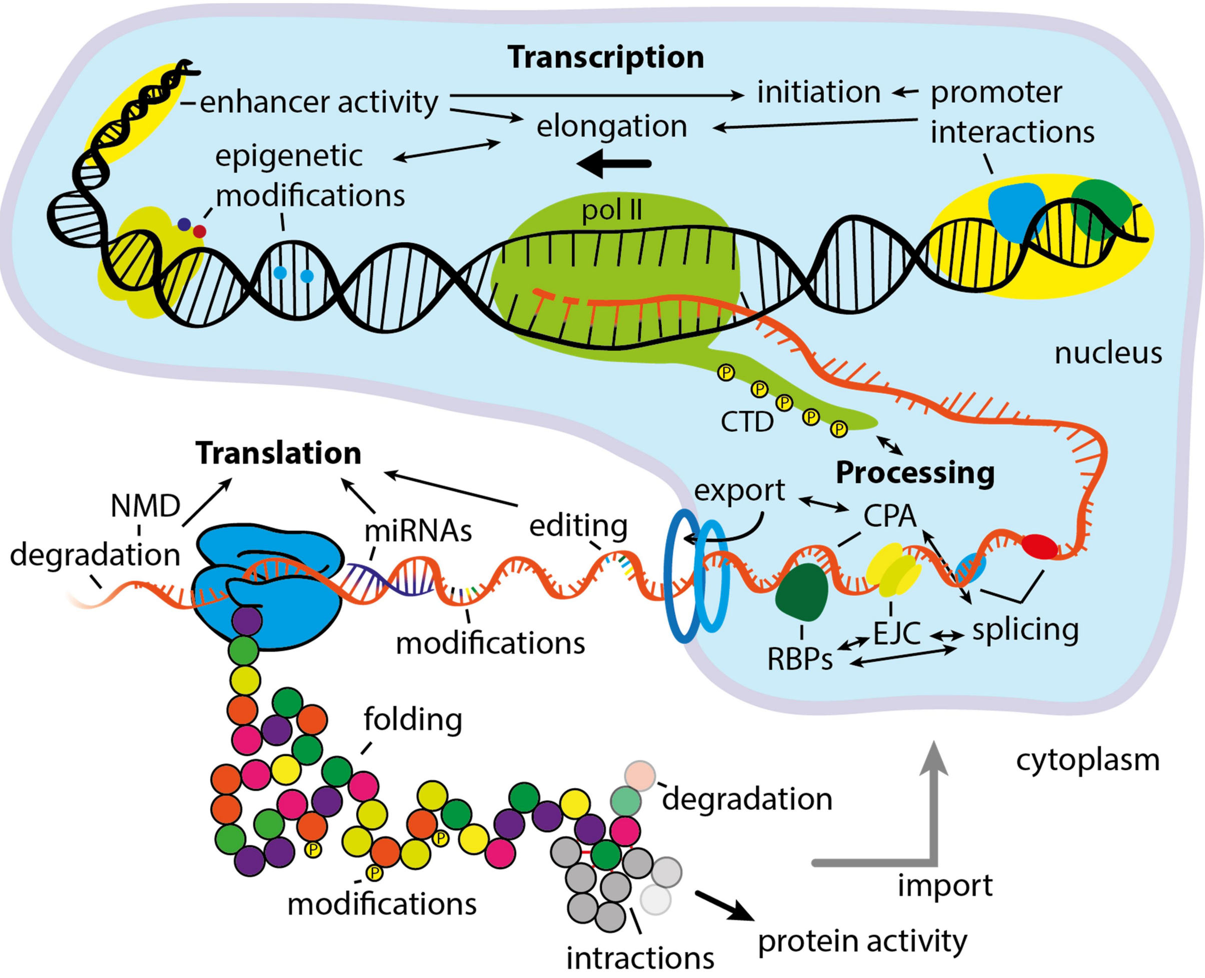


Figure 1: Regulatory layers of gene expression from genome to proteome. **Transcription initiation** is regulated by enhancer elements, epigenetic modifications, such as histone acetylation or DNA methylation, and protein-promoter interactions. **Transcriptional elongation** is controlled by various factors that modify stalling, dissociation, processing and elongation speed. **RNA processing**, the next layer with regulatory capacity involves 5'capping, splicing, cleavage and polyadenylation (CPA) and deposition of RNA binding proteins (RBPs), such as the exon junction complex (EJC) or polyA-binding protein. These RBPS induce mRNA export. Cytoplasmic mRNA can be further modified by RNA editing, RNA modification (epitranscriptome) and regulatory miRNAs. If **translation** is initiated, translation speed and efficiency determine the mRNA stability. Nonsense transcripts will be quickly turned over by the nonsense-mediated degradation pathway. The amino acid chain produced in the translation process is subjected to further regulatory procedures, involving protein folding, posttranslational modifications, protein-protein interactions and, eventually, degradation.

RNA PROCESSING

The transcriptome undergoes constant evolution and is far from being a static entity. The first modifications to each RNA molecule occur during its biogenesis, i.e. co-transcriptionally. Depending on the transcribing polymerase, its processing speed and CTD state, the nascent RNAs sequence and protein occupation, the transcript can undergo various interacting modifications.

5'CAPPING

Each transcriptional initiation event is followed by a transcriptional pause before polII either dissociates or is released into elongation. Predominantly during this promoter proximal pausing, the 5’ triphosphate of the first nucleotide of the RNA is modified by a phosphatase and a guanylyltransferase mediates the addition of a guaninemonophosphate. This results in a triphosphate bond between the first nucleotide and the newly added guanine. Then a methyltransferase (RNMT) transfers a methyl group to the N7 position of the guanine. Therefore the cap has the structure: m7G (5’) - ppp - (5’)[A,U,C,G].

The capping enzymes, which are partly recruited to the elongating polII, as well as the cap itself with its binding proteins, influence downstream processing, including: transcriptional elongation, splicing, 3’ end processing, stability, nuclear export and translation (reviewed by Cowling et al. (2009)[14]) and thereby provide the first elaborate regulatory platform of the nascent RNA.

SPLICING

In 1977, the (physical) alignment of adenoviral RNA with its respective DNA origin led to the discovery of the discontinuity of genetic information and the necessity to process mRNA for protein synthesis[15,16]. Discovering the concept of split genes and gene splicing was fundamental for the understanding of mRNA biogenesis and, later, for explaining the proteomic complexity of numerous eukaryotes by means of alternative splicing[17]. The protein coding sequences were termed ‘exons’, the removed sequences ‘introns’.

Similar to transcription, the basic sequence requirements to initiate splicing are limited to a small set of sequence patterns. 5’ and 3’ splice sites (ss) flank the intron and a short branchpoint consensus followed by a polypyrimidine (pY-) tract defines the position for lariat formation. These sequence patterns (see **Figure 2A**) are short and show limited variation in their nucleotide composition which allows only for limited affinity tuning. Therefore the major regulatory impact on splice site selection and splicing efficiency lies within the interplay of the spliceosomal components with signals of the transcriptional machinery, transcription speed, chromatin and nascent RNA binding proteins. Before diving into these regulatory interactions, we will have a look at core spliceosome assembly:

Spliceosome assembly and activity

5’ splice site definition is highly sequence dependent as base pairing between the U1snRNA 5’ end and the nascent RNA defines the beginning of an intron. Therefore, the U1snRNA sequence defines the 5’ splice site consensus sequence: 5’-$_{m3}G\mathit{ppp}A_{m}U_{m}AC\psi\psi ACCUG$ (U1snRNA) binds most efficiently to 5’-$_{C}{}^{A}AG|GU_{A}{}^{G}AGU$-3’ [18]. This binding event is also the first step in spliceosome assembly. The 3’ end of the intron is occupied by SF1 and U2AF, binding to the branchpoint and pY-tract, which will be later replaced by U2snRNP, defining the 3’ splice site and forming the A-complex (**Figure 2B**). Depending on the length of the intron, the presence and interaction of U1, SF1 and U2AF can suffice to define an intron. If interaction between these factors is distorted by spatial distance, U2 and U1 can interact via SR-proteins covering the exon, leading to exon definition.[19–22]

In the next step the snRNPs U4, U5 and U6 are recruited together with a multitude of proteins. This preformed tri-snRNP, its recruitment and its conformational changes of the B-complex, resulting in its activation, have recently been structurally solved due to the advances in cryo-EM[23]. This allows for a very precise understanding of the steps involved in generating a functional spliceosome. Similar work, on other spliceosome conformations allow for a very complete structural understanding of spliceosome composition, transformation and activity[23–28].

The activated spliceosome (B*-complex) consists of the core subunits U2, U5 and U6, after the dissociation of U4snRNP. Before catalysis of the first splicing reaction can take place, components of the exon junction complex (EJC) are recruited to the spliceosome. Then the nucleophillic attack of the branchpoint A's 2'OH on the 5' phosphate of the 5'ss guanine is catalysed. This results in the formation of a lariat intermediate that is still connected to the 3' part of the nascent RNA. Further conformational reconfiguration of the U2-U6-U5 complex and interaction of further spliceosomal proteins lead to the second nucleophillic attack of the free 3'OH of the 5' part of the RNA on the phosphate at the 3' end of the 3'ss guanine. This results in a covalent linkage between the exonic nucleotides and release of the branched lariat structure. The EJC proteins are deposited on the RNA[29] whereas the post-splicing complex, consisting of U2, U5, U6 and numerous spliceosomal proteins dissociates, allowing for recycling of the spliceosomal components. Very comprehensive reviews of this process have been published by Wahl & Lührmann (2015)[22,30,31], Papasaikas et al. (2015)[20] and Matera & Wang (2014)[19].

The complexity of the spliceosome composition (>100 proteins involved) and its dynamic nature provide a perfect platform for regulatory tuning.

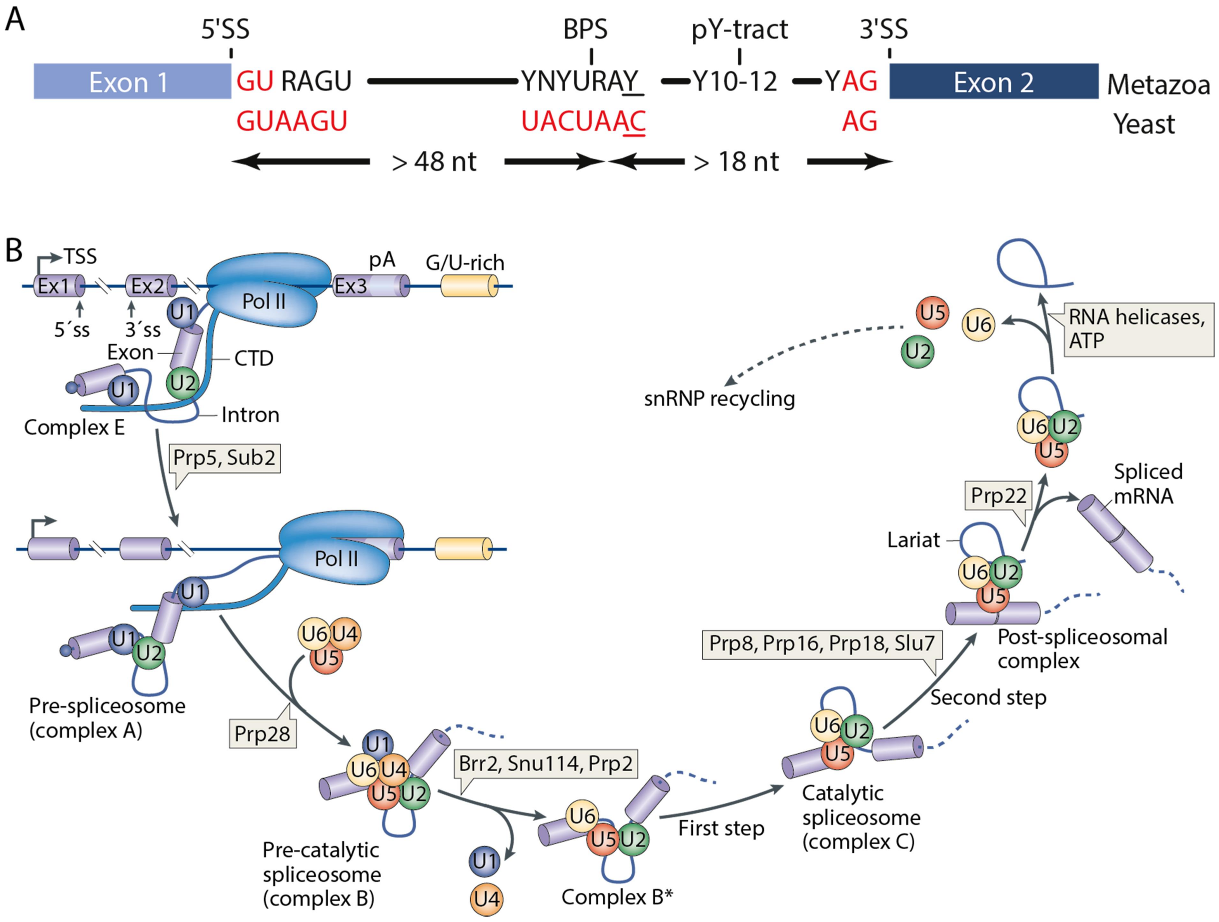


Figure 2: Consensus sequences of the basic intronic reactive sites, defining exon-intron boundaries and binding sites of spliceosomal components (A, adapted from Wahl & Lührmann (2015)[22]; B) depicts a typical and simiplified spliceosomal assembly-splicing-dissocation live cycle. Initial splice site recognition by U1 and U2 forms the A-complex. Deposition of the tri-snRNP U4-U5-U6 leads to the dissociation of U1 and U4 forming the B-complex. Structural rearrangements and further protein interactors activate the spliceosome and facilitate the two nucleophillic attacks required for intron removal and exon ligation. The remaining snRNPs dissociate with the excised lariat and are recycled for subsequent splicing steps (adapted from Matera & Wang (2014)[19]).

Transcription and splicing

Transcription and splicing are coupled processes. Before any splicing commitment can occur, the splice sites need to be transcribed. The influence of polII transcription rates and pausing on splice site selection are well studied[32,33]. This especially impacts on alternative splicing. For example, a low polymerase elongation speed favours the inclusion of weak splice sites, if they are located upstream of a potentially stronger competitor splice site. The weaker splice site will be bound by respective cofactors and committed to splicing, before the downstream splice site is even transcribed. As stated earlier, transcriptional elongation speed depends mainly on the promoter and the type of proteins associated with the transcriptional machinery. Thus promoter modifications or a change in local concentrations of promoter interactors can influence alternative splicing patterns.

A more direct influence on splicing is carried out by the RNA polymerase II itself. Experiments by

McCracken et al.[34] showed that removal of the RNA polII carboxyl terminal domain (CTD) results in loss of RNA processing, including 5' capping, splicing and 3' end processing. Follow up studies revealed more details and showed that many of the spliceosome-recruitment components, such as U2AF (recruits U2, Prp19), Prp40 (U1) or PSF can be found directly associated with the polII CTD. Vice versa, there is proof that splicing also feeds back to transcription, stimulating elongation (reviewed in [35]). These two observations are strong indicators for widespread occurrence of co-transcriptional splicing, or at least splicing commitment.

The ever expanding repertoire of RNA-seq protocols includes also an approach that allows for the quantification of nascent RNA and thus elucidates co-transcriptional RNA processing. A study by Khodor et al.[36] unveils that 87% of introns in *D. melanogaster* are fully spliced during transcription. Exceptions are very long introns and introns involved in alternative splicing, as they might require additional factors determining their splicing commitment. Very similar results have been obtained by Oesterreich et al. (2016) in yeast, by tracking polII position relative to splicing of the nascent intron[37]. They found an onset of splicing after, in average, 26 nt and completion after 129 nt continued transcription. This allows for an estimation of 1.4 seconds intron dwell time before splicing is completed. This tight spatial and temporal boundary suggests a very strong link between transcription and co-transcriptional splicing. Advancing methods such as full length, direct RNA sequencing will allow for a more complete picture of this process.

Further layers of splicing regulation

The strong connection between transcription and splicing implies an at least indirect connection between chromatin state and splicing. Whether or not there is also a direct interaction was unclear until a recent study by Allemand et al. (2016)[38] showed that U2snRNP engaged in splicing interacts with chromatin factors and remodelers. Additionally they showed that knockdown of various chromatin factors impacts on splicing. Yet, the precise mechanisms of these interactions remain to be solved, but likely different chromatin factors interact with specific splicing factors, modulating their activity.

Another layer of splicing regulation, and more specifically splice site selection is added by short sequence elements on the nascent RNA that recruit splicing modulators. These sequence elements, called splicing enhancers or silencers, based on their activity can be located in the exon (exonic splicing enhancer/silencer: ESE/S) or intron (ISE/S). Enhancer elements are mainly bound by members of the serine-arginine (SR) rich protein family of splicing factors which promote usage of proximal splice sites whereas members of the heteronuclear ribonucleoproteins (hnRNPs) have been shown to prevent U1 binding and thus negatively influencing splice site selection (reviewed by Lee and Rio (2015)[39]). These sequence elements play a major role in alternative splice site selection and are therefore a highly valuable potential therapeutic target.

Alternative splicing

Most of the above mentioned regulatory layers are not tailored to initiating splicing but rather to fine tune splice site selection and modulate isoform expression based on external signal transduction. Oftentimes, this signal is merely a low or high concentration of a certain subset of splicing factors, which in return promotes a certain splicing pattern.

These splicing patterns are the bridge between the low genomic and the high proteomic diversity. The rise of RNA sequencing technologies over the last decade allowed for the estimation that around 95% of all human multi-exonic genes produce more than one mRNA isoform. This diversity plays a crucial role in generating tissue specific transcriptomes and proteomes with the brain holding the majority of tissue specific splicing events as well as splicing factors[40]. The process of alternative splice site selection and other mechanisms of shaping isoform expression have been reviewed by Cáceres & Kornblihtt (2002)[41], Wang et al. (2008)[4], Dujardin et al. (2013)[32], Roy et al. (2013)[17], Kornblihtt et al.(2013)[42], Lee & Rio (2015)[39] and Francastel & Hubè (2015)[43]. As this topic is extensively covered in the literature I only want to point out one example that shows the power of alternative splicing and then discuss a recent controversy that arose over the role of alternative splicing in generating proteome diversity.

Animals with high cognitive capabilities have complex neuronal networks where single cells transduce signals to other single cells via extensive dendritic arborisation. During the arborisation process, it has to be avoided that a branch of a dendrite comes into contact with another branch of the same cell and thus forming a short circuit. The mechanism of this self-avoidance was first identified in the fruit fly *D. melanogaster*. The gene dscam1 codes for a surface protein with a homophilic extracellular domain that, if it binds to its exact same isoform, induces a repulsion response in the dendritic branch, preventing synapse formation. Due to the number of dendrites in a neuronal complex, the number of isoforms of dscam1 has to be immense. The genomic organization is the following: the first variable is a set of 12 mutually exclusive cassette exons, followed by other sets of 48, 33 and 2 exons, respectively (**Figure 3**). Of each of these cassettes, one exon is picked allowing for a total of >38.000 isoforms. The resulting diversity in homophilic protein binding domains is the key for dendritic self-avoidance (reviewed in Hattorin et al. (2008)[44] and Zipursky & Grueber (2013)[45]) and clearly demonstrates the regulatory power of alternative splicing.

Recent studies dealing with the diversity of the transcriptome started questioning the translational engagements of all these isoforms. Basis for this claim is the evaluation of a set of large scale mass-spectrometry experiments that aimed to achieve a full picture of the human tissue specific proteome. Of all genes identified (12,716) only 246 showed peptides of more than one isoform[46]. One major isoform per gene is found and derivations are declared untranslated. But the authors also state the difficulties of comparing deep sequencing data with mass spectrometry. Short RNA-reads allow for a very sensitive approach, yet, isoform assembly remains problematic. Mass spectrometry on the other hand is not as sensitive and gene coverage is often incomplete. These two properties make it challenging to claim translational inertness of the majority of isoforms detected by RNA-seq. Another approach towards this question has been taken by Weatheritt et al. (2016)[47]. They used ribosome profiling to determine, which fraction of the transcriptome is actively undergoing translation or at least engaged with a ribosome. According to their finding, the majority of annotated alternative isoforms is engaged in translation and therefore contribute to proteome diversity. The only exception are intron-retention isoforms, which due to nuclear retention or nonsense-mediated decay show reduced or no ribosome engagement. These results therefore clearly confirm the role of alternative splicing as a major contributor to proteome diversity.

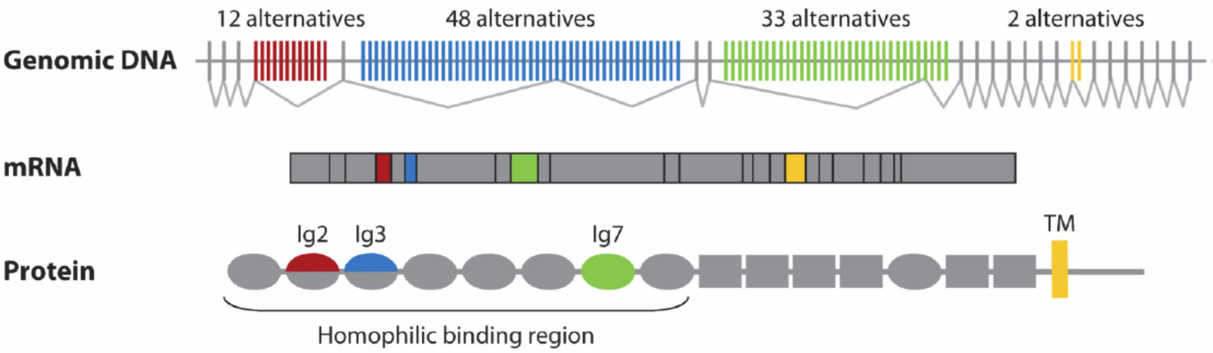
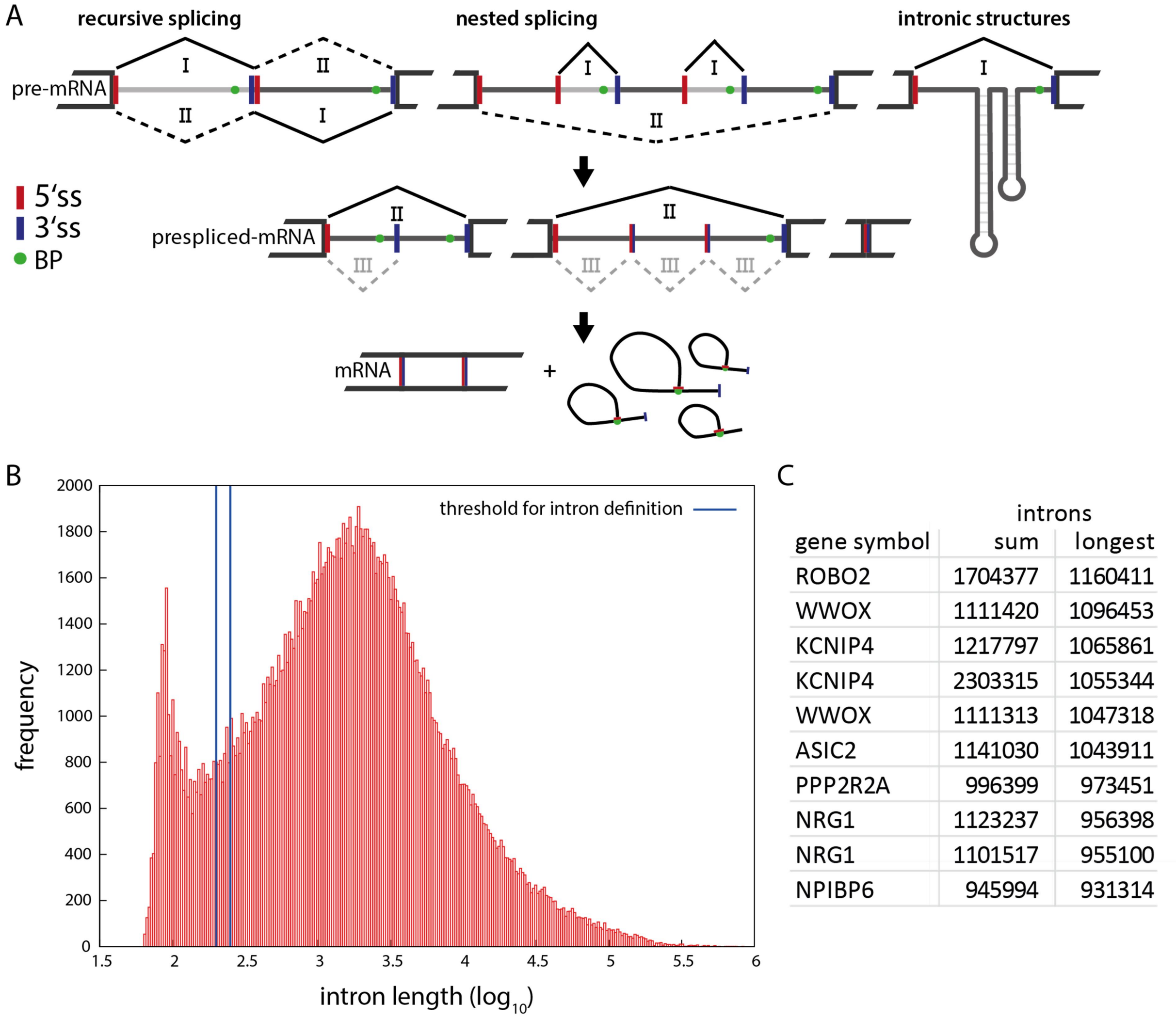


Figure 3: The genomic organisation of DSCAM shows the potential of alternative splicing. From 4 adjacent exon blocks with 12, 48, 33 and 2 potential exons respectively, one is selected each to constitute the protein binding domain and the transmembrane domain. This mutually exclusive splicing pattern allows for more than 38,000 isoforms which are required for the function DSCAM has in neurogenic self-avoidance (adapted from Hattori et al. (2008)[44]).

Splicing of long introns

As outlined above, interaction of RNA binding proteins, snRNPs and RNA sequences is key to spliceosome assembly and consequently efficient intron removal. Since proper genome annotations of complex eukaryotic organisms are available, the length of some introns seems to put an unanticipated challenge to sequence-based ribonucleoprotein-complex assembly. The genecode.v24 annotation[48] for example annotates close to 195 million nucleotides as exonic and 2,97 billion as intronic, which are arranged in 359,586 introns with an average length of 8262 nucleotides. 72 introns are larger than 500,000 and 12,123 larger than 50,000 nucleotides (**Figure 4B,C**). Besides impacting on transcription times, these large introns increase the spatial distance between functional splice sites. In genes with long introns, the exons get defined and prone to splicing. Yet, the intronic ends must still be rearranged to get into contact and allow for spliceosome assembly. Multiple potential solutions have been hypothesized and experimentally proven (**Figure 4A**), yet to date, a full understanding of all mechanisms involved in long intron removal is beyond reach. As outlined above, splicing is often an immediate consequence of splice site transcription. As such it appears unlikely that some splice sites remain inert until the corresponding counterpart is transcribed which can take hours in the longest introns. In 1998 the first experimental verification of recursive splicing was achieved[49]. A 74kb intron of ultrabithorax in *D. melanogaster* was found to be removed stepwise with the usage ratcheting points (RPs), sequences that reconstitute a functional splice site after a first splicing step. These “zero-length” exons were found in other fruit fly genes, yet are a rare occurrence in mammalian genes. Therefore an alternative mechanism for mammalian long intron removal has been proposed based on intron structuring[50]: this study found an abundance of stem-loop structures formed by intronic repetitive elements, which would lead to intron shortening and facilitate splice site interaction. Yet, this computational prediction still waits for its experimental verification, proving its general applicability.

In the meantime two studies set out to discover genome wide recursive splicing patterns in humans and *Drosophila*[51,52]. As projected by previous reports[49,53], many recursive splicing events were found in long introns in the fruit fly, yet only a few in vertebrates. For a few human genes, the presence of so called RS-exons was confirmed in this study. RS-exons are basically “non-zero exons” which are temporarily included in the spliced RNA and then eventually removed. This allows for bridging long introns with an additional splicing step and implements additional regulatory capacity as an alternative exon is available to shape the final isoform.



	introns	
gene symbol	sum	longest
ROBO2	1704377	1160411
WWOX	1111420	1096453
KCNIP4	1217797	1065861
KCNIP4	2303315	1055344
WWOX	1111313	1047318
ASIC2	1141030	1043911
PPP2R2A	996399	973451
NRG1	1123237	956398
NRG1	1101517	955100
NPIBP6	945994	931314

Figure 4: Removal of long introns could occur in various ways. Three plausible hypotheses are depicted in (A): recursive splicing (RS) is a splicing cascade with multiple subsequent splicing steps to intronic ratcheting points. Nested splicing utilizes only intronic splice sites and thus leads to intron shortening. Intronic secondary structures formed by repetetive sequence elements that are often located in introns could also physically contribute to intron shortening, bringing the exonic splice sites in closer proximity. (B) The intron length distribution in humans shows two distinct peaks. One just below 100 nucleotides, defining an intron class that splices via intron definition, and another, broader peak between 1000 and 4000 nucleotides. Thus the vast majority of introns are smaller than 4 Kb, yet a few extend to up to 1 Mb as shown in (C). Here the genes with the 10 longest introns according to genecode.v24, are listed. Figure B is adapted from Osella & Caselle (2009)[54].

Irregular splicing events

Recursive splicing of long introns is not the only splicing event that deviates from the initially anticipated splicing model. The first study on how splicing can influence gene expression and modulate isoform selection outside the framework of alternative splicing was carried out by Parra et al. (2008)[55]. They discovered that the human 4.1R gene harbours three alternative, mutually exclusive first exons (1A, 1B and 1C) of which 1A always splices to a distal splice site within exon 2 whereas 1B and C always splice to a proximal splice site at the intron-exon boundary of exon 2. Minigene constructs and splice site analysis led to the conclusion that in order to achieve this specific splice site selection of exon 1A, two consecutive splicing events are required. When exon 1A is transcribed, the first splicing step occurs from exon 1B to the proximal exon 2 splice acceptor, the second splicing step ligates exon 1A to the distal splice site in exon 2. In this way, the first splicing step, promoted by a strong splicing acceptor, removes splicing-relevant sequences at the exon 2 proximal site, and the second splicing step to the weaker acceptor can be carried out precisely. This type of splicing was due to the nested nature of splicing events termed nested- or intrasplicing. A follow-up study confirmed this specific type of splicing in other vertebrates and narrowed down the mechanism and regulatory sequence elements involved in splice site selection[56].

In the case of the 4.1R gene, intrasplicing serves the purpose of modulating the N-terminal domain of the protein, yet originally, this mechanism was proposed to serve, similar to recursive splicing, as a means of splicing large introns consecutively[57]. Only recently a first evidence of splicing events, presumably serving this purpose, was reported[58]. Suzuki et al. report the presence of lariats arising from intrasplicing events in human *dmd* intron 7 and 8. These cover a small fraction of the total intron length and are therefore questionable candidates for intron shortening. Also, it has not been shown so far whether these intrasplicing events are a necessity for full intron removal or merely an occasional by-product of snRNP binding.

Another type of splicing process that does not obey the conventions of regular intron removal is trans-splicing. This type of event ligates two exons which are not part of the same transcript. Genic trans-splicing generates chimeric transcripts by ligating exons of two different genes, whereas SL-trans-splicing fuses a splice leader (SL) sequence to another exon. In humans, trans-splicing events are often associated with cancer whereas nematodes rely on SL-trans-splicing for efficient operon processing. Even though the mechanism of SL-trans-splicing is well established, the precise steps of genic-trans-splice site selection remain convoluted. The SL-RNA is a highly specialized transcript that closely associates with the splice acceptor and the spliceosomal components, allowing for a regulated trans-splicing event (reviewed by Lasda & Blumenthal (2011) [59]).

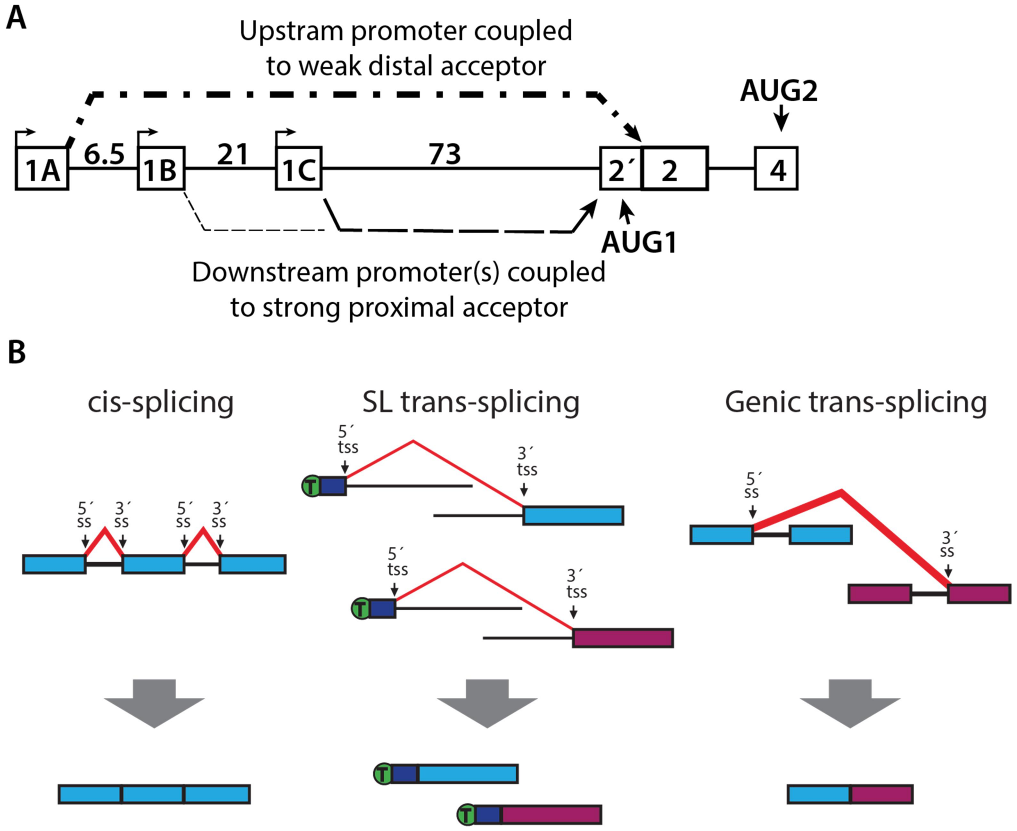


Figure 5: The first functional intrasplicing event was identified in the 4.1R gene (A). If the promoter of exon 1A is used, a first splicing step splices from exon 1B or 1C to 2’, then a second splicing step ligates exon 1A to exon2 (adapted from Parra et al. (2008)[55]). Another splicing mechanism, deviating from regular intron removal is trans-splicing (B). Here a splice leader (SL) of a different transcript or an exon from another gene is ligated to a 3’ exon. This generates modulated 5’ termini or chimeric transcripts (adapted from Lasda & Blumenthal (2011)[59]).

Indirect splicing modulation

A pre-requirement for intrasplicing is an early binding of U1 and U2 snRNPs to intronic splice sites. U1 is known for scanning the nascent RNA and binding numerous splice-site-like sequences with varying affinity. This predominantly serves the purpose of interfering with premature cleavage and polyadenylation (PCPA)[60], a process in which intronic sequences resemble 3’ end processing signals, such as the AAUAAA polyA signal. Via a to date not precisely understood mechanism, U1 snRNP, or rather one of its subcomponents, interferes with the CPA-machinery and prevents cleavage and polyadenylation (CPA) in the intron. As a result, frequent U1 intron binding might be a trigger for intrasplicing commitment, if a corresponding 3’ss is occupied, leading to intra-intron definition.

Another more indirect influence is taken by lariat stability. As briefly mentioned above, lariats are to a large extend quickly turned over to provide resources (nucleotides, RBPs, splicing factors, snRNPs) for subsequent transcription and processing events[61]. A key step in lariat degradation is breaking the 2’-5’ bond of the branch point[62]. This task is carried out by a specialized debranchase enzyme, encoded by the human gene *dbr1*. It has been reported that lariats that evade debranching evolved to take over additional functions in the human transcriptome. The most apparent being serving as a reservoir for RBPs that associated with the intron pre-splicing. Triggering the degradation of the lariat would release these proteins, making them available for further processing steps. Also, lariats have been reported to be exported from the nucleus, persisting in the cytoplasm, playing a major role in virus latency[63].

As one of the approaches to generate a rich splicing-focused RNA-seq dataset for intra-splice site detection (manuscript 1 of this thesis) I attempted to enrich lariats artificially via a transient dbr1 knockdown. A similar strategy was used successfully in *S. pombe* where the increased lariat stabilization lead to the discovery of numerous novel splicing and alternative splicing events[64]. However, it turned out that human cells with more complex genomes, i.e. more genes, with more and longer introns than yeast, do not tolerate a full dbr1 knockout and a knockdown only for a few days. Sequencing the total RNA pool of dbr1 knockdown cells additionally showed a strong change in the splicing landscape as compared to untreated cells (unpublished, see **Figure 6**). This observation is a strong indicator, that lariat turnover influences splicing patterns. Further experiments are required to validate this initial observation and to determine candidate lariats that exhibit a regulated degradation response.

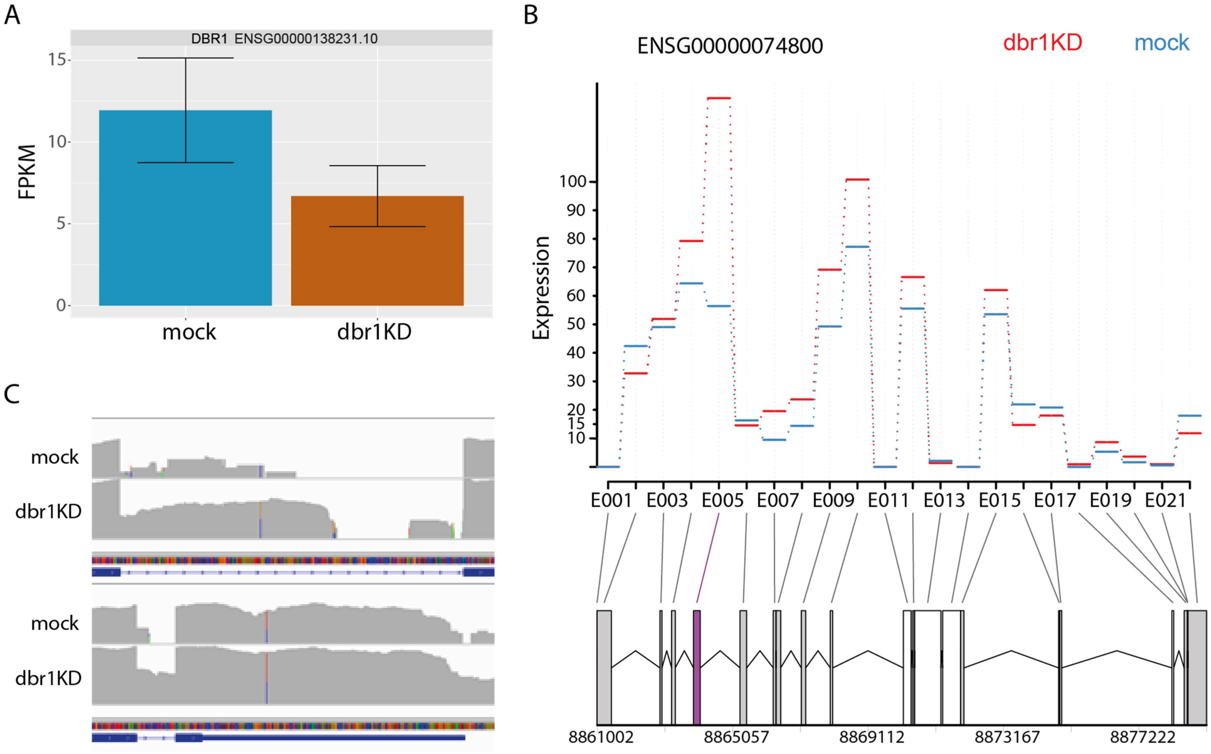


Figure 6: Knockdown of dbr1 (A) by 50% leads to altered isoform expression. DEXseq analysis shows a clear change in exon usage in the example gene (B). Red bars reflect the coverage of this exon in the dbr1 knockdown. The knockdown efficiency can also be determined by comparing general sequence coverage of introns (C). In the example introns shown, intron coverage is increased compared mock. Notably, coverage deacreases after two thirds of the intron, which indicates the branchpoint position. The knockdown does not lead to intron retention but stabilization of the lariats.

Methods for the detection of rare splicing events

The above described findings on *dmd* intrasplicing demonstrate the difficulties and obstacles in the discovery of rare or lowly abundant splicing events. Splicing detection is deteriorated by low transcript abundance, quick turnover of lariats and in the case of pre-splicing events such as recursive or nested splicing, commencing transcript processing. Lariat turnover and further transcript processing are processes believed to occur rapidly as there is mostly no function of intermediate RNA species. Recent approaches to overcome these limitations mainly focus on enriching the desired RNA species. This can be done biochemically, using the 3'-exonuclease RNAseR that digests most linear RNA, leaving circRNA and lariats intact[65–67] which can be sequenced and mapped to identify splicing events. Another approach is based on a genetic mutation that disrupts the functionality of *dbr1*, the gene encoding the debranchase responsible for lariat cleavage and thus induction of their degradation. This approach, applied in yeast, led to the discovery of a more complex splicing landscape than previously assumed[64]. Yet, the attempt to apply this approach in more complex, multicellular organisms is difficult as an increased number of introns and therefore an increased retention of lariats in a *dbr1* knockout depletes the cell of nucleotides and cofactors bound to excised lariats. This renders the knockout lethal and a partial knockdown is problematic as the resulting depletion of splicing factors heavily impacts on further processing steps (unpublished).

The approach I want to present here also aims for the enrichment of sequences of interest, but does so during selective RNA-seq library preparation steps. Targeted RNA-seq uses labelled probes that bind to a certain RNA species (e.g. branchpoints in [67]) or a certain gene (*dmd* in [68]) and allows to focus the entire sequencing depth of modern NGS approaches on this limited sequence pool. This results in a superior coverage of the sequences of interest and in consequence novel analytical approaches. The software suite SplicePie was specifically developed to make use of such an increased coverage, in combination with paired-end sequencing. It allows to determine splicing patterns, splicing sequentiality and also detects aberrant splicing events, such as recursive splicing[68]. Yet, due to the requirement of high coverage, it cannot be applied on a genome-wide scale.

Therefore the two most recent studies dealing with genome-wide splicing annotation in humans both apply RNAseR digest in order to obtain a more detailed view on the splicing landscape[66,67]. But both studies focus on branchpoint localization, conservation and sequence in proximity to annotated splice sites (i.e. 3' splice sites of annotated introns) and therefore miss to identify novel splicing events, such as recursive or intrasplicing. Thus, there is still need for an approach that utilizes biochemical and computational enrichment of splicing-relevant sequences and yet delivers the confidence to allow for novel splice site detection without the need of relying on available splice site annotations. This is exactly what the main work of this thesis focuses on and what can be found in the first manuscript attached to this introduction: *Genome-wide identification of intrasplicing events* *in the human transcriptome and hints to their regulatory potential.*

The exon junction complex

In order to understand and solve conflicts between the conventional splicing model and the accumulating proof of the abundance of intrasplicing events, we have to take a closer look at the role and regulatory impact of the exon junction complex (EJC). This multiprotein complex takes over a number of regulatory roles in the lifespan of an mRNA between splicing and translation. As briefly described above, the EJC is deposited on the pre-mRNA during spliceosome assembly. This implies that every splicing event results in EJC deposition. In fact CLIP-seq experiments with EJC component eIF4AIII confirmed this assumption[69]. After deposition, it modulates polII elongation rate by interactions with CTD kinases and facilitates mRNA export by interaction with export receptor TAP und thus tethering the RNA to the nuclear pore complex. The EJC is also the quality control milestone during translation: if an EJC is bound downstream of a stop codon, nonsense mediated decay is initialized. This mechanism prevents the translation of incompletly spliced RNAs. Most of these functions (elaborately reviewed by Woodward et al. (2016)[29]) deem the EJC deposition a signal for splicing completion. Intrasplicing, as described above, uses the very same spliceosomal pathway as regular, full intron splicing and thus also leads to the depositon of the EJC within the intron, next to intrasplicing events. In fact, the eIF4AIII-CLIPseq study shows a clear co-occurrence of EJC deposition sites and intrasplicing events (see manuscript 1). The role of the EJC in these scenarios is completely unknown as intrasplicing gained attention only recently. Therefore there is a need for studies that elucidate the role of the EJC bound to introns and how it interplays with the facilitation of another, subsequent splicing step.

Why introns?

As discussed above, the roles of introns are often of regulatory nature. They allow for isoform determination and gene expression regulation, provide the resource for downstream processing steps such as circRNA, snoRNA or miRNA biogenesis. Also mRNA stability, 3’end processing, export and translation is influenced by a preceding splicing step. Yet, the abundance and specifically the massive length of some introns and the resulting energy investment during this extense transcription seem staggering. Various hypotheses about the origin of introns and reasons for the cell to sustain them have been proclamated, yet due to the nature of evolution and the complexity of the human genome, it is hard to follow how introns ended up in their current form. One proposal sees splicing in the light of the recent discovery that the majority of the euchromatic genome is constantly transcriptionally engaged (reviewed by Fedorova and Fedorov (2005)[70]). The dense occupation of nascent pre- and mRNA by splicing factors is seen as a means of allowing the cell to distinguish between mRNA transcripts and transcriptional noise, arising from undirected transcription events.

So there seem to be many benefits of having functional introns. But why do some of them need to be so long? The presence of transposable and repetetive elements in introns provide explanation for intron-genesis and lengthening but it does not answer the question of why the cell tolerates this massive energy and time investment of transcribing up to megabases of intronic nucleotides in some genes, which can delay mRNA synthesis by several hours. Unless the hereby gained benefits from increased regulatory capacity outcompetes the energy investment. As this is quite unmeasurable, we might not get a complete picture in this respect and have to rely on hypotheses that proximate our observations.

3' END PROCESSING

The main focus of this thesis is RNA processing with emphasis on splicing. But the transcript would not be completed without 3'end processing. The mammalian consensus sequences defining the end of an mRNA or a polyadenylated ncRNA are the polyA-signal AAUAAA, the dowstream cleavage site CA and flanking sequence elements, termed upstream and downstream sequence elements (U/DSE). The main protein interactors are CPSF (cleavage and polyadenylation specificity factor, binding to the polyA signal), PAP (polyA polymerase, extending the polyA tail), CF I and II (cleavage factor, initiating cleavage of the transcript at the cleavage site CA), CstF (cleavage stimulatory factor, interacting with the DSE and CPSF) and the polII CTD. The process of transcription termination and polyadenylation and its regulatory influence on gene expression is reviewed in detail by Hollerer et al. (2014)[71]. Here I would like to focus on the influence on splicing on 3' end processing. For long time already, inclusion of the last intron in a minigene construct was known to increase gene expression. For some genes, last intron splicing is even a necessity for functional 3'end processing. Studies on the human β-globin and preproinsulin II gene revealed that last intron splicing, besides EJC deposition leads to physical interactions between U2snRNP and CPSF[72,73], prompting a stimulatory effect on 3' end processing efficiency. Vice versa, splicing rates decreased, when the AAUAAA polyA signal was mutated[74], making this interaction effective in both directions. This interplay between 3' end processing and splicing adds yet another regulatory layer to the already complex framework of RNA processing. With this step, an efficient gene regulation as well as a quality control mechanism is introduced that surveys mRNA abundance and splicing efficiency.

AIMS AND SUMMARY

Transcription is the process that implements the information of the evolutionary memory of our cells. This process creates the diversity in celullar regulators we observe and itself is highly regulated in every step involved. Figure 1 demonstrates some of these regulatory layers, yet the full extend of molecular interactions to maintain a system as complex as a biological cell is far from being fully understood or even discovered. This thesis aims to extend the knowledge on regulatory interactions in three directions:

1. Self-regulating RNAs: The presented studies on RAPs, spearheaded by the respective first authors, show an unprecedented diversity in RNA polymerase binding aptamers that allow for an extensive regulatory impact by the nascent RNA itself. This ranges from self-supression of repetetive sequence elements to self-regulation of protein-coding mRNAs and potentially to alterations of splicing patterns.
2. As Lucia Aronica, the first author of the respective study deciphered: Nrl1 couples DNA damage response, splicing and the transcriptional machinery to homologous repair mechanisms. These interplays are examplatory for the tight interactions that occur during transcription and their necessity to maintain genomic integrity during critical events, such as R-loop formation.
3. The main focus of this thesis: splicing. Even though much is already known about splicing regulation, factors impacting splice site selection, inter-splicing, alternative splicing, interactions between splicing and transcription, etc., a lot is still left to understand. Such as intrasplicing and its impact on not only long, but also shorter introns. The central study of this thesis aims to add to understanding this novel regulatory layer by which individual splicing events gain the potential to determine transcriptional output, not only quantitatively but also qualitatively. This includes induction of NMD, exon in- or exclusion and tuning splicing efficiencies.

Besides the in-detail research performed in the attached studies, the RAP database and the intrasplicing extraction pipeline are made puplically available to provide resources, others can use to complement or base their research on.

MANUSCRIPTS

1. Genome-wide identification of intrasplicing events in the human transcriptome and hints to their regulatory potential (accepted for full submission at eLife, accepted at bioRxiv: 159350; doi: https://doi.org/10.1101/159350 (2017))

   pages 27 - 60

2. Human α satellite transcripts are substrates for RNA Pol II and contain remnants of snoRNAs (manuscript in preparation)
   pages 61 - 93

3. RNA polymerase II-binding aptamers in human ACRO1 satellites disrupt transcription in *cis (under revision)*
   pages 94 - 121

4. The spliceosome-associated protein Nrl1 suppresses homologous recombination-dependent R-loop formation in fission yeast (Nucleic Acids Res. 44, 1703–1717 (2016))
   pages 122 - 136

eLIFE

24 Hills Road, **P** 01223 855346
Cambridge **W** elifesciences.org
CB2 1JP **T** @elife_sciences

Genome-wide identification of intrasplicing events in the human transcriptome and hints to their regulatory potential

Tracking no: 28-06-2017-RA-eLife-29990

Maximilian Radtke (Max F. Perutz Laboratories), Ismet Srndic (Max F. Perutz Laboratories), and Renée Schroeder (University of Vienna)

Abstract:
Alternative splicing is one of the major regulators of transcriptome diversity and individual isoform abundance. Identifying and understanding the scope and variety of splicing events is still an ongoing process. We established a novel pipeline to extract splicing events from specific RNA-sequencing datasets and identified numerous splicing events that did not span the entirety of the respective annotated intron. These events occurred in introns of all lengths. After confirmation of these splice events by conventional methodologies, we analyzed the impact of intrasplicing on full intron removal. This showed that these intronic splicing steps can be beneficial, deleterious or neutral for full intron removal. This in part confirms recent findings on recursive splicing events in humans and other vertebrates and further uncovers an additional level of transcriptome regulation based on a yet undiscovered level of flexibility and regulation of splice site selection and its impact on gene expression.

Impact statement: A novel regulatory function of intrasplicing events as well as a comprehensive dataset of a genome-wide identification approach of said events are provided and may serve as the basis of future research.

Competing interests: No competing interests declared

Author contributions:
Maximilian Radtke: Conceptualization; Software; Investigation; Visualization; Methodology; Writing—original draft; Project administration Ismet Srndic: Validation; Investigation; Visualization; Methodology Renée Schroeder: Supervision; Funding acquisition; Project administration; Writing—review and editing

Funding:
Austrian Science Fund (FWF): Maximilian Radtke, Renée Schroeder, F4308 F4301; RNA Biology DK: Maximilian Radtke, W1207-B09; University of Vienna: Maximilian Radtke, Ismet Srndic, Renée Schroeder The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Datasets:
Previously Published Datasets: Identification of expressed and conserved human non-coding RNAs: Nielsen MM, Pedersen JS, 2014, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45326, GSE45326; [E-MTAB-513] Illumina Human Body Map 2.0 Project: European Bioinformatics Institute, 2011, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30611, GSE30611; A survey of human brain transcriptome diversity at the single cell level: Darmanis S, Enge M, Quake SR, Sloan SA, Barres BA, Zhang Y, Caneda C, Hayden Gephart MG, Shuer LM, 2015, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67835, GSE67835; CircleSeq in Human Fibroblasts: Sharpless, N., 2012, https://trace.ddbj.nig.ac.jp/DRASearch/submission?acc=SRA050270, SRA050270; Genome-wide characterization of long nonpolyadenylated RNA: Chen LL, Yang L, Duff MO, Graveley BR, Carmichael GG, 2012, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22666, GSE22666; Genome-wide discovery of human splicing branchpoints: Mercer TR, Clark MB, Andersen SB, Brunck ME et al., 2014, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53328, GSE53328

Ethics:
Human Subjects: No Animal Subjects: No

Author Affiliation:
Maximilian Radtke(Department of Biochemistry and Molecular Cell Biology,Max F. Perutz Laboratories,Austria) Ismet Srndic(Department of Biochemistry and Molecular Cell Biology,Max F. Perutz Laboratories,Austria) Renée Schroeder(Biochemistry and Cellbiology,University of Vienna,Austria)

Dual-use research: No

Permissions: Have you reproduced or modified any part of an article that has been previously published or submitted to another journal?
No

Genome-wide identification of intrasplicing events in the human transcriptome and hints to their regulatory potential

Maximilian Radtke[1,] Ismet Srndic[1,] Renée Schroeder[1]

[1] Department of Biochemistry and Cell Biology, Max F. Perutz Laboratories, University of Vienna,Dr. Bohr-Gasse 9/5, 1030 Vienna, Austria

Keywords:

Intrasplicing, human transcriptome, nested splicing, recursive splicing

Abstract

Alternative splicing is one of the major regulators of both, transcriptome diversity and individual isoform abundance. Therefore regulation of alternative splicing is crucial and yet, due to the complexity of the human genome (23.000 genes, most of which can be alternatively spliced) diverse and multileveled. Identifying and understanding the scope and variety of splicing events is still an ongoing process. We established a novel pipeline to extract splicing events from specific RNA-sequencing datasets and identified numerous splicing events that did not span the entirety of the respective annotated intron but used intronic splice sites. These splicing events could be generally categorized into three groups: 5'recursive (using the exonic splice donor and intronic acceptor), 3'recursive (using an intronic splice donor and an exonic acceptor) and nested (using two intronic splice sites). Surprisingly, the splicing events we found occurred in introns of all lengths, but generally followed the abundance scheme of all introns, i.e. most were found in introns between 500 and 5000 bps. After confirmation of these splice events by conventional methodologies, we further analyzed the impact of intrasplicing on full intron removal. For this we established a luciferase-based reporter which showed that these intronic splicing steps can be beneficial, deleterious or neutral for full intron removal. Thus intrasplicing events can be crucial for determining the transcriptional output. This in part confirms recent findings on recursive splicing events in humans and other vertebrates and further uncovers an additional level of transcriptome regulation based on a yet undiscovered level of flexibility and regulation of splice site selection and its impact on gene expression.

Introduction

Precursor messenger RNA (pre-mRNA) splicing is not only a key step in protein production, but is also critical for the regulation and expansion of the functional proteome. The human genome contains an unexpectedly low number of genes, approximately 23,000 and alternative splicing in combination with posttranslational modifications is known to diversify these into 90,000 proteins[1]. High-throughput sequencing studies suggest that up to 100% of human multi-exon genes produce at least two alternative mRNA isoforms. The majority of human introns contain 5'- and 3'-splice sites with a consensus sequence of variable conservation providing room for many ways how to define a splice site[2]. Combinatorial events including RNA-protein interactions and RNA-RNA interactions have been shown to be essential for determining splicing events, especially in long introns. The longer the intron, the higher the precision needed to control splice site selection.

Mechanisms of long intron splicing

Different mechanisms of long intron splicing in higher eukaryotes have been proposed over the last two decades, with two main models emerging: recursive and nested splicing[3–6] (Figure 1). Nested splicing is an intron-internal splicing event that does not use the primary or exonic splice sites. Whether or not this type of splicing is a necessity for effective splicing of a subset of (longer) introns is still unsolved. Splice site-like sequences, dispersed over introns are able to prevent premature cleavage and polyadenylation (PCPA)[7] by binding U1snRNPs and could promote intra-splice site usage. In a single case, the regulatory impact of intrasplicing has been demonstrated: in the 4.1R gene alternative promoter usage in combination with subsequent and first-exon-dependent intrasplicing determines the expressed isoforms[8].

Recursive splicing (RS), in contrast to intrasplicing, involves incremental splicing reactions. One of the primary splice sites is involved and intronic splicing reconstitutes a functional, recursive splice site (RS-site) that is utilized for subsequent splicing steps (Figure 1A). Two recent studies analyzed recursive splicing in depth and on large scale in drosophila and vertebrates, demonstrating that this mechanism is prevalent in the longer introns. In humans, RS does not seem to be a necessity for long intron removal but might rather have a regulatory impact on in- or exclusion of RS-exons[9]. And as shown in *D. melanogaster*, intron length cannot be the only requirement to induce recursive splicing cascades[10], making space for regulatory potential of these splicing events.

Detection of splicing events

One of the major obstacles in detecting and verifying intermediate splicing events such as nested and recursive splicing, is the temporary nature of their products. Partially spliced or pre-spliced pre-mRNA is subjected to immediate further processing (i.e. subsequent splicing steps). Lariats arising from recursive or intrasplicing events are, presumably as most lariats, turned over by debranching with Dbr1 and exonucleolytic degradation of the intron RNA within minutes after splicing[11,12]. A combination of PCR-based methods, lariat enrichment and deep sequencing can be applied to partially overcome these restrictions[13,14]. Lariats can also be stabilized by interfering with the degradation pathway, mainly by inhibiting debranching. Lariat sequencing as performed by Awan et al.(2013)[14] took advantage of a Δ*dbr1* strain, resulting in the identification of hundreds of novel splicing events in *S. pombe*. In more complex genomes which require a more stringent splicing regulation due to the increased abundance of potential alternative splicing events and longer host introns, we found that Dbr1 knockdown to non-lethal levels (~40% of wildtype) resulted in a major shift in genome wide splicing patterns and introduction of new, aberrant splicing events (unpublished). Therefore this approach is not applicable, at least in human cells.

Another restriction in lariat identification, if approached with RNA-sequencing, is an increased occurrence of mapping artifacts due to repetitive elements within introns and general sequence similarities at splicing regulatory sequence elements. These artifacts deteriorate when reads spanning either exon-exon or branchpoint-5'ss junctions (split-reads) are required to identify lariats and splice sites, reducing the effective mapping length (Figure 2).

Previous studies[15–18], dealing with the detection of branchpoints[16] or novel splicing events strongly rely on splice site annotations, which allows an event prediction with very high precision, but does not aim at the detection of novel, intronic splicing events. A promising approach in this direction is targeted RNA-seq with extremely high coverage and subsequent analysis by SplicePie[17] that allows to draw a wealth of conclusions about splicing sequentiality, unusual splicing events and intron dwell times. Yet, this analysis focuses on single genes to achieve the required high read coverage and does not allow for a genome wide analysis.

Lariat Spanning Linear Split Reads

In this work, we performed an in depth analysis of existing RNA-sequencing datasets of human cells using detailed bioinformatic analyses to identify unusual splicing events in a genome wide manner. To overcome the above described restrictions, we applied an approach that makes use of the vast quantity of published RNA-seq datasets, focusing on those generated with a library preparation protocol that enriches RNA species that are likely to hold information on unusual splicing events. The novelty of this procedure is the combination of reads covering the linear splice junction with reads that arise from the branchpoint-5' splice site junction of lariats (Figure 2). These **la**riat **s**panning **li**near **s**plit reads (LaSLis) therefore represent a set of high confidence splicing predictions. This approach resulted in the identification of more than 90,000 splicing events (neglecting alternative branch points), of which 5,693 are not covering the full length of annotated introns and are therefore potential recursive or intrasplicing events. As recursive splicing in the context of long introns has already been thoroughly investigated[5,9,10], we focused on those events found in shorter introns (<5000 nucleotides), which comprised the majority of the extracted splicing events (see Figure 3A). The logic behind the model of recursive splicing as a means for facilitating long intron splicing excludes the necessity of such events for short intron splicing. Therefore, these splicing events could hold regulatory capacity.

Results

Extraction of novel splicing events

The aim of this work was to identify rare and non-canonical splicing events in the human transcriptome and to test if these events affect splicing output, which would point to a potential regulatory function. To achieve this, we established a novel extraction pipeline, designed to extract transient splicing events with high confidence. This approach utilizes reads from RNA sequencing datasets that cover both, linear splice junctions and circular branchpoint junctions (Figure 2). The combination of these two split-read types under certain splicing-specific criteria such as branchpoint distance and splice site positioning (see Methods), allows for stringent filtering and strongly reduces mapping artifacts that accompany split-read alignment. This approach additionally maximizes the alignment length and thus allows for the identification of novel splicing events, outside the known intron-exon boundaries.

The datasets we mainly focused our analysis on were generated by library preparation protocols that enriched either for long non-polyadenylated nuclear RNA[19], by size selection and fractionation, or circular RNA by exonuclease treatment or CaptureSeq[16,20]. The former includes, besides other RNA species, precursor- and partially processed mRNA, which sets the foundation for extraction of linear splice junctions. The latter holds, besides circRNAs, reads of lariats, which are the basis for branchpoint split read extraction.

Splicing event classification

Using this approach we identified 97,338 splicing events (neglecting alternative branchpoints), most of which span full introns. 5,693 events are shorter than the respective annotated intron and therefore may originate from recursive or nested splicing. The majority of non-full length splicing events are found in introns shorter than 5000 basepairs (bp) (Figure 3A). Further classification of this dataset subdivided the splicing events into 4 classes:

1) Full length (**FL**): covering a fully annotated intron, allowing for alternative splice sites within 50 nucleotides of the annotated splice site (88,410 events).
2) 5' recursive (**5rec**): the splicing event uses an annotated 5' splice site (or one close to it) but does not extend throughout the entire intron, using an unannotated, intronic 3' splice site, thus being a potential first splicing event of a recursive splicing cascade (2,699).
3) 3' recursive (**3rec**): like 5rec, with the difference, that the 5' splice site is within the intron and the 3' splice site is annotated (2,515).
4) Nested splicing event (**nest**): both, the 5' and the 3' splice sites lie within the intron and neither is annotated (479).

The distribution of all splicing events identified over these four classes is depicted in Figure 3B and C. These numbers do not take into account alternative branchpoints, but show only those events with distinct 5' and 3' splice sites.

In order to keep naming of different types of splicing events consistent, in this paper we use the following designations: **FL**, **5rec**, **3rec** and **nested** as described above; **intrasplicing** refers to a splicing event that does not span the entire annotated intron and can therefore be a recursive or nested splice; **LaSLi** stands for Lariat Spanning Linear split and describes splicing events identified in our extraction pipeline; Splice sites that do not mark an annotated exon-intron border are termed **intra-splice sites**.

Data availability

The full dataset, including splicing event classification, splice site scoring and alternative branchpoint positioning is available in the supplementary materials. Note that circular split read positions, i.e. potential branchpoint positions, are only available in the set *all_laslis_BP_bed8.bed* as block coordinates. Splice site scoring is available as a separate set of files with the name column (4) of the bed format as the 5'ss score and the score column (5) as 3'ss score.

Characteristics of intronic splicing events

In order to gain an overview about the splice site quality of LaSLis, we determined the splice site strength of splice sites of each splicing class with Xmaxentscan[21], a tool that scores a given sequence by the maximum entropy principle and references to a predetermined splice site consensus (**http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html**).

The score distributions for each intrasplicing class as well as annotated splice sites are given in Figure S 1. Average scores for annotated splice sites are 7.84 for 5' and 7.99 for 3'ss. 5rec and 3rec exonic sites are within that range, with the intronic splice site scoring at 4.31 and 5.70, respectively. Both intra-splice sites of nested LaSLis have a much more spread out score distribution with a strong difference between 5’ and 3’ splice sites (Figure S 1C). We used these scores, to adjust splice site definition and overcome the error prone split-read mapping. In this approach we allowed for a maximum adjustment of 2 nucleotides at each splice site and selected the highest scoring position. Especially 3’splice site assignment benefitted from this approach and resulted in a final dataset with much higher splice site assignment fidelity.

Next, we performed a meta-analysis on these 4 classes of splicing events to investigate and compare their properties, with respect to annotated splicing sites. This also serves as a measure of the performance of our extraction pipeline as the unique properties of introns and lariats, such as branchpoint and splice site consensus sequence, branchpoint – 3’ss distance, conservation, etc., are well studied and can be referenced to. The vast majority of splicing events that are in contact with an exon, such as full length, 5rec and 3rec, fit the annotation in terms of splice site position (Figure 4A). Interestingly, those events that show an offset to the annotated splice site follow a pattern of peaks every third nucleotide up- and downstream of the annotated splice site. This is indicative of alternative splice sites, which do not break the reading frame. Notably, alternative 3’ss seem to have a strong preference for a shift of -12 nucleotides into the intron, leading to additional 4 amino acids included upstream of that exon.

We next analyzed the sequences surrounding the splice sites to resolve any deviating splice site motifs for intronic splicing events (Figure 4B). Full length, 5rec and 3rec splice sites showed a strong GT-AG motif, very close to that of annotated splice sites, accounting for the precision of splice event extraction. Nested splicing events deviated strongly from the consensus sequence and reevaluation of the mapping data revealed that a lot of sequences originating from ribosomal RNA were falsely mapped to intronic regions. The abundance of these sequences allowed for a number of reads to evade our filtering criteria and get included in the nested splicing event subset. To remove these false positives, nested splicing events were further filtered by presence of a splice site consensus, i.e. GT at the 5’ and AG at the 3’ end. We have preliminary results, that also showed other, non-GT-AG nested splicing events to be spliced out, but these are subject to further studies and are therefore not included in the present dataset.

The conservation of splice sites was determined by extracting the corresponding genomic positions from the phastCons100 dataset, that calculates element conservation scores for 100 vertebrates[22,23]. Conservation scores of splice sites at intron exon boundaries show an expected drop in conservation towards the intronic region (Figure 4C). Splice sites located within the intron such as 3’ splice sites of 5’ recursive, 5’splice sites of 3’ recursive or both splice sites for nested events have a generally lower conservation, which still exceeds that of intron background and show, for the putative recursive splice sites, a similar conservation pattern as exonic splice sites. The conservation scores drop towards the site of the intrasplicing event (Figure 4D), which could be indicative of cryptic or RS exons. In contrast, nested splice sites display an increased conservation for approximately 10 and 6 nucleotides surrounding the 5’ and 3’ splice site, respectively.

The next characteristic we took a closer look at was branchpoint properties. The distance between the branchpoint and the downstream 3’ splice site is limited due to the structure of the spliceosome[24] and can be used to determine the quality of our extracted splicing events. As shown in Figure 5A, the branchpoint distance peaks between 20 and 30 nucleotides, which concurs with literature[25]. Also the branchpoint sequences mainly follow what is known from literature, namely CUNAN or UUNAN[16]. Nested splicing events deviate from displaying a strong branchpoint adenosine. Deviations from consensus are likely to be introduced by the imprecision of split read mapping. The polypyrimidine tract downstream of the branchpoints is apparent for all splicing event classes (Figure 5C).

The conservation pattern at branchpoints shows an increased score for the branchpoint A and flanking nucleotides as compared to intron or local background (Figure 5B). Nested and 5’ recursive splicing events, and therefore branchpoints of exon free, intronic 3’ splice sites, have a similarly high score, yet do not show the characteristic drop in conservation before and after the branchpoint. Examples of a highly conserved branchpoint and intra-splice site are depicted in Figure 5D.

Confirmation of unusual splicing events by lariatPCR

In order to confirm the existence of the computationally extracted splicing events, we performed lariatPCR on cDNA after RNA extraction from human cells and reverse transcription using random primers. This approach uses diverging primer pairs, binding downstream of the 5’ splice site and upstream of the branchpoint so that a PCR product is only obtained if a circularization event, i.e. lariat formation, has occurred (Figure 6A). LariatPCR further allows to precisely determine the site of the splicing event. We were able to obtain PCR products for the majority of lariats tested (Figure 6B). In many cases nested PCR was required due to the low abundance of lariats in the cells. RNA from HeLa, Hek293, K562 and Hep2G cells was used to increase the chances of detecting the lariats. PCR products were then sanger-sequenced and mapped against the genome to confirm the nature of the lariat and control for the precision of the splicing event extraction pipeline (Figure 6C). In case of the putative 3’recursive splicing event in *prkab2*, intron 7, we could also detect the linear splice junction on the preprocessed pre-mRNA (Figure 6D-F). The transient nature of lariats and splicing intermediates makes this confirmation process tedious, as even annotated, full length introns of highly expressed genes could not always be detected, presumably due to sequence repetitiveness and quick turnover. These properties inherit to and deteriorate in rarer or less abundant (due to low gene expression) splicing events. Thus our rather high confirmation rate (10/14) demonstrates the fidelity and robustness of our splicing event extraction approach.

The impact of intrasplicing on full intron removal in a reporter assay

To investigate the impact of intronic splicing events on full intron splicing efficiency and potentially transcript fate, we constructed a splicing reporter using the renilla luciferase, on the basis of the phRL-TK vector (Promega). Small introns (< 5000nts) were selected that showed at least one putative nested or recursive splicing event in our dataset, and cloned into the renilla gene at a position that favors splice site recognition (see Methods). The intra- or recursive splice sites were mutated or, in case of a 3'splice site, deleted together with the upstream polypyrimidine tract. Further in depth analysis was performed by removing the nested splicing event on DNA level, elucidating the fate of the pre-spliced transcripts. Vice versa, we also generated constructs, containing only the nested splicing event to evaluate their splicing efficiency, allowing us to draw conclusions about intra-splice site strengths (see Figure 7A). Using this approach, we identified three different types of effects of intrasplicing events on full intron splicing.

Intrasplicing competes with full intron splicing suggesting a novel type of gene expression regulation

The two genes *gnb2l1* and *prkab2* both contained putative recursive splicing events in intron 7. Our dataset included a 5'rec and a 3'rec event for *gnb2l1* and a 3'rec event for *prkab2*. Mutation of these intra-splice sites in our reporter construct increased renilla expression and transcription in all three cases (Figure 7B, Figure S 2), meaning that full intron removal is more efficient in the absence of intrasplicing. Thus, these events are unlikely to be part of a recursive splicing cascade. Removal of the 3'end of intron 7 of both, *gnb2l1* and *prkab2*, does not lead to the reconstitution of a new 3'splice site and the polypoyrimidine tract is removed, rendering the remainder of both introns splicing incompetent. The level of increase of renilla expression by 2-fold for 5'ss mutants allows a rough estimate that 50% of transcripts of this reporter, are spliced at the LaSLi. The inefficient downstream processing of pre-spliced introns is confirmed by three more constructs that only contained the 5' or 3' remainder sequence that is left after an initial intra-splicing event. These constructs are unable to express functional luciferase (Figure 7C, "pre"). Contrary, the intrasplicing event of *prkab2*, intron 7 alone is effectively spliced (Figure 7C,"3'nest"). These results are further supported by semi-quantitative RT-PCR (Figure S 2). As these splicing events do not facilitate full intron removal, as the model of recursive splicing would suggest, but rather counteract it, they represent, in concert with a downstream degradation mechanism, a potential new mode of regulation of gene expression.

Silent intrasplicing

We also tested two candidate introns containing putative splicing events at the 5'end, where the pre-splice step does not seem to influence the full intron removal efficiency in the cell lines tested (Figure 7D). This may be either due to this event not taking place in the cell lines we tested, weak and therefor easily outcompeted intra-splice sites or a recursive splicing cascade that is, due to the shortness of the full intron, not a necessity for intron removal.

Intrasplicing boosts full intron removal

A third type of intrasplicing effect was identified in intron 8 of rbm17, which harbours 3 putative intrasplicing events. Mutation of the intra-splice sites, be it 5' or 3'ss, led to a strong reduction of renilla expression, both on transcript and protein level (Figure 8A, B). Therefore, these splicing events might be starting points of a recursive splicing cascade, yet due to their positioning, presumably of three different cascades. This is an interesting case, as intron 8 of rbm17 is only slightly over 1kb long and thus, recursive splicing should not be a necessity for this intron's removal.

More to discover

Besides the above described modes of intrasplicing, we found several other interesting intrasplicing localizations in our dataset: e*pb41l5* expresses several isoforms, two of which are significantly shorter and terminate with an exon that is skipped in the longer isoform. In this terminal intron, we found a recursive splicing event that reconstitutes a RS-site and induces a second splicing step that skips the alternative terminal exon, allowing transcription of the long isoform (Figure 8: rbm17, intron 8 harbours 3 potential intrasplicing events. Mutation or deletion of the intra-splice sites leads to a strong reduction in transcriptional (A, one representative of two replicates shown) and translational (B) output ($p<0.05$ for all constructs in Hek cells and 213 in Hela).

Figure 9A). Attempts to mask this RS-site with antisense oligomers failed to show an isoform shift. Similar has been observed in other studies on recursive splicing, where ASOs were used[9], either rendering this approach inapplicable or hinting to technical difficulties when targeting ASOs to deep intronic sites.

Another example shows two adjacent splicing events in *nudc*. This appears to be a classical recursive splicing situation, yet the intron length is with 2500bp quite short and arguably, would not require recursive splicing as a means of full intron removal.

Discussion

Currently available gene annotations have a very exon-focused view on the transcriptome. Our splicing dataset aims at extending available annotations and promoting a more processing-aware approach to transcription. Generation of mini-genes or mini-introns as well as deep intronic mutations can result in unprocessed or inefficiently transcribed RNAs due to insufficient understanding of intronic elements controlling RNA processing and thus stability. In combination with other available datasets or tools such as eCLIP-seq[26], splicePie[17] or computational prediction of regulatory sequence elements by tools like RBPmap[27], a broader understanding of the processing steps required to achieve regulated gene expression can be obtained.

Current research, dealing with splice junction discovery on a massive scale[28] clearly showed that splicing cascades are by far more complex than current annotations imply. Yet, the extensiveness of this dataset makes it close to impossible to determine biological relevance of individual splice reactions. In our study we attempted, with a more restrictive approach, to extract high confidence events that can be more easily tested and we evaluated their impact on gene expression. Our pipeline utilizes tools designed for the discovery of circular RNAs to supplement the information a regular split read discovery gains on splice site location. Besides increasing confidence in the discovered splice junctions, this also allows for the localization of the branchpoint. Even though we only took a closer look at a handful of unusual splicing events, our results anticipate a much broader range of impacts of intrasplicing events on gene expression. Based on our data and on the diversity of mechanisms we found, we can speculate on the diverse effects of intrasplicing events on full intron splicing. Recursive splicing allows for clean full intron removal. Yet, the necessity of RS-sites for efficient full intron removal remains debated. In some cases, recursive splicing might be a means of processing a mis-spliced intron, to retain transcript functionality. This might be especially true for short introns, where recursive splicing seems not be an approach to spanning long distances between splice sites. Mis-splicing could be promoted by an abundance of U1-binding in many introns, as premature cleavage and polyadenylation (PCPA) inhibitor[7], what might direct the splicing machinery to falsely recognize intronic splice sites and commit to an intrasplicing event.

If the transcript cannot be further processed or processed in a way that does not interrupt the reading frame, situations like in *prkab2*, intron 7 splicing can occur. Here the intrasplicing event effectively removes the functional 3'splice site, branchpoint and polypyrimidine tract, rendering the remainder of the intron retained (Figure 10A). Induction of such a dead-end processing in combination with downstream degradation pathways, such as nonsense-mediated decay, can be a powerful regulatory tool for gene regulation. It also lays the foundation for exon skipping induction, as depicted in Figure 10B. Vice versa, speeding up processing of dwelling introns can have the opposite effect by overcoming nuclear retention and facilitating mRNA export. Even with the precise mechanism unknown, such a situation seems to occur in *rbm17*, intron 8, presumably via recursive splicing. Yet unknown is the role of the exon junction complex (EJC) in recursive and intrasplicing. CliP experiments determining genome-wide EJC deposition[29] clearly show that our intrasplicing events are associated with proximal intronic EJC binding (Figure S 3). The consequences of EJC deposition are broad and range from modulating splicing patterns, transcription speed, mRNA stability to export and translation efficiency. The effect of EJC deposition on introns has not been studied to date but it seems likely that interference with spliceosome assembly would occur. In order to effectively splice the full intron in recursive or intrasplicing patterns, this interference has to be handled. Therefore, further studies are required to elucidate the role of EJCs in recursive and intrasplicing.

Recent studies utilizing nascent RNA sequencing to determine the co-transcriptional nature of splicing have shown that intron removal is in many cases a prompt event and occurs seconds after the splice sites have been transcribed[30,31]. If this observation can be extended to human transcription, it explains high occurrence of intrasplicing we observe, despite partially weak splice site interactions (Figure S 1).

The variety of mechanistical implications makes intrasplicing a promising new regulatory layer of gene expression that adds to the already complex and dynamic regulatory landscape of multi-exon gene regulation. Future studies will focus on the interactors involved in intra-splice site recognition and a potential cell-type specific cocktail of splicing factors that allows fine tuning of the transcriptome via intra-splicing.

Materials and Methods

Cell culture and transfection

HeLa and Hek293T cells have been grown in DMEM medium, supplemented with 10% FCS (Sigma-Aldrich). Cells were transfected with Lipofectamine 3000 (Thermo Fisher) according to the manufacturer's specifications.

RNA preparation, reverse transcription and PCR

RNA was isolated 48 hours post-transfection with Tri-reagent (Sigma-Aldrich), DNAse I (Roche) treated with 2x 20U per 50µg RNA for 30 minutes and then phenol chloroform extracted. Reverse transcription was carried out with superscript III (Invitrogen) on 1 µg total RNA for 60 minutes at 50°C. 1/20th of this reaction was then used as a template for subsequent PCR (oneTaq, NEB) or qPCR (Evagreen, Medibena). Lariat PCR was carried out with diverging primers, spanning the putative branch point. In some cases, nested PCR with cycle numbers ranging from 20 to 30 cycles was necessary to amplify lowly abundant lariats. All primers used for PCR, qPCR and cloning can be found in the supplementary data.

Cloning and site directed mutagenesis

The phRL-TK constructs containing various introns have been generated with the InFusion cloning kit (Clontech) and NEBuilder (NEB). All introns have been inserted at position 414 of the renilla gene. The splice site mutations and nested intron deletions have been introduced with the Q5 site-directed mutagenesis kit (NEB).

Dual luciferase assay

For the luciferase assay, the experimental phRL-TK vector with intron insertions has been cotransfected with pmirGLO -ren as an internal control in 24-well plates. pmirGLO –ren is a modified version of pmirGLO, where the renilla gene and its promoter have been removed by restriction digest and religation. 48 hours after transfection the medium, except 100 µL was removed and the luciferase assay was carried out according to the manufacturer's specifications. Luminescence was measured on a luminometer (Robion Solaris 3170) and relative renilla luminescence was calculated by background subtraction and normalization to the internal firefly control.

Bioinformatic pipeline for splicing event extraction

The focus on pre-mRNA and lariats implied the usage of datasets that are potentially enriched in these RNA species. Therefore the main focus of this approach was on two RNA-seq datasets: nuclear, non-ployadenylated long RNA and RNAse R digested circular RNA. The raw RNA sequencing data of these datasets, together with several others, to extend the read pool (Table 1), were obtained from the gene expression omnibus (https://www.ncbi.nlm.nih.gov/geo) and have been processed by the following pipeline:

a) quality control: adapter clipping, quality trimming and duplicate removal where necessary with fqtrim (http://ccb.jhu.edu/software/fqtrim/index.shtml) and FastUniq (https://sourceforge.net/projects/fastuniq/).
b) Mapping to human reference genome hg38 with segemehl 0.1.9[18]. The segemehl tool has been modified to allow detection of split reads that are up to 1.2 million nucleotides apart, which ensures detection of split reads covering the largest introns in the human genome.
c) For the extraction of split reads, segemehl's realign routine was run. The resulting bedfile *.splice.bed was processed with bedtools[32] and custom scripts, extracting and sorting linear and circular split reads and filtering those out that did not overlap with annotated introns and extended beyond gene boundaries.

d) In order to obtain high confidence splicing events, intronic circular and linear splits were intersected and filtered by following requirements: the 5’ ends of intersecting linear and circular splits cannot deviate more than 2 nucleotides in any direction, to tolerate a given uncertainty in split-read mapping; the 3’ end of the circular split has to be located within 100 nucleotides upstream of the 3’ end of the linear split read to allow for a certain range of branchpoint-splice site distances.

It is not possible to precisely allocate lariats and their respective linear split if, for example an alternative 3’ splice site is used and two branchpoints and therefore 3’ ends of two different circular splits lie within 100 nucleotides upstream of this splice site. This results in a small number of redundant splicing events, varying in BP and 3’ss position.

The resulting splicing events of this extraction procedure were termed lariat-spanning linear splits (LaSLiS) and provide the basis of all subsequent analyses.

Splicing event analyses

LaSLiS were classified based on the position within the hosting intron. LaSLiS with exon contact and therefore utilizing one primary splice site were classified as potentially recursive and LaSLis without any exon contact as intra- or nested splicing events.

Sequences of LaSLiS splice sites (i.e. the 5’ and 3’ ends and surrounding nucleotides) were extracted and motifs were generated with Weblogo 3[33]. As a reference, 5’ and 3’ ends of annotated introns were analyzed with the same procedure.

Conservation scores for seven vertebrates were obtained from the phyloP7 project[34] (http://hgdownload.cse.ucsc.edu/goldenPath/hg38/phyloP7way/) and intersected with the 20 nucleotides surrounding the 5’ and 3’ ends of LaSLiS. The reference is another window of 20 nucleotides from within every intron (>500 nts) in the human genome.

Splice site scoring was achieved with Xmaxentscore[21]. Respective sequences were extracted and used as input. If, due to poor split-read mapping, a higher splice site score was found within +/- 2 nts of the 5’ or 3’ end of any LaSli, the higher scoring position was assumed as the proper splice site. This was done to increase fidelity in splice site annotation. Both, the original and the optimized dataset are available.

Table 1: datasets accessible via GEO or SRA, used for splicing event extraction.

Dataset	GEO/SRA accession	Library preparation	reference
12tissues	GSE45326	total RNA	Nielsen et al., 2014 [35]
Illumina body map 2.0	GSE30611	total RNA	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM759490
Brain single cell	GSE67835	total RNA, single cell	Darmanis et al., 2015 [36]
circSeq	SRP011042	total RNA, RNAse R	Jeck et al., 2013 [20]
polyA minus	GSE22666	polyA minus RNA	Yang et al., 2011[19]
BPseq	GSE53328	lariatSequencing + BPcapture	Mercer et al., 2015[16]

References

1. Roy, B., Haupt, L. M. & Griffiths, L. R. Review: Alternative Splicing (AS) of Genes As An Approach for Generating Protein Complexity. *Curr. Genomics* **14,** 182–94 (2013).

2. Nellore, A., Jaffe, A. E., Fortin, J.-P., Alquicira-Hernández, J., Collado-Torres, L., Wang, S., Phillips III, R. A., Karbhari, N., Hansen, K. D., Langmead, B. & Leek, J. T. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biol.* **17,** 266 (2016).

3. Ott, S., TAMADA, Y., Bannai, H., Nakai, K. & MIYANO, S. INTRASPLICING-ANALYSIS OF LONG INTRON SEQUENCES. *Pacific Symp. …* **350,** 339–350 (2003).

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

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

6. Shepard, S., McCreary, M. & Fedorov, A. The peculiarities of large intron splicing in animals. *PLoS One* **4,** (2009).

7. Berg, M. G., Singh, L. N., Younis, I., Liu, Q., Pinto, A. M., Kaida, D., Zhang, Z., Cho, S., Sherrill-Mix, S., Wan, L. & Dreyfuss, G. U1 snRNP determines mRNA length and regulates isoform expression. *Cell* **150,** 53–64 (2012).

8. Parra, M. K., Tan, J. S., Mohandas, N. & Conboy, J. G. Intrasplicing coordinates alternative first exons with alternative splicing in the protein 4.1R gene. *EMBO J.* **27,** 122–31 (2008).

9. Sibley, C. R., Emmett, W., Blazquez, L., Faro, A., Haberman, N., Briese, M., Trabzuni, D., Ryten, M., Weale, M. E., Hardy, J., Modic, M., Curk, T., Wilson, S. W., Plagnol, V. & Ule, J. Recursive splicing in long vertebrate genes. *Nature* (2015). doi:10.1038/nature14466

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

11. Masaki, S., Yoshimoto, R., Kaida, D., Hata, A., Satoh, T., Ohno, M. & Kataoka, N. Identification of the specific interactors of the human lariat RNA debranching enzyme 1 protein. *Int. J. Mol. Sci.* **16,** 3705–21 (2015).

12. Hesselberth, J. R. Lives that introns lead after splicing. *Wiley Interdiscip. Rev. RNA* **4,** 677–91 (2013).

13. Coombes, C. E. & Boeke, J. D. An evaluation of detection methods for large lariat RNAs. *RNA* **11,** 323–31 (2005).

14. Awan, A. R., Manfredo, A. & Pleiss, J. A. Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 12762–7 (2013).

15. Bitton, D. A., Rallis, C., Jeffares, D. C., Smith, G. C., Chen, Y. Y., Codlin, S., Marguerat, S. & Bahler, J. LaSSO, a strategy for genome-wide mapping of intronic lariats and branch-points using RNA-seq. *Genome Res.* (2014). doi:10.1101/gr.166819.113

16. Mercer, T. R., Clark, M. B., Andersen, S. B., Brunck, M. E., Haerty, W., Crawford, J., Taft, R. J., Nielsen, L. K., Dinger, M. E. & Mattick, J. S. Genome-wide discovery of human splicing branchpoints. *Genome Res.* gr.182899.114 (2015). doi:10.1101/gr.182899.114

17. Pulyakhina, I., Gazzoli, I., Hoen, P.-B. t., Verwey, N., den Dunnen, J., Aartsma-Rus, A. & Laros, J. SplicePie: a novel analytical approach for the detection of alternative, non-sequential and recursive splicing. *Nucleic Acids Res.* gkv242- (2015). doi:10.1093/nar/gkv242

18. Hoffmann, S., Otto, C., Doose, G., Tanzer, A., Langenberger, D., Christ, S., Kunz, M., Holdt, L. M., Teupser, D., Hackermüller, J. & Stadler, P. F. A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biol.* **15,** R34 (2014).

19. Yang, L., Duff, M. O., Graveley, B. R., Carmichael, G. G. & Chen, L.-L. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol.* **12,** R16 (2011).

20. Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., Marzluff, W. F. & Sharpless, N. E. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19,** 141–57 (2013).

21. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. in *Proceedings of the seventh annual international conference on Computational molecular biology - RECOMB '03* 322–331 (ACM Press, 2003). doi:10.1145/640075.640118

22. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20,** 110–121 (2010).

23. Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W. & Haussler, D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).

24. Luukkonen, B. G. M. The role of branchpoint-3' splice site spacing and interaction between intron terminal nucleotides in 3' splice site selection in Saccharomyces cerevisiae. *EMBO J.* **16,** 779–792 (1997).

25. Taggart, A. J., DeSimone, A. M., Shih, J. S., Filloux, M. E. & Fairbrother, W. G. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nat. Struct. Mol. Biol.* **19,**

719–21 (2012).

26. Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., Blue, S. M., Nguyen, T. B., Surka, C., Elkins, K., Stanton, R., Rigo, F., Guttman, M. & Yeo, G. W. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **13,** 508–514 (2016).

27. Paz, I., Kosti, I., Ares, M., Cline, M. & Mandel-Gutfreund, Y. RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Res.* **42,** W361–W367 (2014).

28. Nellore, A., Jaffe, A. E., Fortin, J.-P., Alquicira-Hernández, J., Collado-Torres, L., Wang, S., Phillips III, R. A., Karbhari, N., Hansen, K. D., Langmead, B. & Leek, J. T. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biol.* **17,** 266 (2016).

29. Saulière, J., Murigneux, V., Wang, Z., Marquenet, E., Barbosa, I., Le Tonquèze, O., Audic, Y., Paillard, L., Roest Crollius, H. & Le Hir, H. CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. TL - 19. *Nat. Struct. Mol. Biol.* **19 VN-r,** 1124–1131 (2012).

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

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

32. Quinlan, A. R., Quinlan & R., A. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. in *Current Protocols in Bioinformatics* 11.12.1-11.12.34 (John Wiley & Sons, Inc., 2014). doi:10.1002/0471250953.bi1112s47

33. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–90 (2004).

34. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20,** 110–21 (2010).

35. Nielsen, M. M., Tehler, D., Vang, S., Sudzina, F., Hedegaard, J., Nordentoft, I., Orntoft, T. F., Lund, A. H. & Pedersen, J. S. Identification of expressed and conserved human noncoding RNAs. *RNA* **20,** 236–251 (2014).

36. Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., Hayden Gephart, M. G., Barres, B. A. & Quake, S. R. A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 7285–90 (2015).

Figure Legends

Figure 1: Proposed splicing models for long intron splicing. Recursive splicing (top) uses intronic recursive splice sites (RS-sites) that reconstitute a new 5’ or 3’ splice site after a first splicing step, depending on the directionality of the recursive splicing cascade. Nested splicing (bottom) uses two intronic splice sites that remove part of the intron before the final splicing step removes the entire intron.

Figure 2: Split-reads that cover exon-exon or branchpoint-5’splice site junctions are extracted from segemehl-aligned RNA-seq datasets. Split reads originating from the same splicing event (exon-exon junction reads and lariat reads) can be combined in order to obtain splice site and branchpoint positions (see Materials and Methods). This allows for a higher confidence mapping of splice sites than previous approaches which rely solely on linear split reads.

Figure 3: Overview of the extracted splicing events. (A) shows the length of the extracted splicing event plotted against the length of the corresponding intron. This plot is shown as a density plot to allow for visualization of overlapping splicing events, grading the number of splicing events with similar properties by the colour grade shown in the legend. The majority of splicing events, both, full length and shorter are found in introns < 5000 nts. In (B) the same splicing events are classified based on their position in the corresponding intron: full length (blue), putative initial 5’ recursive (green), 3’ recursive (red) and nested splicing events (yellow). The abundance of each class of splicing event is shown in (C).

Figure 4: Splice site analysis of the 4 classes of splicing events. (A) shows the distance of the newly identified splice sites to the closest annotated splice site. (B) Splice site motifs of the four splicing classes as compared to the reference motif of 10,000 annotated introns. (C) shows the phastCons100 score across the splice sites. Intronic splice sites are generally less conserved and yet show a specific increase in conservation just around the splice site, as can be seen in (D), which is a close up of the indicated area of (C).

Figure 5: Analysis of branchpoint properties. (A) distance between the branchpoint and the respective 3‘ splice site plotted against the fraction of events identified in that respective class. (B) branch point motif of the 4 LaSLi classes. (C) branchpoint conservation with phastCons100 score. (D) shows an example of a highly conserved 3‘ splice site and branchpoint of a 5rec LaSLi in the *HSPG2* gene and a 5‘ splice site of a 3rec LaSLi in *CARM1*. The bold bar indicates the circular, the thinner bar the linear split junction, corresponding to branchpoint and 3‘ or 5‘ splice site, respectively.

Figure 6: LariatPCR confirms the computationally extracted splicing events. (A) in order to obtain lariat specific PCR products, diverging primers were designed. Only in the case of a circularization during splicing and lariat formation, a PCR product will be obtained. (B) lariat specific PCR products obtained in a nested PCR. (C) larPCR products were sanger-sequenced, the obtained sequences were aligned to the human genome with BLAT and compared to the LaSLi dataset. Green arrows indicate the primer positions used for the larPCR. In few cases, linear splice junctions of nested splicing events could also be detected. (D) shows the primer positions in intron 7 and exon 8 of *prkab2*, (E) shows the pcr product obtained over the splice junction, compared to genomic DNA and (E) visualizes the mapping of the sequenced PCR product.

Figure 7: (A) Overview of mutational analysis of intrasplicing event containing introns cloned into the renilla luciferase gene. Putative 3'recursive splice events had their 5'ss mutated and, in some cases, the intron was sliced into the fractions that contain only the intrasplicing event and those that contain the remainder of the intron. These were termed pre-spliced and 3'/5' lasli, respectively. In case of a 5' intra splicing event, the 3' splice site was rendered dysfunctional by deleting the polypyrimidine tract and the downstream AG of the 3'intrasplicesite. (B) Intron 7 of gnb2l1 contains two intrasplicing events. Mutation of the respective intronic splice site leads to an increase in renilla expression (left panels, $p<0.05$). Similar effects are observed for intron 7 of prkab2, which contains one putative 3'recursive splice event ($p<0.05$). (C) Prespliced constructs of gnb2l1, intron 7 and prkab2, intron 7 are splicing incompetent and do not produce a functional renilla luciferase ($p<0.05$). The intronic splicing event of prkab2, intron 7 on its own is well capable of splicing, though apparently with a lower efficiency than the full intron. (D) Example of two genes, rad54L and megf6, where the mutation of the intra-splice sites did not affect full intron removal efficiency ($p>0.1$).

Figure 8: rbm17, intron 8 harbours 3 potential intrasplicing events. Mutation or deletion of the intra-splice sites leads to a strong reduction in transcriptional (A, one representative of two replicates shown) and translational (B) output ($p<0.05$ for all constructs in Hek cells and 213 in Hela).

Figure 9: UCSC genome browser screenshots of splicing events. (A) *epb41L5* holds a recursive splicing event in a terminal intron, that induces terminal exon skipping and thus a shift towards expression of a longer isoform. (B) shows a classical recursive splicing situation in a notably short intron (2500bp).

Figure 10: Two models illustrating the potential impact intrasplicing events can have on gene expression, via regulation of transcript abundance and isoform expression by inducing exon skipping events. (A) In absence of a functional intra-splice site, the intron is efficiently spliced. An intrasplicing event on the other hand removes the py-tract and does not re-establish and RS-site. As a result, the

remainder of the intron is retained, leading to nuclear retention or NMD. (B) If a downstream intron is present, such an intrasplicing event can induce exon skipping by utilizing the next functional 3’ss.

Figure S 1: Maximum entropy score distribution for each given class of splice sites. The computed score is given on the x-axis and the number of splicing events in each given subset is shown on the y-axis. (D) shows the scores of annotated splice sites and is used as a reference set.

Figure S 2: RT-PCR of prkab2 and gnb2l1 intron 7 in phRL-TK including splice site mutations. This semi-quantitative PCR reproduces the result of the renilla assay and sheds additional light on RNA processing. A mutation of the 5’ss of prkab2, intron 7 leads to alternative splice site selection within the renilla exon. This splice site is also used when only the 3’intrasplicing event is cloned into the reporter. Here about 50% of transcripts use the native, the rest the exonic splice site. This reflects the reduced intra-splice site strength. The longer product of the pre-spliced transcript arises due to the incomplete processing and splicing incompetence.

Figure S 3: Exon junction clip data from Saulière et al. (2012)[29] shows clear deposition of exon junction complexes in genomic regions flanking intrasplicing events (blue boxes).

Figure 1

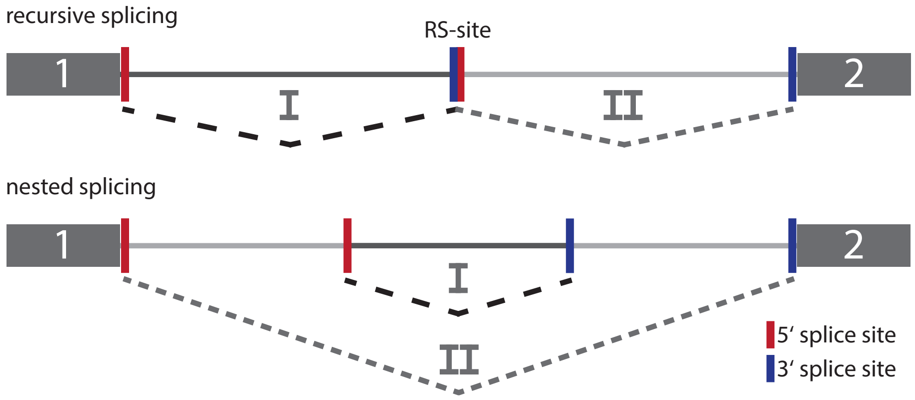

recursive splicing
RS-site
1
2
I
II
nested splicing
1
2
I
II
5′ splice site
3′ splice site

Figure 2

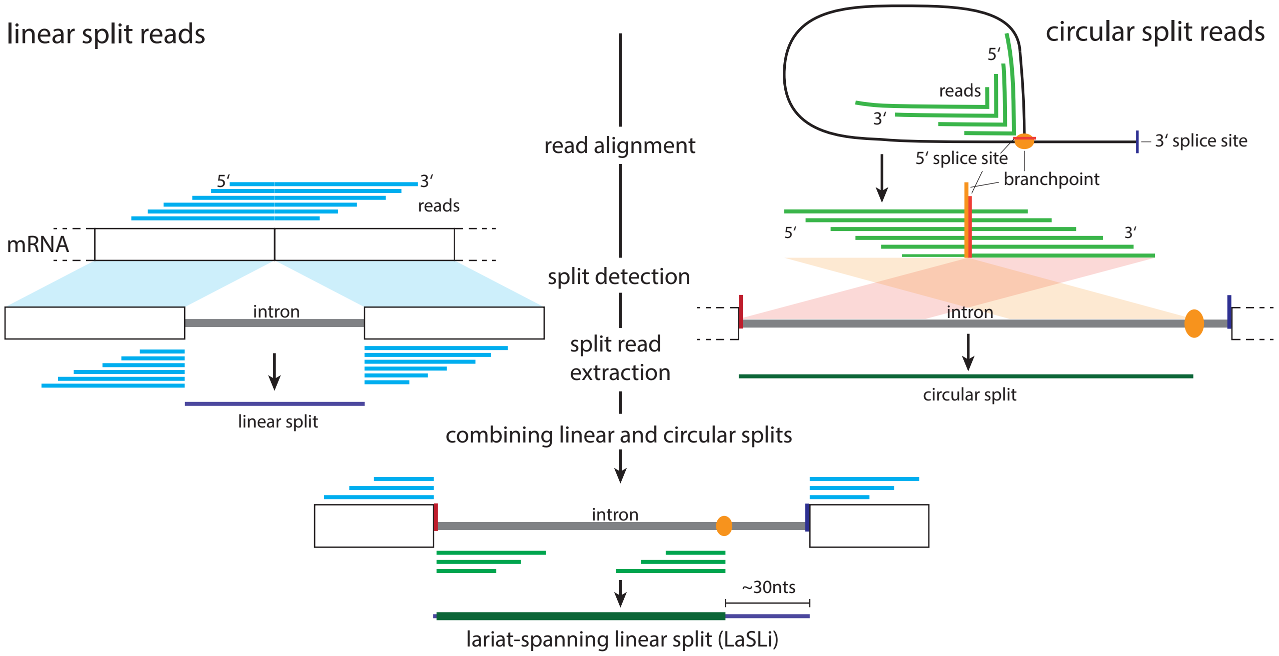
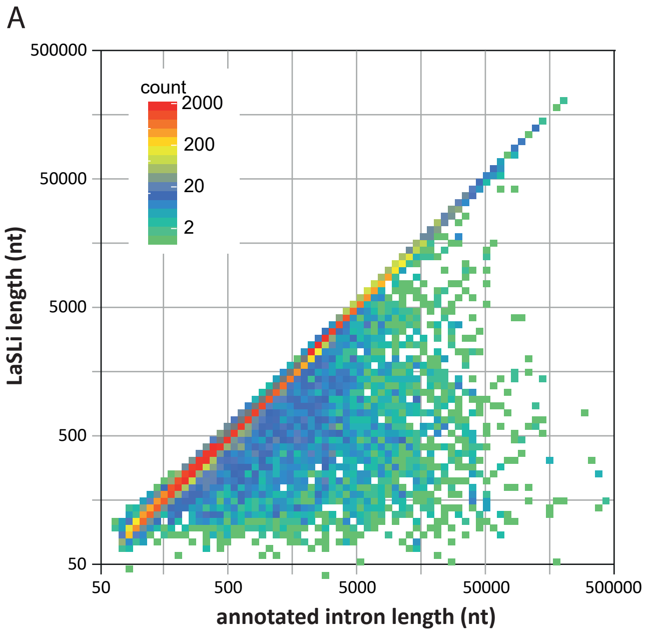

linear split reads
read alignment
circular split reads
5′
3′
reads
mRNA
intron
linear split
split detection
split read
extraction
5′
reads
3′
5′ splice site
branchpoint
3′ splice site
5′
3′
intron
circular split
combining linear and circular splits
intron
~30nts
lariat-spanning linear split (LaSLi)

Figure 3


A
count
2000
200
20
2
500000
50000
5000
500
50
LaSLi length (nt)
50
500
5000
50000
500000
annotated intron length (nt)

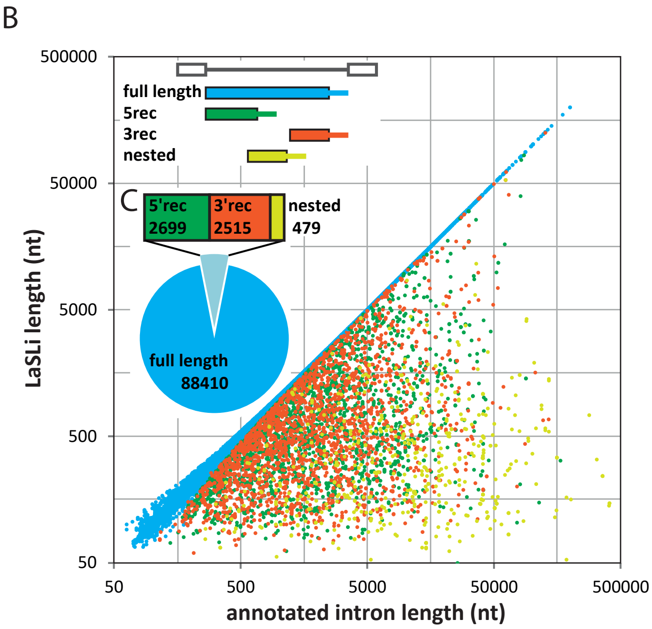

B
full length
5rec
3rec
nested
C
5'rec
2699
3'rec
2515
nested
479
full length
88410
500000
50000
5000
500
50
LaSLi length (nt)
50
500
5000
50000
500000
annotated intron length (nt)

Figure 4

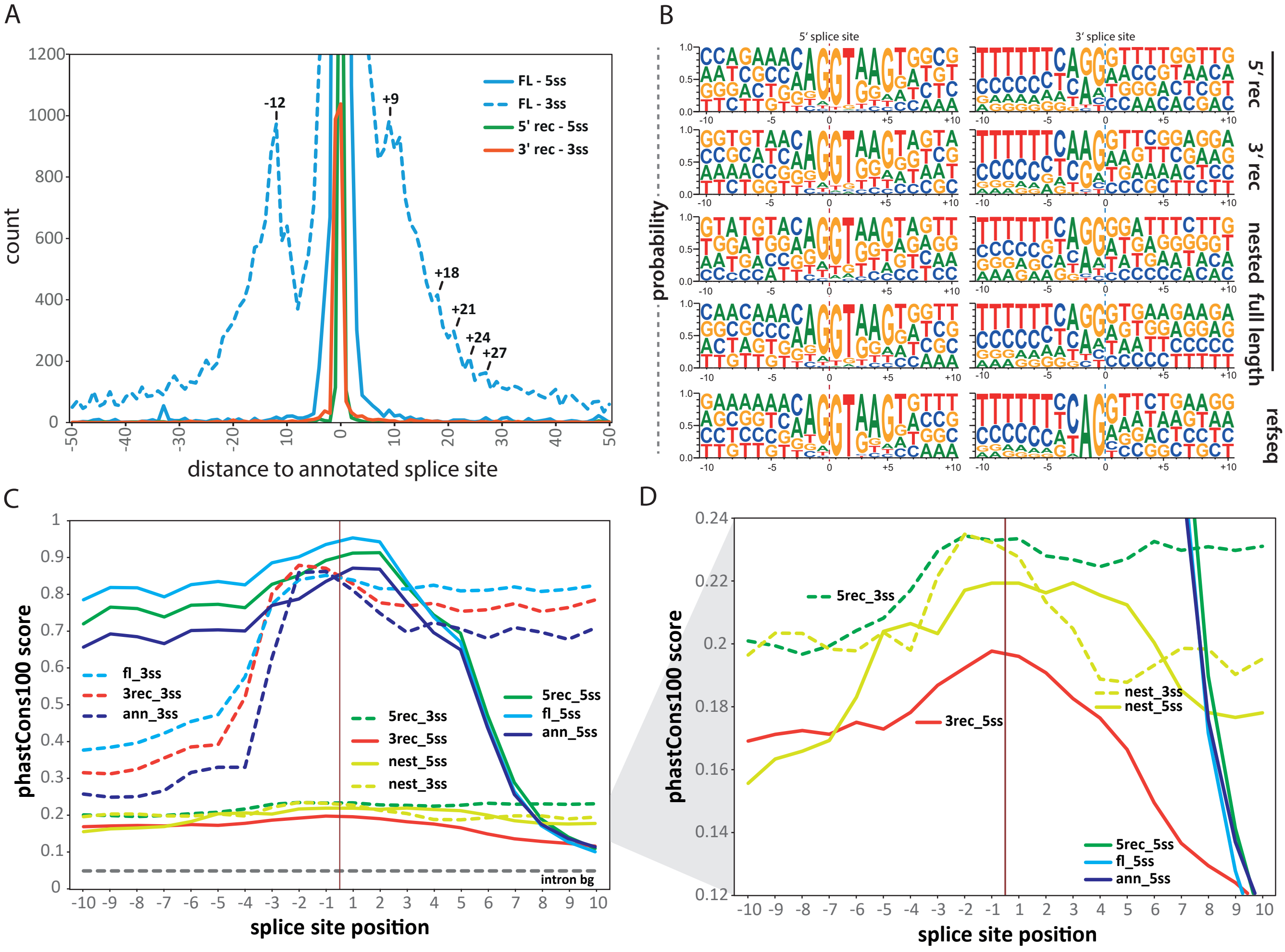

A
count
distance to annotated splice site
FL - 5ss
FL - 3ss
5' rec - 5ss
3' rec - 3ss
-12
+9
+18
+21
+24
+27
1200
1000
800
600
400
200
0
-50
-40
-30
-20
-10
0
10
20
30
40
50
B
5' splice site
3' splice site
probability
5' rec
3' rec
nested
full length
refseq
LASLIs
C
phastCons100 score
splice site position
fl_3ss
3rec_3ss
ann_3ss
5rec_3ss
3rec_5ss
nest_5ss
nest_3ss
5rec_5ss
fl_5ss
ann_5ss
intron bg
D
phastCons100 score
splice site position
5rec_3ss
3rec_5ss
nest_3ss
nest_5ss
5rec_5ss
fl_5ss
ann_5ss

Figure 5

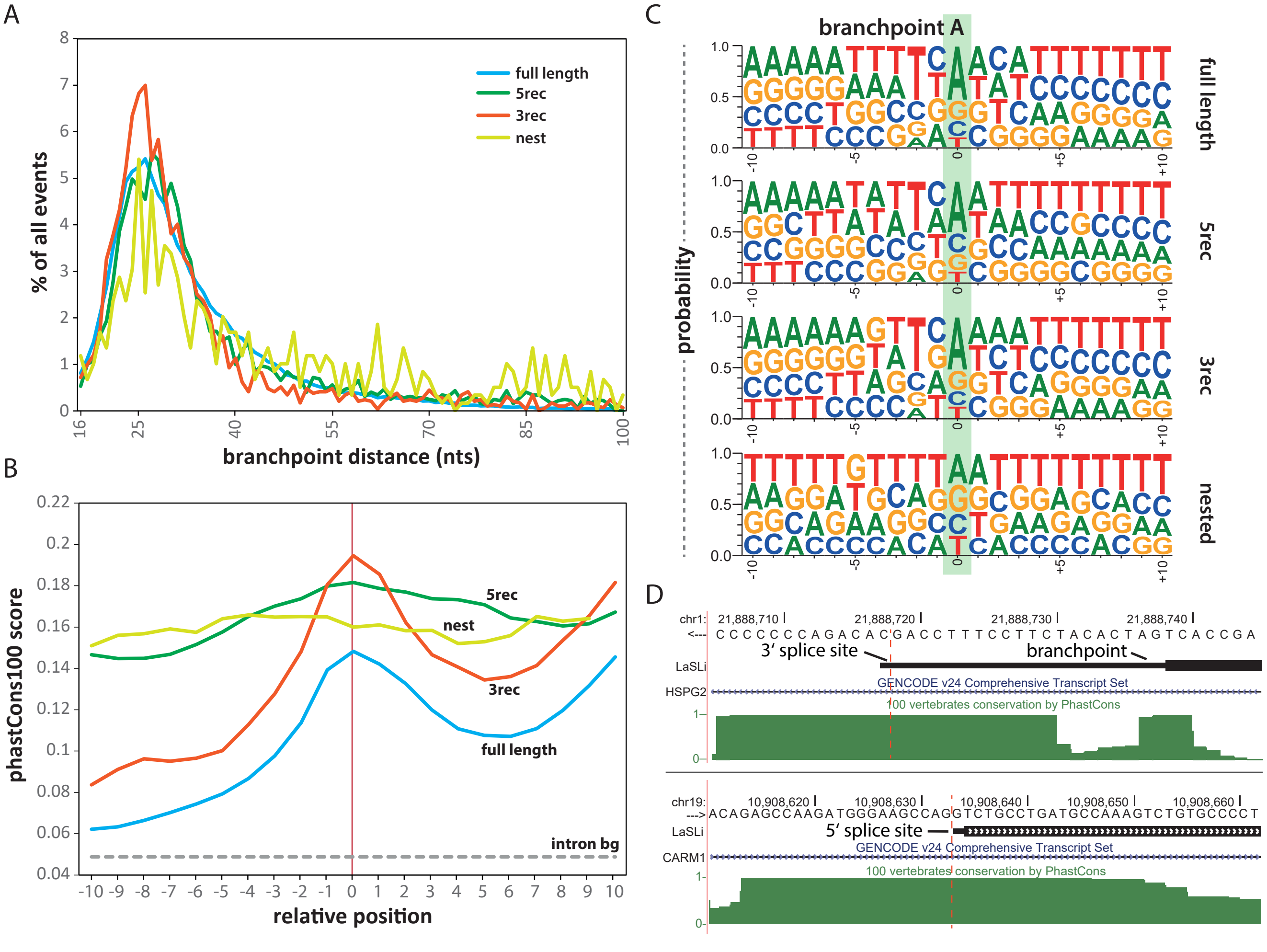

A
% of all events
branchpoint distance (nts)
0
1
2
3
4
5
6
7
8
16
25
40
55
70
85
100
full length
5rec
3rec
nest
B
phastCons100 score
relative position
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0.2
0.22
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10
5rec
nest
3rec
full length
intron bg
C
branchpoint A
probability
0.0
0.5
1.0
-10
-5
0
+5
+10
full length
5rec
3rec
nested
D
chr1:
21,888,710
21,888,720
21,888,730
21,888,740
<--- C C C C C C C A G A C A C G A C C T T T C C T T C T A C A C T A G T C A C C G A
3' splice site
branchpoint
LaSLi
HSPG2
GENCODE v24 Comprehensive Transcript Set
100 vertebrates conservation by PhastCons
1
0
chr19:
10,908,620
10,908,630
10,908,640
10,908,650
10,908,660
---> ACAGAGCCAAGATGGGAAGCCAGGTCTGCCTGATGCCAAAGTCTGTGCCCCT
5' splice site
LaSLi
GENCODE v24 Comprehensive Transcript Set
CARM1
100 vertebrates conservation by PhastCons
1
0

Figure 6

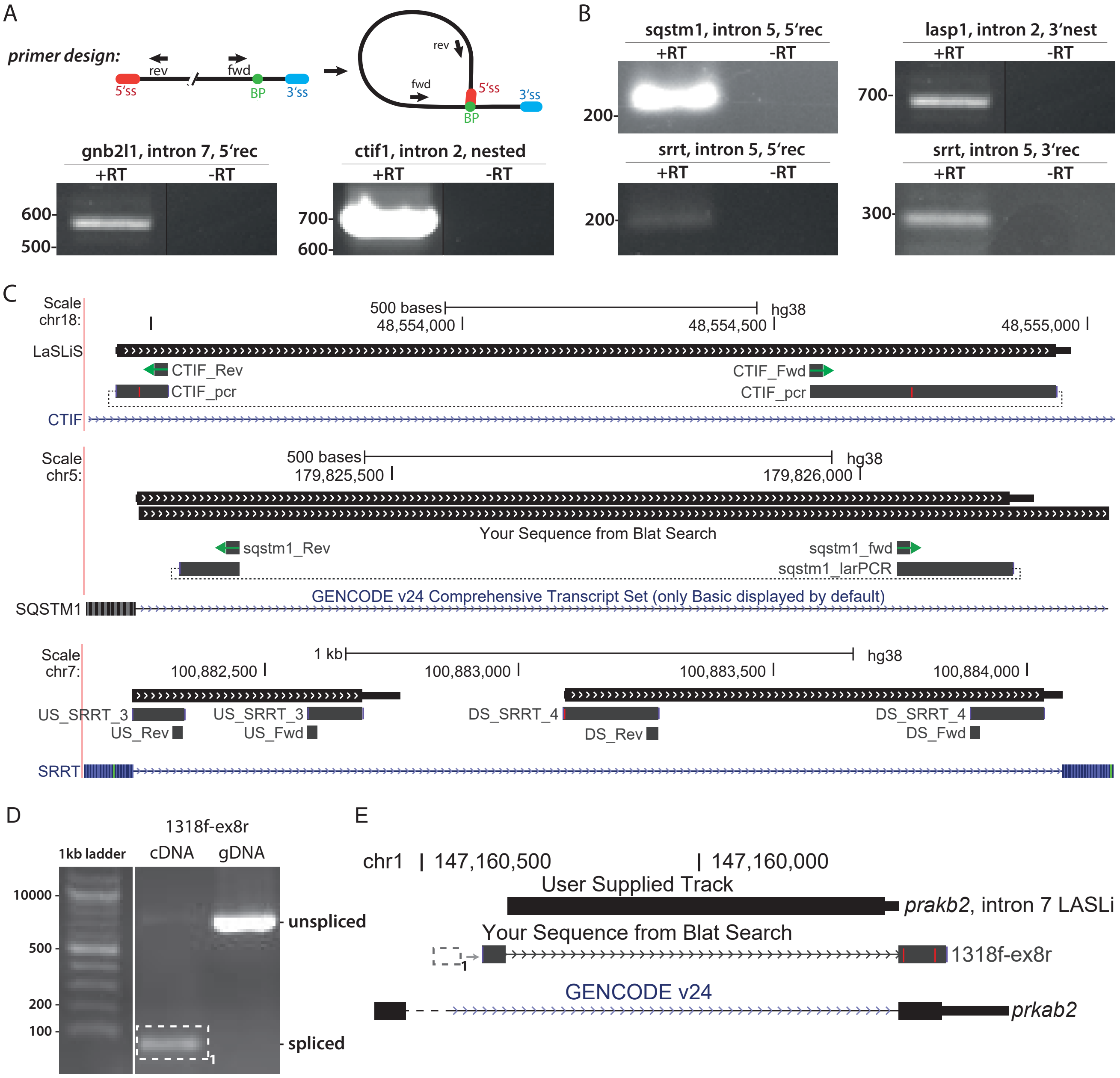
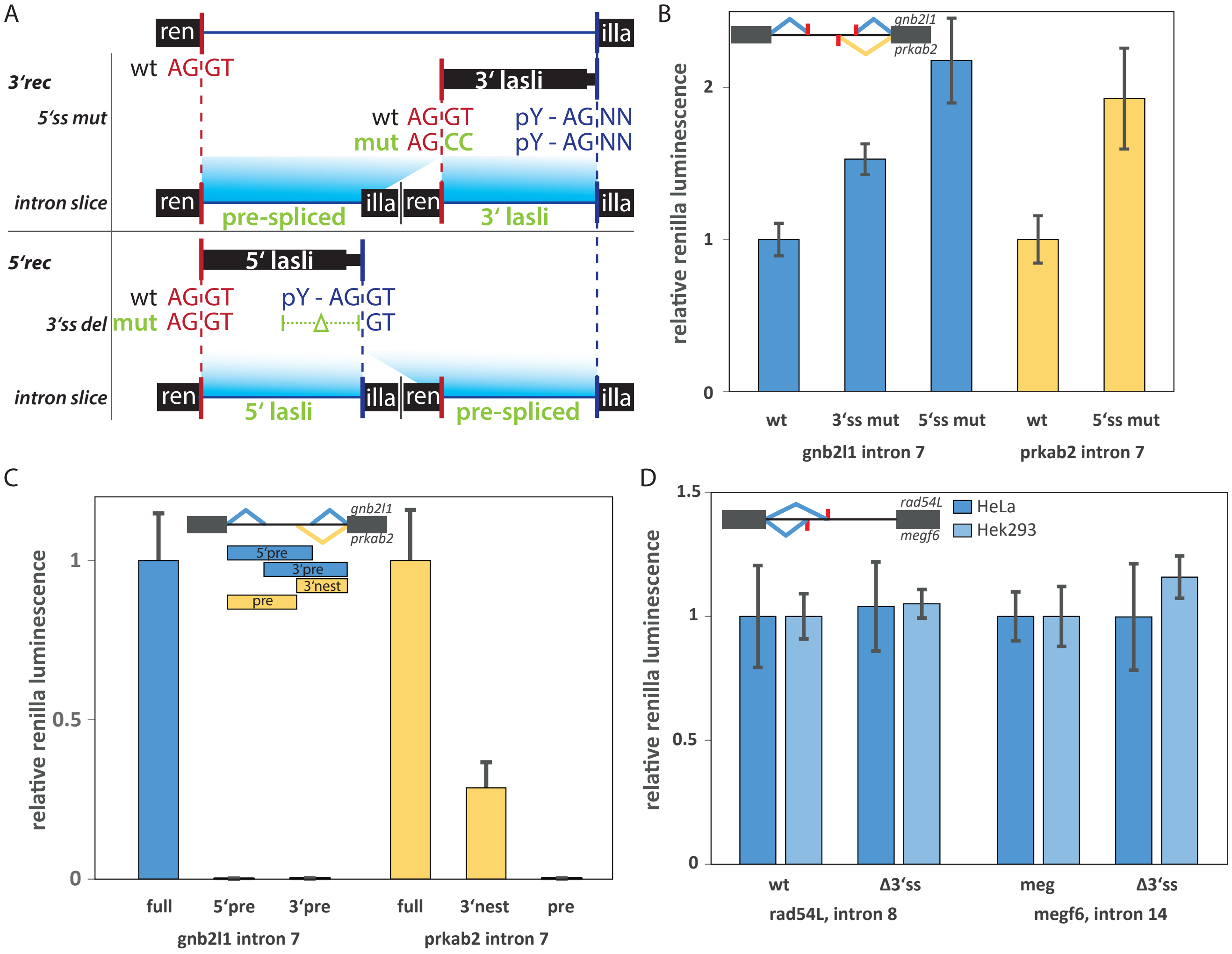

A
primer design:
rev
fwd
5'ss
BP
3'ss
rev
fwd
5'ss
3'ss
BP
gnb2l1, intron 7, 5'rec
+RT
-RT
600-
500-
ctif1, intron 2, nested
+RT
-RT
700-
600-
B
sqstm1, intron 5, 5'rec
+RT
-RT
200-
lasp1, intron 2, 3'nest
+RT
-RT
700-
srrt, intron 5, 5'rec
+RT
-RT
200-
srrt, intron 5, 3'rec
+RT
-RT
300-
C
Scale
chr18:
500 bases
hg38
48,554,000
48,554,500
48,555,000
LaSLiS
CTIF_Rev
CTIF_Fwd
CTIF_pcr
CTIF_pcr
CTIF
Scale
chr5:
500 bases
hg38
179,825,500
179,826,000
Your Sequence from Blat Search
sqstm1_Rev
sqstm1_fwd
sqstm1_larPCR
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
SQSTM1
Scale
chr7:
1 kb
hg38
100,882,500
100,883,000
100,883,500
100,884,000
US_SRRT_3
US_SRRT_3
DS_SRRT_4
DS_SRRT_4
US_Rev
US_Fwd
DS_Rev
DS_Fwd
SRRT
D
1318f-ex8r
1kb ladder
cDNA
gDNA
10000
500
200
100
unspliced
spliced
1
E
chr1
147,160,500
147,160,000
User Supplied Track
prakb2, intron 7 LASLi
Your Sequence from Blat Search
1318f-ex8r
1
GENCODE v24
prkab2

Figure 7



A
ren
illa
3‘rec
wt AG GT
3‘ lasli
5‘ss mut
wt AG GT
pY - AG NN
mut AG CC
pY - AG NN
intron slice
ren
pre-spliced
illa
ren
3‘ lasli
illa
5‘rec
5‘ lasli
wt AG GT
pY - AG GT
3‘ss del
mut AG GT
Δ
GT
intron slice
ren
5‘ lasli
illa
ren
pre-spliced
illa
B
gnb2l1
prkab2
relative renilla luminescence
0
1
2
wt
3‘ss mut
5‘ss mut
wt
5‘ss mut
gnb2l1 intron 7
prkab2 intron 7
C
gnb2l1
prkab2
5‘pre
3‘pre
3‘nest
pre
relative renilla luminescence
0
0.5
1
full
5‘pre
3‘pre
full
3‘nest
pre
gnb2l1 intron 7
prkab2 intron 7
D
rad54L
megf6
HeLa
Hek293
relative renilla luminescence
0
0.5
1
1.5
wt
Δ3‘ss
meg
Δ3‘ss
rad54L, intron 8
megf6, intron 14

Figure 8

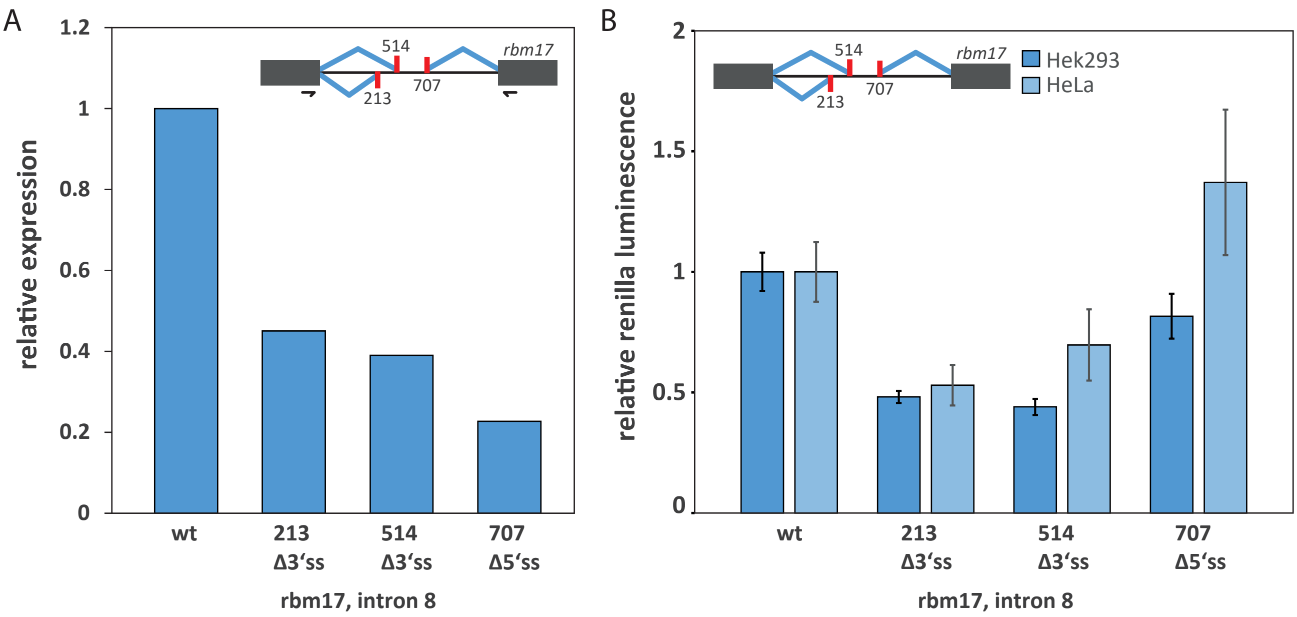

A
1.2
1
0.8
0.6
0.4
0.2
0
relative expression
514
213
707
rbm17
wt
213 Δ3'ss
514 Δ3'ss
707 Δ5'ss
rbm17, intron 8
B
2
1.5
1
0.5
0
relative renilla luminescence
514
213
707
rbm17
Hek293
HeLa
wt
213 Δ3'ss
514 Δ3'ss
707 Δ5'ss
rbm17, intron 8

Figure 9

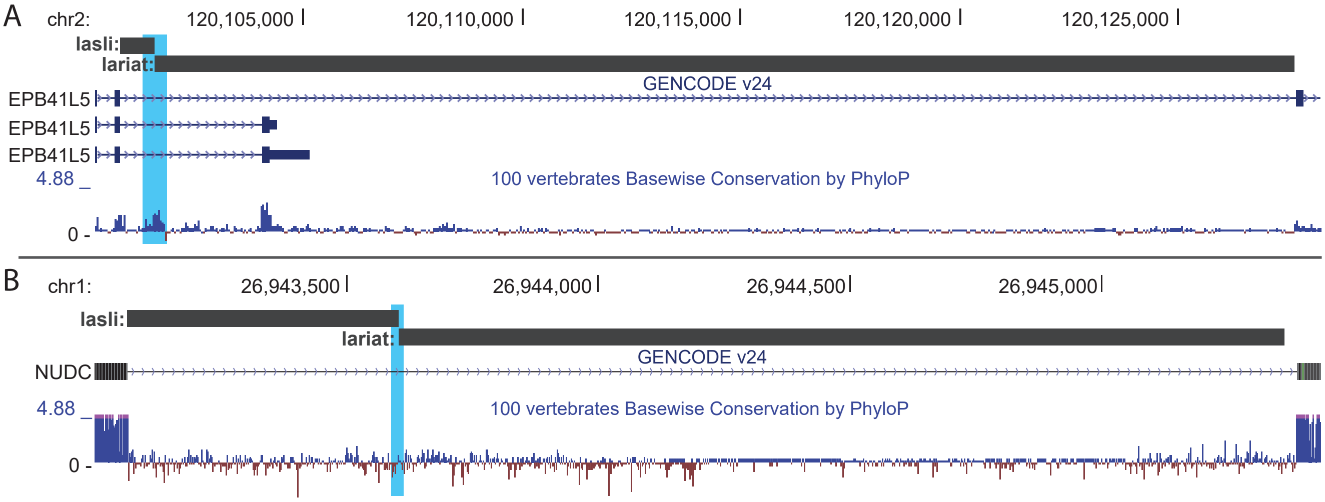

A
chr2:
120,105,000
120,110,000
120,115,000
120,120,000
120,125,000
lasli:
lariat:
GENCODE v24
EPB41L5
EPB41L5
EPB41L5
4.88
100 vertebrates Basewise Conservation by PhyloP
0
B
chr1:
26,943,500
26,944,000
26,944,500
26,945,000
lasli:
lariat:
GENCODE v24
NUDC
4.88
100 vertebrates Basewise Conservation by PhyloP
0

Figure 11

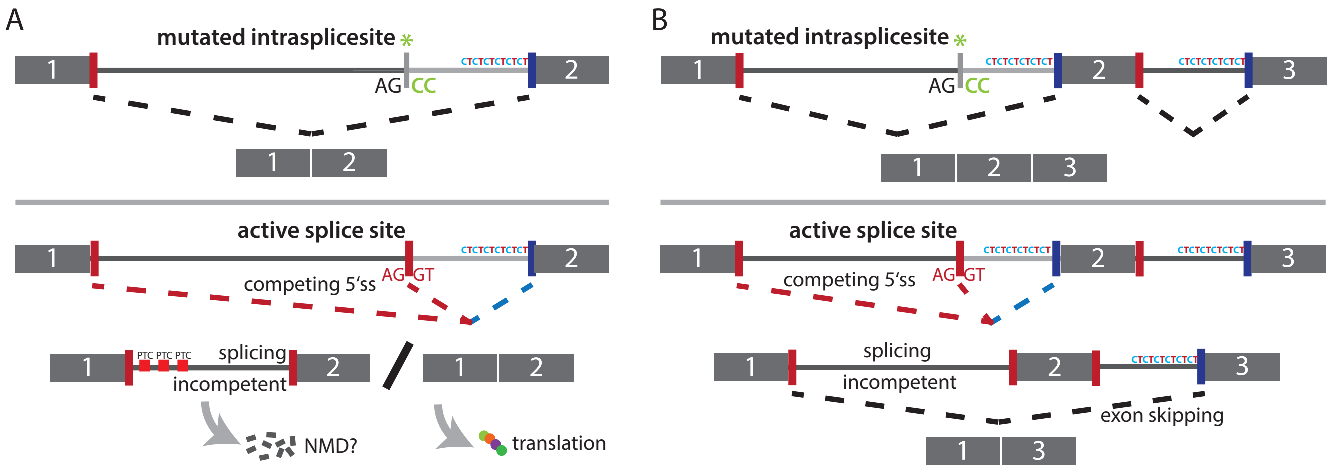

A
mutated intrasplicesite
*
1
AG
CC
CTCTCTCTCTCT
2
1
2
active splice site
1
CTCTCTCTCTCT
2
competing 5'ss
AG
GT
1
PTC PTC PTC
splicing
incompetent
2
1
2
NMD?
translation
B
mutated intrasplicesite
*
1
AG
CC
CTCTCTCTCTCT
2
CTCTCTCTCTCT
3
1
2
3
active splice site
1
CTCTCTCTCTCT
2
CTCTCTCTCTCT
3
competing 5'ss
AG
GT
1
splicing
incompetent
2
CTCTCTCTCTCT
3
exon skipping
1
3

Figure S 1

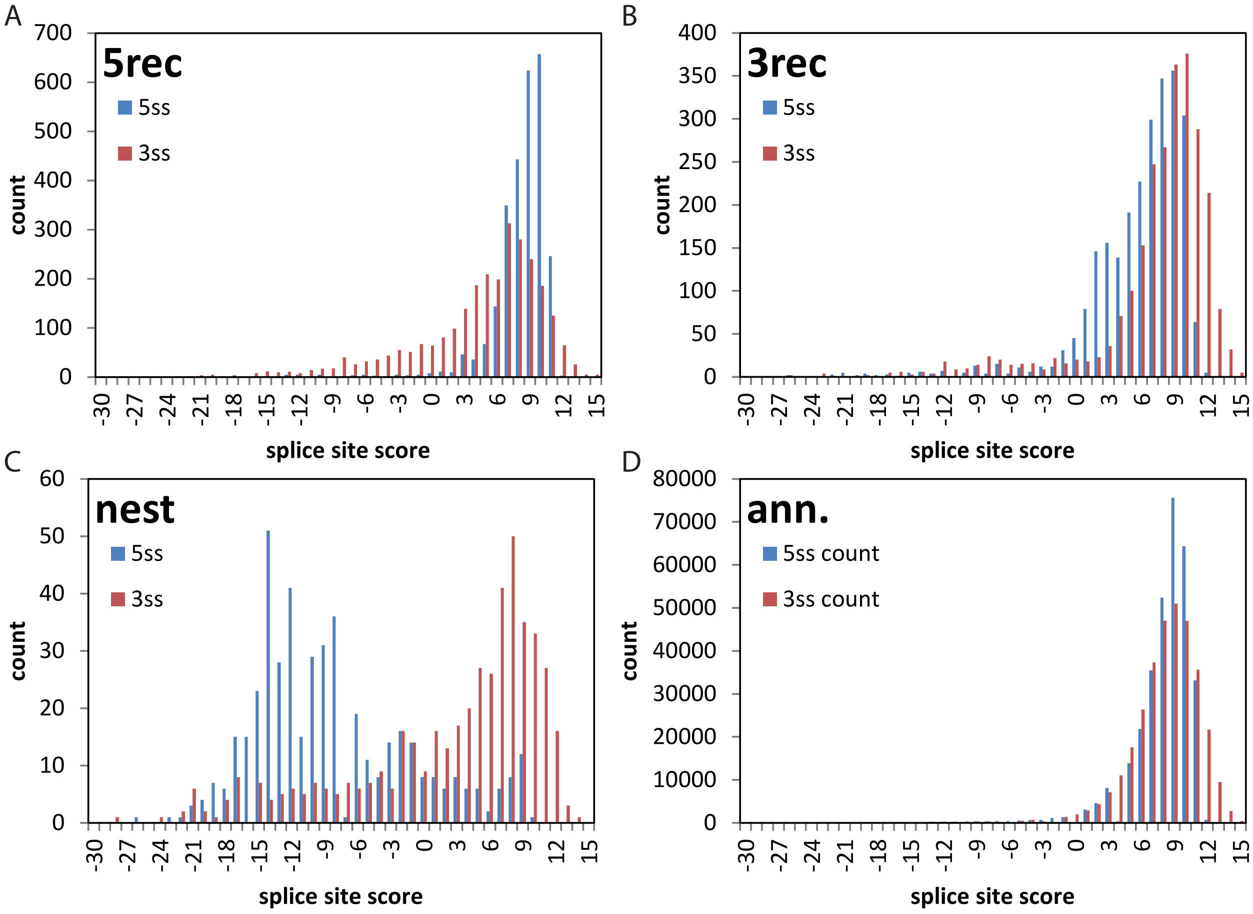

A
5rec
■ 5ss
■ 3ss
count
0
100
200
300
400
500
600
700
-30
-27
-24
-21
-18
-15
-12
-9
-6
-3
0
3
6
9
12
15
splice site score
B
3rec
■ 5ss
■ 3ss
count
0
50
100
150
200
250
300
350
400
-30
-27
-24
-21
-18
-15
-12
-9
-6
-3
0
3
6
9
12
15
splice site score
C
nest
■ 5ss
■ 3ss
count
0
10
20
30
40
50
60
-30
-27
-24
-21
-18
-15
-12
-9
-6
-3
0
3
6
9
12
15
splice site score
D
ann.
■ 5ss count
■ 3ss count
count
0
10000
20000
30000
40000
50000
60000
70000
80000
-30
-27
-24
-21
-18
-15
-12
-9
-6
-3
0
3
6
9
12
15
splice site score

Figure S 2

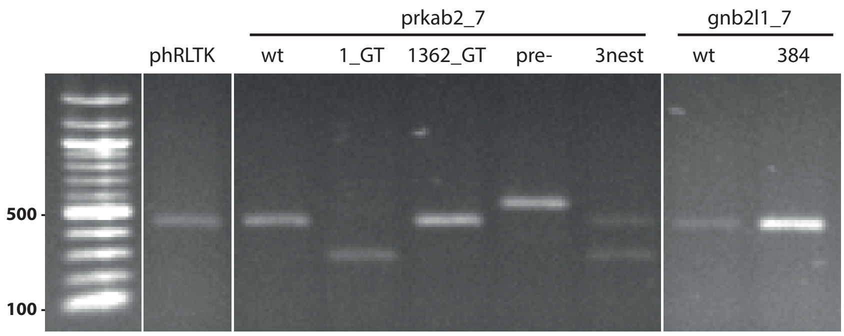
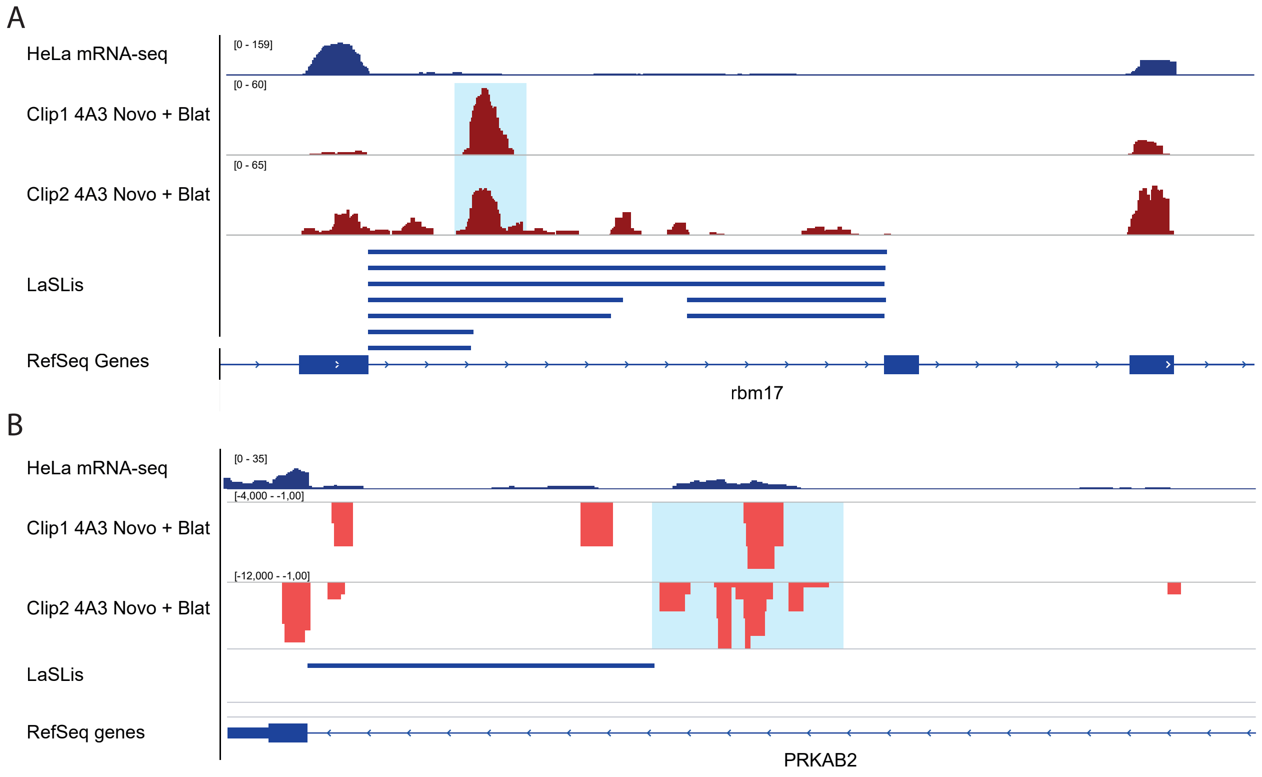

prkab2_7
gnb2l1_7
phRLTK
wt
1_GT
1362_GT
pre-
3nest
wt
384
500
100

Figure S 3


A
HeLa mRNA-seq
[0 - 159]
Clip1 4A3 Novo + Blat
[0 - 60]
Clip2 4A3 Novo + Blat
[0 - 65]
LaSLis
RefSeq Genes
rbm17
B
HeLa mRNA-seq
[0 - 35]
Clip1 4A3 Novo + Blat
[-4,000 - -1,00]
Clip2 4A3 Novo + Blat
[-12,000 - -1,00]
LaSLis
RefSeq genes
PRKAB2

Human α satellite transcripts are substrates for RNA Pol II and contain remnants of snoRNAs

Katarzyna Matylla-Kulinska[1], Hakim Tafer[1], Maximilian Radtke[1], Bob Zimmermann[1], Jennifer L.Boots[1] and Renée Schroeder[1,*]

[1]Department of Biochemistry and Cell biology, Max F. Perutz Laboratories, University of Vienna,

*Corresponding author: Department of Biochemistry and Cell biology, Max F. Perutz Laboratories, University of Vienna, Dr. Bohrgasse 9/5 1030 Vienna, Austria. Tel: +431427754690; Fax: +43142779528; E-mail: renee.schroeder@univie.ac.at

running title: Function and origin of human alpha satellite RNAs

Article character count: 6.219 words

Abstract

Alpha satellites belong to repeats that cover human centromeres and surrounding pericentromeric heterochromatin and play an important biological function in chromosomes segregation. Here, we show that α satellites are transcribed from both DNA strands into long transcripts of at least 8 Kb. α satellite expression is more pronounced under cellular stress conditions and is confined to the S phase of the cell cycle. Satellite transcription is sensitive to α amanitin. α satellite RNAs (αsatRNAs) are capped but not polyadenylated and are retained in the nucleus. Aptamers derived from α satellites were isolated via genomic SELEX with RNA Pol II as bait, indicating that αsatRNAs interact with RNA Pol II. We further show that αsatRNAs bound to the active site of RNA Pol II are extended and/or transcribed resulting in 3' labeling and/or second strand synthesis. Bioinformatic prediction revealed that αsatRNAs contain structural features of H/ACA snoRNAs, hinting at their origin. The phylogeny of αsatRNAs may serve as an example of how repeat-derived repRNAs may constitute a reservoir for the evolution of novel functional RNAs.

Keywords: alpha satellites/ genomic aptamers/ repeat-derived RNA (repRNA)/ RNA-dependent RNA synthesis/ RNA Pol II/ snoRNA

Introduction

The human genome project resulted in an unexpected picture of the genomic content with less than 2 % being protein coding and almost two thirds belonging to repeat elements (Djebali *et al*, 2012; Prasanth & Spector, 2007; Mattick, 2005). These repeat elements have mainly been considered inert and therefore were often referred to as junk. Compared to protein- and RNA-coding regions, the analysis of highly repetitive parts of the genome is still lagging behind mainly due to technical obstacles. As a consequence repeats are most often excluded from experimental design and masked in the analysis.

Interest in repeat elements was boosted after the ENCODE project consortium reported that 60-70 % of the DNA was transcribed into RNA (Djebali *et al*, 2012). Transcripts derived from highly repetitive regions were reported to be involved in genome evolution (Bennetzen, 1996), gene expression (Weil & Martienssen, 2008; Feschotte, 2008; Lai *et al*, 2005) and may serve as a reservoir for the evolution of novel functional RNAs (Gao & Voytas, 2005). Especially interesting is the fact that many of these repeats might have derived from functional RNAs retrotransposed and amplified into genomes.

Human α satellites belong to repeat elements that consist of 171 bp monomers arranged tandemly into higher order arrays spanning hundreds of kilobases to megabases. α satellites localize predominantly to centromeres which often are poorly annotated due to their repetitive character. Although centromere function is highly conserved, centromere DNA sequences do not show conservation between species. Human centromeres contain α satellites, which are primate specific. However, no DNA sequences have been shown to be either sufficient or necessary for centromere assembly, α satellites included. Instead, studies on neocentromeres highlighted that epigenetic signals are crucial for centromere specification (Amor & Choo, 2002) and suggested that this process is highly context-dependent (Hayden *et al*, 2013). Centromeric α satellite arrays are bound by CENP-A (histone H3-like, centromere protein A) and are a site for spindle attachment assuring proper chromosome segregation process during mitosis and meiosis (Verdaasdonk & Bloom, 2012). Repetitive arrays are mainly enclosed within constitutive heterochromatin domains and therefore to date have been considered to be transcriptionally silent. However, it was recently shown that transcription does occur in the centromeric regions and that RNA is present at the mitotic kinetochore (Wong *et al*, 2007). Active transcription and an essential role of RNA Pol II at the centromere during mitosis are currently being discussed (Chan *et al*, 2012).

In a genomic SELEX experiment using RNA Pol II as bait (Boots et al., in preparation) we isolated RNA aptamers derived from human α satellites, which suggests that α satellite transcripts may have a function at the RNA level. This prompted us to analyse the transcriptional output derived from α satellites and to characterize their transcripts. Furthermore, we show that αsatRNAs harbour RNA Pol II binding elements that interact with the active site of the enzyme leading to 3' elongation of the RNA and/or second strand synthesis via RNA-dependent RNA polymerase activity (RdRP). Finally, we report that human α satellites contain hallmarks of snoRNAs suggesting that αsatRNAs, like other repeats, may derive from non-coding RNAs.

Results

Long transcripts arise from both DNA strands of α satellite arrays

Before analysing α satellite-derived transcripts, we reannotated the α satellite repeat sequences using the dfamscan.pl script (Wheeler *et al*, 2013) and found 44058 genomic loci matching α satellite sequences, which are clustered into higher order repeats covering around 0.1 % of the human genome. α satellite monomers show single nucleotide variation of 20-40 % (Wayel & Willard, 1987) and are organized in a head-to-tail manner.

Transcription from human α satellite arrays was studied in HeLa cells. To assess whether α satellite arrays are transcribed, strand specific RT-PCR was primarily chosen due to its sensitivity. Since low abundance of transcripts is to be expected, a degenerated "consensus" primer pair was used, designed on the consensus α satellite sequence reported in (Prosser *et al*, 1986). This consensus primer pair recognizes different α satellite species allowing their detection *en masse.* To avoid unspecific amplifications, PCR products were subsequently cloned, sequenced and mapped to the hg19 reference genome to confirm their α satellite origin (Supplementary Table S1). Consensus primers amplify not only one α satellite monomer (171 bp), but also hybridize into flanking units resulting in a typical ladder-like pattern of amplification products as schematically shown in Figure 1A. Using strand-specific RT-PCR in the presence of radioactively labelled [α-^{32}P] dGTP, α satellites derived from both DNA strands were amplified (Figure 1B). Radioactive PCR products enable sensitive detection with significant reduction of number of amplification cycles (from over 35 down to 18) lowering the probability of unspecific amplification. α satellite transcripts in a direct orientation (D-αsatRNAs) to the consensus unit were reverse transcribed with the reverse primer (Rev), whereas those in reverse complement (RC-αsatRNAs) orientation to the consensus with the forward primer (Fwd). To ensure strand specificity RT reaction without a primer was carried out (-) to exclude RNA snapping back upon itself to serve as a template.

To estimate the size of α satellite transcripts Northern blot analyses was performed with probes derived from the consensus sequence. These probes hybridize to approximately 1700 of the annotated α satellite genomic loci with 100 % identity, again enabling detection of αsatRNAs *en masse*. Northern blots were performed on total RNA isolated from HeLa cells grown at the following conditions: i) 37 °C, ii) heat shocked at 45 °C for 30 minutes and iii) from heat shocked HeLa cells recovered at 37 °C for 60 minutes. The Northern blot experiments reveal that αsatRNAs are transcribed into large products of more than 8 Kb from both DNA strands (Figure 1C). Moreover, comparison of the band intensity between samples from control, heat shocked and recovered conditions, suggests that αsatRNA accumulate upon cellular stress, in this case heat shock. Furthermore, the analysis of αsatRNA expression in synchronized HeLa cells showed that their transcription is more pronounced in the S phase of the cell cycle (Supplementary Figure S1).

αsatRNAs localize to the nucleus

To assess the cellular localization of αsatRNAs, total RNA from HeLa was separated into nuclear and cytosol fractions using sucrose gradient centrifugation. Fractionated RNA was then analysed via Northern blot with LNA-modified consensus probes in D and RC orientations. As shown in Figure 2A, high molecular weight bands are detected in the nuclear fraction demonstrating that αsatRNAs localize to the nucleus. These findings were further confirmed via RT-PCR analysis (Figure 2B). To control the fractionation accuracy *Kcnqt1ot1*, which is a nuclear lncRNA, was detected in the nuclear fraction; *Gapdh* was detected either as a spliced transcript in the cytoplasm or together with the unspliced one in the nucleus.

αsatRNAs are RNA Pol II transcripts, capped but not polyadenylated

In order to investigate which polymerase is engaged in transcribing α satellite arrays, we analyzed αsatRNAs levels in the presence of α amanitin at a 20 µg/ml concentration, which specifically blocks RNA Pol II transcription (Bortolin-Cavaillé *et al*, 2009). As presented in Figure 3A, 6 hours post α amanitin treatment the levels of α satellite transcripts stayed constant. After 24 hours D-αsatRNAs were hardly detectable, whereas the levels of RC-αsatRNAs reduced dramatically after 48 hours. In parallel, the expression rate of well characterized RNA Pol II and RNA Pol III products were monitored as controls. As a positive control, the level of RNA Pol II-transcribed *Gapdh* was analyzed and shown to reduce after 48 hours. As expected, the amount of 5S RNA, a RNA Pol III product, remained unchanged.

To assess whether αsatRNAs contain poly(A) tails, pull down experiments with biotinylated oligo(dT) probes were performed and the presence of αsatRNA either in the pulled down (PD), or flow through (FT) fractions was tested. As shown in Figure 3B, αsatRNAs derived from both DNA strands were more enriched in the FT, implying that they are devoid of poly(A) tails. Moreover, a bioinformatic analysis of α satellite arrays could not detect any canonical polyadenylation signals. The low amount of αsatRNA detected in the PD fraction might be due to the fact that α satellites are A-rich resulting in some affinity to oligo(dT).

The nature of the 5' terminus of αsatRNAs was determined by 5' adaptor ligation reaction. As shown on the schematic representation of the experimental set up (Figure 3C), the ability of RNA to ligate the 5' adaptor strictly depends on the presence of the cap structure. RT-PCR analysis on CIP/TAP treated ligation reactions, followed by cloning and sequencing of PCR products collected from the gel, revealed α satellite-derived products only in samples treated with TAP (Figure 3D). This data suggests that α satellite transcripts possess a 5' cap structure.

Taken together our results demonstrate that αsatRNAs are atypical RNA Pol II transcripts. Like most RNA Pol II products they are sensitive to α amanitin and they possess a cap structure at the 5' terminus; but they are devoid of a poly(A) tail. The lack of a poly(A) tail commonly correlates with decreased RNA stability. However, an estimated half-life of over 12 hours can be calculated from the α amanitin experiments. In addition, the lack of poly(A) tail may cause their nuclear retention.

To assess the expression levels of the αsatRNAs relative to other RNA families, deep sequenced nuclear RNA libraries from THP1 cells (Taft *et al*, 2010) and 5-8F cells (Liao *et al*, 2010) were analysed. For each considered RNA family: i.e. snoRNA, snRNA, 10 nuclear lnRNAs (KCNQ1OT1, NEAT1, MALAT1, HOTAIR, MIAT, SRA1, AIRN, HOTTIP, NRON and XIST) and αsatRNA, the total number of overlapping reads, the total number of reads versus the number of annotation elements in each family as well as the total number of reads versus the total number of nucleotides of the RNA family was compared. This approach allows a rough estimate of the order of αsatRNAs expression. Depending on the dataset, there are 25 to 1000 times more reads overlapping with snRNAs than with αsatRNAs, 60 to 540 times more reads overlapping with snoRNAs than with αsatRNAs and 23 to 70 times more reads overlapping with annotated nuclear long non-coding RNAs than on αsatRNAs (Supplementary Table S2A).The low expression levels of αsatRNAs may suggest a local function of those transcripts, i.e. in changing the chromatin state of the loci of their origin.

RNA Pol II-binding aptamers derived from α satellites

In a previous Genomic SELEX experiment aimed at isolating regulatory RNAs with high affinity to RNA Pol II (Boots et al. unpublished results), several human α satellite-derived aptamers were identified (Supplementary Table S3). After seven rounds of selection and amplification, the enriched pool was sent to deep sequencing and the obtained reads were mapped to the hg19 human genome. The reads that mapped to α satellite regions were further analysed for this study in order to shed light on a potential activity of α satellites-derived transcripts. The selected aptamers were aligned to the α satellite consensus sequence to identify the RNA Pol II binding platform (Figure 4**)**. Interestingly, aptamers against RNA Pol II map to both strands of α satellites. Most of the aptamers span the junction of α satellite units from position 163 (or minus 9) on the upstream unit to position 6 on the downstream unit for the aptamers derived from the D-strand and from position 158 (minus 14) to position 8 on the RC-strand.

To isolate a sequence motif among RNA Pol II aptamers within α satellite regions the MEME suite (Bailey & Elkan C, 1994) of tools was used. For both D- and

RC-αsatRNAs the best motif was selected (Figure 4). Interestingly, both motifs show an overrepresentation of G/A on their 5' and T/C on their 3' half. The ability of endogenous αsatRNAs to bind RNA Pol II was verified by mobility shift electrophoresis and by Co-IP using an antibody against RNA Pol II subunit 1 (Supplementary Figure S2). The presence of RNA Pol II-binding aptamers encoded within α satellites points to a transcription linked activity within α satellite transcripts. While several RNA Pol II-binding aptamers disrupt transcription *in cis,* the αsatRNA aptamers did not affect their own transcription (Boots el al, unpublished).

αsatRNAs interact with the active site of RNA Pol II resulting in second strand synthesis via RNA dependent RNA polymerase activity in vitro and in vivo and in 3' extension

We show that α satellites are transcribed in D- and RC-orientation, and that they contain RNA Pol II-binding aptamers. Considering that RNA Pol II harbours RNA-dependent RNA polymerase (RdRP) activity (Filipovska & Konarska, 2000), we tested whether αsatRNA is used as a substrate by RNA Pol II in the cellular nucleus (Figure 5, Supplementary Figure S3). For this purpose, chimeric RNA templates were designed for nucleofection into HeLa cells that consist of a αsatRNAs-derived RNA Pol II aptamer (αsatPBE ♯111) and a short artificial non-human sequence. The artificial non-human sequence allows the discrimination between *de novo* synthesised and endogenous αsatRNAs. After nucleofection of this chimeric RNA into HeLa cell nuclei and incubation for 24 hours, total RNA was extracted and analysed via strand-specific RT-PCR. A clear RdRP product was detected via strand-specific RT-PCR in the total RNA isolated from transfected cells, which was not present in the input chimeric RNA used for the nucleofection (Figure 5A). When incubating the same chimeric RNA in HeLa nuclear extract in the presence of [α-^{32}P] rNTPs labelled products of the size of the RNA template can be observed. The reaction was inhibited by the adenosine analogue DRB (5,6-dichloro-1-β-D-ribofuranosylbenzimidazole), an inhibitor of elongating RNA Pol II (Figure 5B). These data suggest that RNA Pol II binds the αsatRNAs-derived aptamer (αsatPBE ♯111) to modify it.

To investigate whether the αsatRNA aptamer sequence serves as a specific target for RdRP activity, the short artificial non-human fragment was shuffled from the

5' end of the template to its 3' terminus. As presented in the Figure 5C, incubation of both chimeric RNA templates in the HeLa nuclear extract in the presence of [α-^{32}P] rNTPs resulted in radioactively labelled transcripts that differ in size, depending on the position of the RNA Pol II aptamer on the RNA template. As schematically explained in the Figure 5D, RNA Pol II recognizes the binding platform in the αsatRNA aptamer and synthesizes the second strand. Polymerase omits the non-human artificial sequence when it is positioned at the 3' end of the RNA template and therefore labelled transcripts correspond solely to the size of the αsatRNA fragment (70 nucleotides). When the non-human sequence is located upstream to the αsatRNA aptamer, the newly synthesized strand corresponds by size to the entire template (85 nucleotides).

We further observed that the incubation of αsatRNA in HeLa nuclear extract with [α-^{32}P] rUTPs led to extension of the αsatRNA by a few residues at the 3' end (Figure 6). *In vitro* transcribed αsatRNA was incubated in HeLa nuclear extract, then RNA was isolated and a reverse RNase protection assay (RPA) was performed (Figure 6A). The isolated RNAs were hybridized with a probe complementary to the template RNA and subjected to RNases A/T1 degradation. The fragment protected by the RPA probe was shorter than the input RNA (Figure 6B). The designed probe hybridizes completely to the αsatRNA; therefore A/T1 RNases cleave the single stranded 3' overhang, which was detected as a shift in size of the RNA.

Taken together our results demonstrate that αsatRNAs interact with the active site of RNA Pol II serving as template for a 3' end extension of the RNA and/or second strand synthesis via RNA-dependent RNA polymerase activity.

αsatRNAs contain remnants of snoRNAs

It has been suggested that many non-coding RNAs might be remnants from the RNA world (Brosius & Tiedge, 1996; Brosius, 1999, 2003). Alu elements originating from the 7SL RNA (Ullu & Tschudi, 1984), and several SINEs, like the BC1 RNA, derived from tRNAs (DeChiara & Brosius, 1987) are good examples. In order to characterize a potential ancestor sequence of αsatRNA we looked for hallmarks of functional RNAs within α satellite transcripts by *in silico* structural and sequence analysis. For each Dfam (Wheeler *et al*, 2013) α satellite family, the corresponding consensus structure was derived by clustalw (Larkin *et al*, 2007) aligning the seed sequences used to generate the hidden markov models in Dfam. The alignments were subsequently folded with RNAalifold. α satellites families alrA and alrB showed the highest degree of structuredness, both harbouring a conserved stem located in the first 65 nucleotides. In order to reduce the effect of the large variability within α satellite sequences on the structure computation, the consensus structure was recomputed by aligning only the consensus sequences of all three α satellite families. The resulting consensus structure exhibited two stems separated by an unstructured region containing an H-Box sequence and an ACA-Box found directly downstream of the 3' stem - a structure similar to that of H/ACA-snoRNAs (Figure 7A).

In order to study phylogeny of α satellites we searched for αsatRNAs homologs in primates' genomes. Dfam α satellite seed sequences were blasted against primates' genomes in the reverse order of the phylogenetic tree. Firstly, human α satellites were blasted against the chimpanzee genome. All sequences found to be homologous were than added to the query sequences and used to screen the next closest genome, in this case of gorilla. The same procedure was repeated for genomes of gorilla, orang-utan, gibbon, macacca, phillipine tarsier, grey mouse lemur and greater galago. This strategy allowed us to identify alphoid sequences up to the common marmoset, but not earlier in the evolution, as previously reported (Shepelev *et al*, 2009).

In the marmoset genome, a total of 30 alphoid sequences were identified, however no intersection between human α satellites sequences and the marmoset EST was found. Those alphoid sequences were mapped to 17 distinct loci in the marmoset genome, 5 of them mapped to three different intronic regions of coding genes. Interestingly, the consensus structure of those sequence alignments folds into a structure similar to the hairpin-hinge-hairpin-tail structure of H/ACA snoRNAs (Figure 7B). Putative targets for the consensus sequence of those 5 intronic alphoids were searched using RNAsnoop (Tafer *et al*, 2010) (Figure 7C). Three significant hits were obtained: two located on the 28S rRNA and one the U5 snRNA. Still none of the predicted sites are reported as pseudourydilated, indicating that marmoset homologs to αsatRNAs might not be functional snoRNAs. It should be noted that due to the lack of complete annotation of marmoset rRNA sequences, the human rRNA sequences were taken instead. In order to test for significance of the targets, the α satellites consensus sequence used to find the target was dinucleotide shuffled 5000 times.

The resulting score distribution was then compared to the score of the unshuffled stems to the targets. Only 4 % of the hits from the shuffled sequences scored higher than the best interaction computed for the consensus sequence showing that the isolated targets are significant.

We provide phylogenetic sequence and structure evolutionary evidence that αsatRNAs derive from snoRNAs. These observations would support the Brosius hypothesis of the origin of repeat elements being remnants from the RNA world (Brosius, 2005).

Discussion

With the determination of the human genome sequence, a novel challenge arose due to the high proportion of repeats harbouring our chromosomes. Analysis of repeat-derived transcripts (repRNA) imposes specific challenges and requires special strategies. Classical genetic approaches like mutational analysis and complementation studies cannot be applied. Here we present an unbiased analysis of transcripts derived from α satellites as a first attempt to explore this neglected part of the human genome. We determined detailed properties of αsatRNAs, an important member of human satellites.

We show that α satellite arrays are transcribed by RNA Pol II into non-polyadenylated products over 8 Kb long that localize to the nucleus. Considering the genomic distribution of satellite repeats, we assume that α satellite-containing transcripts detected via Northern blot are of the centromeric higher-order arrays origin. Northern blot analyses did not yield any signals corresponding to the size of α satellite monomers suggesting that α satellites are neither transcribed as single units nor long multimeric transcripts are processed into 171 nucleotide long monomers. Further, we present that α satellites are transcribed from both DNA strands. This observation implies that αsatRNAs in sense and antisense orientation may concurrently exist within the cell. If so, they potentially form double-stranded complexes that could trigger RNA interference response. Data supporting the presence of centromeric small siRNAs has been reported in *S. pombe* (Volpe *et al*, 2002; Hall *et al*, 2002), plants (May *et al*, 2005; Lee *et al*, 2006) and metazoans (Fukagawa *et al*, 2004; Kanellopoulou *et al*, 2005; Pal-Bhadra *et al*, 2004). Centromeric siRNAs alter the local chromatin structure of the centromeric locus that codes for those RNAs and thereby “converts nonspecific sequence information into distinct chromatin states” (Jenuwein, 2002). Yet, in the standard Northern blot analyses we were unable to detect siRNAs with the probe against α satellite, suggesting that either long αsatRNAs are not processed by Dicer into siRNAs, consistently with the notion reported in (Wang *et al*, 2006), or siRNAs are undetectable in our assays.

Northern blot analyses revealed that the steady-state levels of αsatRNAs are higher upon cellular stress. This finding matches results already reported by others that the accumulation of satellites sequences is a consequence of DNA demethylation, cellular stress or genomic instability observed i.e. in cancer (Ting *et al*, 2011; Bouzinba-Segard *et al*, 2006; Jolly *et al*, 2004; Valgardsdottir *et al*, 2008). Importantly, bioinformatics analysis of ENCODE metadata reinforce our data that αsatRNAs expression shown in HeLa cells is not due to the tumor transformation. Reads mapping to α satellite arrays were also found in GM12878 primary cells (Rozowsky *et al*, 2011) (Supplementary Table S2B). It is important to note that levels of αsatRNAs detected prior to heat shock and 1 hour later are comparable. We interpret this as evidence that expression levels from centromeric loci are tightly regulated under normal growth conditions. Studies on human artificial chromosomes (HAC) (Nakano *et al*, 2008) together with some data obtained on neocentromers (Saffery *et al*, 2003; Ishii *et al*, 2008) highlight the close correlation between low transcriptional rates with the centromere function. Low level of αsatRNAs transcription was confirmed by bioinformatics comparison of αsatRNAs amounts relative to other RNA families within deep sequenced nuclear RNA libraries from THP1 (Taft *et al*, 2010) and 5-8F (Liao *et al*, 2010) cells. To technically overcome the problem of low abundance of αsatRNAs, we designed primers and probes on the consensus sequence of α satellite monomer allowing the detection *en masse* from many α satellite genomic loci simultaneously.

With the analysis of αsatRNAs expression throughout the cell cycle, we show that transcripts level reach the peak during the S-phase. In general, the heterochromatin covering repetitive sequences does not have an open structure enabling binding of the transcription machinery. However, when centromeric DNA is being replicated during the late S phase (Shelby *et al*, 2000), silencing marks are diminished for a short period being just deposited on a newly replicated strand, what may allow the transcription to start. Moreover, the chromatin state of the centromeric

DNA shall be open enough to enable the dilution of CENP-A histone variant caused by redistribution of parental CENP-A octamers to the daughter strands (Jansen *et al*, 2007). Lyn Chan et al (Chan *et al*, 2012) provided evidence that there is an active RNA Pol II at the kinetochores of metaphase and anaphase during mitosis in human cells engaged in centromeric α satellite synthesis. This in turn, could suggest that αsatRNAs may be implicated in the kinetochore protein binding, similarly to the maize single-stranded centromeric RNA shown to stabilize CENP-C binding to the DNA (Du *et al*, 2010).

To address a potential function of αsatRNAs, we searched for interacting partners. Genomic SELEX coupled to deep sequencing is an approach that explores genomic regions regardless of their expression levels and allows the identification of functional domains within transcripts without previous knowledge on their structure or sequence. We used this technology to screen the human genome for regulatory RNAs that might interfere with transcription by direct interaction with RNA Pol II. The identification of RNA Pol II-binding aptamers in RNAs derived from α satellites was a first hint that they might contain a functional element on the transcript level. These aptamers, termed PBEs (Polymerase Binding Elements), bind to the core of RNA Pol II. The human genome does not code for a canonical RNA-dependent RNA polymerase (RdRP) and it has previously been shown that human RNA Pol II can exert this activity (Filipovska & Konarska, 2000; Wagner *et al*, 2013; Lehmann *et al*, 2007). Therefore we tested whether αsatRNAs can serve as a template for an RdRP activity. This could clearly been demonstrated, especially when showing that the size of the RdRP product depends on the position of the αsatPBE ♯111 on the RNA template. RdRP activity was detected both *in vitro* (Figure 5C, 5D; Supplementary Figure S4) in a nuclear extract as well as in HeLa cells after nucleofection, and for the first time endogenous RNA was presented as an RdRP template. In addition to RdRP activity, we also detected 3' extension of the αsatRNAs. We do not suggest that the products of these reactions are functional, but propose that they rather represent a general activity of RNA polymerases to dissociate RNA molecules trapped in the active site. This mechanism has previously been shown for the bacterial small 6S RNA, which inhibits sigma70 programmed RNA polymerase under stress conditions (Trotochaud & Wassarman, 2004, 2005). After stress, to be released from the complex, the RNA polymerase uses the 6S RNA as a template for RNA-dependent RNA polymerase activity, leading to the productions of p19 RNAs, for which no function has (yet) been shown (Wassarman & Saecker, 2006). We have previously shown that the bacterial RNA polymerase performs 3' extension on several RNAs that bind to its core regardless if these RNAs interfere with the activity of the enzyme (Windbichler *et al*, 2008). We interpret the above results in such a way that αsatRNAs might be regulators of transcription, because they contain RNA motifs that can recruit RNA Pol II to any RNA that contains such a domain. It remains to be tested if αsatRNAs recruit RNA Pol II to centromeres to potentially promote chromatin remodelling.

A novel and intriguing aspect of our analysis is the evolutionary origin of α satellites. We suggest that α satellites contain remnants of snoRNAs. When inspecting the predicted secondary structure of consensus sequences of Dfam (Wheeler *et al*, 2013) α satellites families, the highest similarity is to H/ACA snoRNAs. Bioinformatically, αsatRNAs fulfil all criteria to be classified as snoRNAs. Additionally, the RNAsnoop target prediction tool (Tafer *et al*, 2010) identified potential pseudouridilation target sites on the 28S rRNA and U5 snRNA. Several other observations support our hypothesis. Marsupials' snoRNAs have been found inserted into a retrotransposable element, called snoRTE (Schmitz *et al*, 2008), suggesting that they can disperse into many new positions. This indicates that at certain time point during evolution snoRNAs invaded retroelements giving raise to the amplification of the snoRNA motif. Moreover, when the evolutionary most distant αsatRNA homologues are folded, a short 40 nucleotides flank is observed. This 3' overhang does not belong to the snoRNA motif, but could be a remnant of the retrotransposable element. Finally, dyskerin, a human homolog of the yeast centromere binding protein Cbfp5, bridges snoRNAs and centromeres. Dyskerin being a pseudouridine synthase is a core component of H/ACA snoRNPs, but intriguingly its depletion in HeLa cells also increases the mitotic index by disrupting the formation of mitotic spindle (Alawi & Lin, 2013).

Taken together, we propose that αsatRNAs have originated from snoRNAs, lost their primary function by accumulating mutations, were reinserted into new locations, likely via retrotransposition, and propagated onto centromeres forming functional higher-order arrays of α satellites with a tightly controlled rate of transcription, presumably triggered by the αsatPBE-RNA Pol II interaction.

The spreading of noncoding RNAs via genomic repeat elements is a wide spread phenomenon. For example, Alu repeats originate from the 7SL signal recognition particle RNA (Ullu & Tschudi, 1984) and belong to SINE retrotransposons. They evolve into alternative splice sites giving rise to new exons when inserted into intronic sequences (Lev-Maor *et al*, 2003). When present in 3' UTRs, Alu elements can be targeted by *trans*-acting Alus via RNA binding protein STAU1 leading to mRNA decay (Gong & Maquat, 2011). Another example is BC1 RNA which is involved in translation repression in dendritic cells (Tiedge *et al*, 1991) and was reported to be derived from tRNAs (DeChiara & Brosius, 1987). These examples strongly sustain the notion that repeat-derived RNAs (repRNAs) originating from functional RNAs represent a rich resource and a huge reservoir for the evolution of RNAs with novel functions.

Materials & Methods

Cell culture

HeLa Ohio cells were seeded onto 10 cm plates in DMEM media supplemented with 4 mM L-glutamine and 10 % FBS and grown at 37 ºC in 5 % CO_2 atmosphere. For the analysis of αsatRNAs expression, HeLa cells were cultured under native as well as stress conditions. 80 % confluent cultured cells were subjected to heat shock at 45 ºC for 30 min with subsequent recovery at 37 ºC for 1 h, when indicated. Control cells were maintained at 37 ºC.

α amanitin treatment

5 x 10^5 HeLa cells were seeded onto 10 cm plates 42 h prior to α amanitin treatment. At time point 0, cells were washed with PBS and incubated with the media supplemented with 20 µg/ml α amanitin (Bortolin-Cavaillé *et al*, 2009). Control cells were grown in regular media. After each time point, total RNA was isolated from cells harvested from a treated and a control dish and subjected to the analysis of αsatRNAs expression.

HeLa cells synchronization

HeLa cells were synchronized with double thymidine block and thymidine-nocodazole treatment according to the protocol previously published by Wendt, K. S. et al (Wendt *et al*, 2008).

HeLa cells nucleofection

1 x 10^6 HeLa cells were nucleofected with 0.5-5 µg *in vitro* transcribed, purified RNA using Amaxa Cell Line Nucleofector Kit R (Lonza) and ATCC program on Lonza Nucleofector II according to the manufacturer's instructions. After 24 h, nucleofected and control cells were harvested for total RNA isolation.

RNA isolation

Total RNA isolation

Total RNA was isolated with the TRI Reagent (Sigma) according to the manufacturer's instructions. The quality of the RNA was assessed using UV absorption at 260 and 280 nm. Then, two consecutive DNase I (NEB) treatments were performed at 37 ºC for 30 min to remove potential genomic DNA contamination.

Nuclear/ cytoplasmic fractionation

Separation of nuclei from cytoplasm was done using the modified Sambrook and Russell protocol. 80 % confluent cells were washed with PBS and spun down at 2000 x*g* at 4 ºC for 5 min. Cell pellet was resuspended in Lysis Buffer (0.14 M NaCl, 1.5 mM $MgCl_2$, 10 mM Tris-Cl pH 8.6, 0.5 % NP-40, 10 mM Vanadyl-Ribonucleoside Complexes) and underlaid with an equal volume of Lysis Buffer containing 24 % w/v sucrose. Nuclei were fractionated by density gradient with ultrafugation at 10,000 x*g* at 4 ºC for 20 min. The cytoplasmic fraction was recovered and subjected to proteinase K digestion (200 µg/ml). The nuclear pellet was resuspend in Lysis Buffer and nuclei were disrupted and the liberated genomic DNA was sheared mechanically through the needle. Then, the nuclear fraction was digested with the proteinase K (200 µg/ml). RNA from both fractions was subjected to two consecutive DNase I (NEB) treatments at 37 ºC for 30 min and purified by standard phenol/chloroform extraction, precipitated and collected by centrifugation.

RNA analysis

Strand specific RT-PCR

For a first strand cDNA synthesis, 0.1-2 µg of DNaseI-treated RNA was used. Mixture of RNA with 1 µM of strand specific or 0.1 µM radioactively end-labelled primer was denatured at 70 ºC for 10 min. Reverse transcription reaction (OmniScript, Qiagen) was performed at 45 ºC for 1 h according to the manufacturer's protocol. "No primer" and “no reverse transcriptase” controls were included. 5 µl of the reverse transcription reaction was amplified by PCR and analyzed on agarose gel or, in case of primer extension, on 8 M urea 8 % PAA gel.

Streptavidin-Biotin pull down

150 pmol biotinylated oligo(dT) probe (Promega) was incubated with prewashed streptavidin beads (Streptavidin MagneSphere® Paramagnetic Particles, Promega) at RT for 10 min according to the manufacturer’s instructions and added to the denatured 200 µg of total RNA resuspended in 0.5 x SSC. Magnetic beads were captured and washed with 0.1 x SSC. Enriched RNA was eluted and 50 ng were analyzed with strand-specific RT-PCR.

Northern blot

15-30 µg of RNA samples were dissolved in 2 x denaturing loading dye, denatured and loaded onto a prerun 0.8 % formaldehyde agarose gel. Electrophoresis was carried out at 175 V, 4 ºC for around 5 hours. RNA was transferred onto nitrocellulose membrane Hybond N+ (Amersham) by capillary transfer o/n and covalently cross-linked to the membrane by 254 nmUV, 120 mJ/cm^2 (UV Stratalinker 2400). After prehybridization in hybridization buffer (Ambion® ULTRAhyb®-Oligo, Ambion® ULTRAhyb®, Ambion), denatured 5’ end-labeled probe was added and hybridized at 42 ºC o/n. The blot was washed according to manufacturer’s protocol and visualized by autoradiography.

Reverse RNase protection assay

Total content of the transcription reaction in the HeLa nuclear extract (see section *Transcription in HeLa nuclear extract*) with [α-^{32}P] rUTP was hybridized to a probe complementary to the input αsatRNA at 42 ºC o/n, digested with A/T1 RNases mixture at 37 ºC for 30 min, and precipitated according to the RPA III Ribonuclease Protection Assay Kit (Ambion) protocol. Samples were analyzed on 8 M urea 8 % PAA gel and visualized on a phosphoimager screen.

In vitro experiments with HeLa nuclear extract

HeLa nuclear extract preparation

HeLa cells grown in suspension were collected at 2000 rpm at 4 ºC for 15 min, washed with PBS and spun down again. Cells pellet was swelled in hypotonic buffer (20 mM HEPES pH 7.9, 1.5 mM $MgCl_2$, 10 mM KCl, 1 mM DTT, 1 mM PMSF) on ice. After Douncer homogenization, nuclei were pelleted at 2800 rpm at 4 ºC for 15 min; resuspended in resuspension buffer (20 mM HEPES pH 7.9, 1.5 mM $MgCl_2$, 420 mM KCl, 0.2 mM EDTA, 1 mM DTT, 1 mM PMSF, 20 % glycerol) and homogenized to disperse clumps. The homogenized suspension was then stirred and spun at 18,000 rpm at 4 ºC for 30 min to remove cell debris. Recovered supernatant was dialyzed to remove salts in dialysis buffer (50 mM HEPES pH 7.9, 100 mM KCl, 1 mM EDTA, 1 mM DTT, 0.1 mM PMSF, 20 % glycerol). Subsequently proteins were precipitated with $(NH_4)_2SO_4$ (0.35 g/ml of extract), collected by centrifugation at 17,000 x*g* at 4 ºC for 20 min and gently resuspended in dialysis buffer. Dialysis was done o/n at 4 ºC with 3,000 MW cut-off. Insoluble debris was pelleted at 14,000 x*g* at 4 ºC for 20 min, whereas the supernatant was snap-frozen to be stored at -80 ºC.

Incubation in HeLa nuclear extract

5-50 nM unlabeled, *in vitro* transcribed αsatPBE #111 RNA or control 50 nM PCR product, were pre-incubated at 30 ºC for 10 min in a reaction mixture containing: 1 x rNTPs mix (0.4 mM rATP, 0.4 mM rGTP, 0.4 mM rCTP, 0.016 mM rUTP), 3 mM $MgCl_2$, 1 x Transcription Buffer (20 mM HEPES pH 7.9, 100 mM KCl, 0.2 mM EDTA, 0.5 mM DTT, 20 % glycerol), 20 U RNase inhibitor. Afterwards, HeLa nuclear extract and [α-^{32}P] rUTP was added to the reaction and incubated at 30 ºC for 1 h. Transcription reaction was terminated by adding stop solution (0.3 M Tris-Cl pH 7.4, 0.3 M NaOAc, 0.5 % SDS, 2 mM EDTA) and RNA was purified by standard phenol/chloroform extraction. Transcription products were separated on 8 M urea 8-12 % PAA gel and visualized on a phosphoimager screen.

Bioinformatics

α satellites annotation

Due to the lack of α satellites annotation on the unplaced contigs of the DFAM α satellites annotation, and the fact that the α satellites annotation changed twice during our study, we decided to re-annotate α satellites on hg19. The script dfamscan.pl was used as described on the DFAM website (http://dfam.janelia.org/help/tools) (Wheeler *et al*, 2013).

Motif search

Motif search was done with MEME (Bailey & Elkan C, 1994). To this aim, all RNA Pol II binding aptamers found to overlap with α satellites were classified into direct and reverse complement sequences depending if the aptamer sequences were parallel (170 unique aptamers) or antiparallel (129 unique aptamers) to α satellite array. The motif search was separately done for both groups. For each group, a random pool of 20000 α satellites was used as negative background. Based on this background distribution, MEME was subsequently used to search for motifs. The search explicitly required finding exactly one motif per sequence, with a maximal motif size of 30 nucleotides, and a minimal length of 6 nucleotides.

snoRNA target prediction

The target prediction for snoRNA-like alphoid sequences was done using RNAsnoop (Tafer *et al*, 2010). The query sequence used was the consensus sequence of the 3 intronic sequences homologues to human α satellites found in marmoset. Target search was run on the marmoset snoRNAs and the human rRNAs, as no complete marmoset rRNA sequences was reported.

Expression of non-coding RNAs in RNAseq data

The set of nuclear long non-coding RNAs were retrieved from (Ip & Nakagawa, 2012).

Acknowledgements

We thank all members of the Schroeder lab for many discussions and especially acknowledge Meghan Lybecker's input in revising the manuscript. We thank the Austrian Science Fund for grants number F4301 and F4308, and the Austrian GenAU programm grant D-110420-010 to R.S.

Author Contribution

Katarzyna Matylla-Kulinska conceived and performed experiments and wrote the paper. Hakim Tafer performed bioinformatic analysis on α satellites and partially contributed to the writing of the article, Bob Zimmermann performed bioinformatic analysis on PBEs, Jennifer Boots helped with experiments, Maximilian Radtke performed an experiment, Renée Schroeder is the project leader, conceived the project and wrote the manuscript.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Alawi F & Lin P (2013) Dyskerin Localizes to the Mitotic Apparatus and Is Required for Orderly Mitosis in Human Cells. *PLoS One* **8:** e80805

Amor DJ & Choo KHA (2002) Neocentromeres: role in human disease, evolution, and centromere study. *Am. J. Hum. Genet.* **71:** 695–714

Bailey TL & Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Int. Conf. Intell. Syst. Mol. Biol.* **2:** 28–36

Bennetzen JL (1996) The contributions of retroelements to plant genome organization, function and evolution. *Trends Microbiol.* **4:** 347–353

Bortolin-Cavaillé M-L, Dance M, Weber M & Cavaillé J (2009) C19MC microRNAs are processed from introns of large Pol-II, non-protein-coding transcripts. *Nucleic Acids Res.* **37:** 3464–73

Bouzinba-Segard H, Guais A & Francastel C (2006) Accumulation of small murine minor satellite transcripts leads to impaired centromeric architecture and function. *Proc. Natl. Acad. Sci. U. S. A.* **103:** 8709–14

Brosius J (1999) Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica* **107:** 209–38

Brosius J (2003) The contribution of RNAs and retroposition to evolutionary novelties. *Genetica* **118:** 99–116

Brosius J (2005) Echoes from the past – are we still in an RNP world ? *Cytogenet. Genome Res.* **24:** 8–24

Brosius J & Tiedge H (1996) Reverse Transcriptase: Mediator of Genomic Plasticity. *Virus Genes* **11:** 163–179

Chan FL, Marshall OJ, Saffery R, Kim BW, Earle E, Choo KHA & Wong LH (2012) Active transcription and essential role of RNA polymerase II at the centromere during mitosis. *Proc. Natl. Acad. Sci. U. S. A.* **109:** 1979–84

DeChiara TM & Brosius J (1987) Neural BC1 RNA: cDNA clones reveal nonrepetitive sequence content. *Proc. Natl. Acad. Sci. U. S. A.* **84:** 2624–8

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, *et al* (2012) Landscape of transcription in human cells. *Nature* **489:** 101–8

Du Y, Topp CN & Dawe RK (2010) DNA binding of centromere protein C (CENPC) is stabilized by single-stranded RNA. *PLoS Genet.* **6:** e1000835

Feschotte C (2008) Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9:** 397–405

Filipovska J & Konarska MM (2000) Specific HDV RNA-templated transcription by pol II in vitro. **6:** 41–54

Fukagawa T, Nogami M, Yoshikawa M, Ikeno M, Okazaki T, Takami Y, Nakayama T & Oshimura M (2004) Dicer is essential for formation of the heterochromatin structure in vertebrate cells. *Nat. Cell Biol.* **6:** 784–91

Gao X & Voytas DF (2005) A eukaryotic gene familly related to retroelement integrases. *Trends Genet.* **21:** 129–33

Gong C & Maquat LE (2011) lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature* **470:** 284–8

Hall IM, Shankaranarayana GD, Noma K-I, Ayoub N, Cohen A & Grewal SIS (2002) Establishment and maintenance of a heterochromatin domain. *Science (80-.).* **297:** 2232–7

Hayden KE, Strome ED, Merrett SL, Lee H-R, Rudd MK & Willard HF (2013) Sequences associated with centromere competency in the human genome. *Mol. Cell. Biol.* **33:** 763–72

Ip JY & Nakagawa S (2012) Long non-coding RNAs in nuclear bodies. *Dev. Growth Differ.* **54:** 44–54

Ishii K, Ogiyama Y, Chikashige Y, Soejima S, Masuda F, Kakuma T, Hiraoka Y & Takahashi K (2008) Heterochromatin integrity affects chromosome reorganization after centromere dysfunction. *Science* **321:** 1088–91

Jansen LET, Black BE, Foltz DR & Cleveland DW (2007) Propagation of centromeric chromatin requires exit from mitosis. *J. Cell Biol.* **176:** 795–805

Jenuwein T (2002) Molecular biology. An RNA-guided pathway for the epigenome. *Science (80-.).* **297:** 2215–8

Jolly C, Metz A, Govin J, Vigneron M, Turner BM, Khochbin S & Vourc'h C (2004) Stress-induced transcription of satellite III repeats. *J. Cell Biol.* **164:** 25–33

Kanellopoulou C, Muljo SA, Kung AL, Ganesan S, Drapkin R, Jenuwein T, Livingston DM & Rajewsky K (2005) Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes Dev.* **19:** 489–501

Lai J, Li Y, Messing J & Dooner HK (2005) Gene movement by Helitron transposons contributes to the haplotype variability of maize. *Proc. Natl. Acad. Sci. U. S. A.* **102:** 9068–73

Larkin M a, Blackshields G, Brown NP, Chenna R, McGettigan P a, McWilliam H, Valentin F, Wallace IM, Wilm a, Lopez R, Thompson JD, Gibson TJ & Higgins DG (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23:** 2947–8

Lee H-R, Neumann P, Macas J & Jiang J (2006) Transcription and evolutionary dynamics of the centromeric satellite repeat CentO in rice. *Mol. Biol. Evol.* **23:** 2505–20

Lehmann E, Brueckner F & Cramer P (2007) Molecular basis of RNA-dependent RNA polymerase II activity. *Nature* **450:** 445–9

Lev-Maor G, Sorek R, Shomron N & Ast G (2003) The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science* **300:** 1288–91

Liao J-Y, Ma L-M, Guo Y-H, Zhang Y-C, Zhou H, Shao P, Chen Y-Q & Qu L-H (2010) Deep sequencing of human nuclear and cytoplasmic small RNAs reveals an unexpectedly complex subcellular distribution of miRNAs and tRNA 3' trailers. *PLoS One* **5:** e10563

Mattick JS (2005) The functional genomics of noncoding RNA. *Science* **309:** 1527–1528

May BP, Lippman ZB, Fang Y, Spector DL & Martienssen R (2005) Differential regulation of strand-specific transcripts from Arabidopsis centromeric satellite repeats. *PLoS Genet.* **1:** e79

Nakano M, Cardinale S, Noskov VN, Gassmann R, Vagnarelli P, Kandels-Lewis S, Larionov V, Earnshaw WC & Masumoto H (2008) Inactivation of a human kinetochore by specific targeting of chromatin modifiers. *Dev. Cell* **14:** 507–22

Pal-Bhadra M, Leibovitch B a, Gandhi SG, Chikka MR, Rao M, Bhadra U, Birchler J a & Elgin SCR (2004) Heterochromatic silencing and HP1 localization in Drosophila are dependent on the RNAi machinery. *Science* **303:** 669–72

Prasanth K V & Spector DL (2007) Eukaryotic regulatory RNAs : an answer to the " genome complexity " conundrum. *Genes Dev.* **21:** 11–42

Prosser J, Frommer M, Paul C & Vincent PC (1986) Sequence relationships of three human satellite DNAs. *J. Mol. Biol.* **187:** 145–155

Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjornson R, Kong Y, Kitabayashi N, Bhardwaj N, Rubin M, Snyder M & Gerstein M (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7:** 522

Saffery R, Sumer H, Hassan S, Wong LH, Craig JM, Todokoro K, Anderson M, Stafford A & Choo KHA (2003) Transcription within a functional human centromere. *Mol. Cell* **12:** 509–516

Schmitz J, Zemann A, Churakov G, Kuhl H, Grützner F, Reinhardt R & Brosius J (2008) Retroposed SNOfall — A mammalian-wide comparison of platypus snoRNAs. *Genome Res.* **18:** 1005–1010

Shelby RD, Monier K & Sullivan KF (2000) Chromatin assembly at kinetochores is uncoupled from DNA replication. *J. Cell Biol.* **151:** 1113–8

Shepelev VA, Alexandrov AA, Yurov YB & Alexandrov IA (2009) The Evolutionary Origin of Man Can Be Traced in the Layers of Defunct Ancestral Alpha Satellites Flanking the Active Centromeres of Human Chromosomes. *PLoS Genet.* **5:** e1000641

Tafer H, Kehr S, Hertel J, Hofacker IL & Stadler PF (2010) RNAsnoop: efficient target prediction for H / ACA snoRNAs. *Bioinformatics* **26:** 610–616

Taft RJ, Simons C, Nahkuri S, Oey H, Korbie DJ, Mercer TR, Holst J, Ritchie W, Wong JJ-L, Rasko JEJ, Rokhsar DS, Degnan BM & Mattick JS (2010) Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans. *Nat. Struct. Mol. Biol.* **17:** 1030–4

Tiedge H, Fremeau RT, Weinstock PH, Arancio O & Brosius J (1991) Dendritic location of neural BC1 RNA. *Proc. Natl. Acad. Sci. U. S. A.* **88:** 2093–2097

Ting DT, Lipson D, Paul S, Brannigan BW, Akhavanfard S, Coffman EJ, Contino G, Deshpande V, Iafrate AJ, Letovsky S, Rivera MN, Bardeesy N, Maheswaran S & Haber DA (2011) Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science (80-.).* **331:** 593–6

Trotochaud AE & Wassarman KM (2004) 6S RNA function enhances long-term cell survival. *J. Bacteriol.* **186:** 4978–4985

Trotochaud AE & Wassarman KM (2005) A highly conserved 6S RNA structure is required for regulation of transcription. *Nat. Struct. Mol. Biol.* **12:** 313–319

Ullu E & Tschudi C (1984) Alu sequences are processed 7SL RNA genes. *Nature* **312:** 171–2

Valgardsdottir R, Chiodi I, Giordano M, Rossi A, Bazzini S, Ghigna C, Riva S & Biamonti G (2008) Transcription of Satellite III non-coding RNAs is a general stress response in human cells. *Nucleic Acids Res.* **36:** 423–34

Verdaasdonk JS & Bloom K (2012) Centromeres: unique chromatin structures that drive chromosome segregation. *Nat. Rev. Mol. Cell Biol.* **12:** 320–332

Volpe T a, Kidner C, Hall IM, Teng G, Grewal SIS & Martienssen R a (2002) Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science (80-.).* **297:** 1833–7

Wagner SD, Yakovchuk P, Gilman B, Ponicsan SL, Drullinger LF, Kugel JF & Goodrich JA (2013) RNA polymerase II acts as an RNA-dependent RNA polymerase to extend and destabilize a non-coding RNA. *EMBO J.* **32:** 781–90

Wang F, Koyama N, Nishida H, Haraguchi T, Reith W & Tsukamoto T (2006) The assembly and maintenance of heterochromatin initiated by transgene repeats are independent of the RNA interference pathway in mammalian cells. *Mol. Cell. Biol.* **26:** 4028–40

Wassarman KM & Saecker RM (2006) Synthesis-mediated release of a small RNA inhibitor of RNA polymerase. *Science* **314:** 1601–3

Wayel JS & Willard HF (1987) Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: a survey of alphoid sequences from different human chromosomes. *Nucleic Acids Res.* **15:** 7549–7569

Weil C & Martienssen R (2008) Epigenetic interactions between transposons and genes: lessons from plants. *Curr. Opin. Genet. Dev.* **18:** 188–192

Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, Tsutsumi S, Nagae G, Ishihara K, Mishiro T, Yahata K, Imamoto F, Aburatani H, Nakao M, Imamoto N, Maeshima K, Shirahige K & Peters J-M (2008) Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451:** 796–801

Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, Smit AFA & Finn RD (2013) Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.* **41:** D70–82

Windbichler N, Pelchrzim von F, Mayer O, Csaszar E & Schroeder R (2008) Isolation of small RNA-binding proteins from E.coli. *RNA Biol.* **5:** 1–11

Wong LH, Brettingham-Moore KH, Chan L, Quach JM, Anderson MA, Northrop EL, Hannan R, Saffery R, Shaw ML, Williams E & Choo KHA (2007) Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere. *Genome Res.* **17:** 1146–60

Legend to Figures

Figure 1. αsatRNAs are transcribed in HeLa cells. A. Schematic representation of α satellite array amplification. Fwd and Rev primers are designed to amplify one α satellite unit. Primers can also hybridize to flanking units, giving rise to a population of products varying by 171 bp. **B. αsatRNAs are derived from both DNA strands.** Total RNA extracted from HeLa cells was subjected to strand specific RT-PCR run in the presence of radioactive [α-32P] dGTP. Amplicons were analyzed on 5 % native PAGE and visualized by autoradiography. Fwd primer in RT step detects RC-αsatRNA, Rev primer primes D-αsatRNA. As a control for strand specificity: RT reaction was performed without any primer (-). RT enzyme was omitted in –RT control to detect genomic DNA contamination. PCR on genomic DNA served as a positive control. **C. αsatRNAs are longer than 8 Kb.** 30 µg of total RNA extracted from HeLa cells grown under normal condition (37 °C, ∞), heat shocked (45 °C, 30') and heat shocked (45 °C, 30') followed by 1 h recovery (37 °C, 60'), was analyzed by Northern blot to estimate the size and relative abundance of αsatRNAs. End-labeled, degenerate LNA-modified DNA probes were used to detect D-αsatRNA and RC-αsatRNA. The profile of ethidium bromide stained ribosomal 18S and 28S served as a loading control and is presented below the blots.

Figure 2. αsatRNAs localize to the nucleus. A. αsatRNAs are detected in the nuclear fraction by Northern blot. 15 µg of fractionated HeLa RNA was analyzed by Northern blot with end-labeled, degenerate LNA-modified DNA detecting αsatRNAs from + and - DNA strands (D-αsatRNA, RC-αsatRNA, respectively). 20 ng of *in vitro* transcribed α satellite unit (inv D-αsat/inv RC-αsat) served as a positive control on each blot. The profile of ethidium bromide stained ribosomal 18S and 28S served as a loading control and is presented the blots below. **B. Nuclear localization of αsatRNAs was verified by strand specific RT-PCR.** αsatRNAs were detected in cytoplasmic and nuclear RNA fractions. Products of strand specific RT-PCR were analyzed on agarose gel. Fractionation accuracy was assessed by *Gapdh* and *Kcnq1ot1* localization.

Figure 3. αsatRNAs are atypical RNA Pol II transcripts. A. αsatRNAs are α amanitin sensitive. Total HeLa RNA extracted from α amanitin-treated (20 µg/ml) and control cells at time points 0 h, 6 h, 24 h and 48 h, was assayed for α satellite expression by strand specific RT-PCR. To verify the specificity of α amanitin treatment, *Gapdh* (RNA Pol II) and 5S rRNA (RNA Pol III) were amplified. PCR on genomic DNA was performed as positive control for D-αsat and RC-αsat amplification. **B. αsatRNAs do not contain poly(A) tails.** Polyadenylated transcripts were pulled down with oligo(dT) probe from total HeLa RNA. Pulled down (PD) and flow through (FT) fractions were examined for expression of D-αsatRNA and RC-αsatRNA by strand specific RT-PCR. Bands marked with an asterisk were confirmed by sequencing to be derived from α satellites. *Gapdh* and U6 transcripts served as controls for pull down procedure**. C. The experimental setup for analysis of 5' terminus of transcripts.** Using 5' adapter ligation reaction from FirstChoice RLM-RACE Kit (Ambion), the 5' terminus of transcripts was analyzed. The ligation of the 5' adapter depends on the presence of the cap structure on the RNA and TAP treatment that removes the cap leaving 5' monophosphate ready for adapter ligation. CIP treatment, prior to TAP, allows discrimination between capped and processed or degradation products. **D. αsatRNAs possess 5' cap structure.** Total RNA isolated from HeLa was subjected to CIP/TAP treatment and analyzed by RT-PCR with α satellite primers. -RT reaction was included to control for genomic DNA contamination. *Gapdh* transcript served as positive control.

Figure 4. RNA Pol II-binding aptamers derived from α satellites. Overlaps between RNA Pol II-binding aptamers and α satellite units are shown in brown. Because the motifs are found to be located on the boundaries between two α satellite units, the position of the overlaps is shown with respect to their location in the offset α satellite, i.e. satellite sequences shifted by 85 nucleotides (offset αsat). The bold arrow indicates the corresponding position in the non-shifted frame. The motifs are shown as weblogo. The p-value corresponds to the p-value return by MAST when scanning the motifs against α satellite dimer.

Figure 5. αsatRNAs interact with the active site of RNA Pol II. A. αsatRNA is a substrate for RNA-dependent RNA polymerase activity *in vivo*. A chimeric RNA

template consisting of the α satellite aptamer for RNA Pol II (αsatPBE ♯111) and a 15-mer non-human sequence fused upstream, was nucleofected into HeLa cells. After 24 h, cells were harvested, total RNA was isolated and subjected to the strands specific RT-PCR. Results obtained on RNA isolated from HeLa cells are presented on the right site. RT-PCR performed on the input RNA is shown on the left site. **B. Detection of labeled αsatRNA products is blocked when DRB is added to HeLa nuclear extract reaction.** 50 nM RNA template consisting of 70-mer α satellite aptamer for RNA Pol II (αsatPBE ♯111) fused with 15 nucleotides non-human sequence (Δ), was incubated in a HeLa nuclear extract with [α-^{32}P] rUTP in the presence or absence of 60 µM DRB inhibitor (+/-). Resulting transcripts were then analyzed on denaturing 8 M urea 8 % PAA gel. Reaction without added template (-) serves as a control for background transcription from remaining nucleic acids in the nuclear extract. End-labeled 70 and 85-mer RNAs served as size markers. **C. αsatRNA is a specific target for RdRP activity.** Templates used are shown on top: double-stranded DNA fragment containing T7 promoter (►), 70-mer αsatPBE ♯111 aptamer fused to 15-mer non human sequence, located at the 3' or 5' terminus of the template (◘, Δ, respectively). 100 nM templates were incubated in HeLa nuclear extract in the presence of [α-^{32}P] rUTP. Products of *in vitro* transcription reaction were analyzed on 8 M urea12 % PAA gel. To control for a background transcription, the reaction with no added template (-) was performed in parallel. End-labeled 70 and 85-mer RNAs served as size markers. **D. Schematic interpretation for the results obtained in C.** RNA Pol II binds to the aptamer αsatPBE ♯111 sequence within αsatRNA either omitting or including the non-human 15-mer to the radioactively labeled product.

Figure 6. Reverse RNase Protection Assay detects 3' extension on the αsatRNA. A. Scheme for the experimental setup. B. αsatRNA is extended by a few residues upon incubation in HeLa nuclear extract. 50 nM RNA template consisting of 70-mer α satellite aptamer for RNA Pol II (αsatPBE ♯111) was incubated in HeLa nuclear extract in the presence of [α-^{32}P] rUTP. After stopping the transcription reaction, the output RNA was precipitated, combined with a complementary RPA probe, subjected to RNases A/T1 treatment and analyzed on 8 M urea 8 % PAA gel. Samples digested with RNases A/T1 (+) were run along with

untreated controls (-). Output RNA precipitated from the HeLa nuclear extract is loaded in the lane 1 as a size reference. P, D in the schematic top panel strand for protection and degradation, respectively.

Figure 7. αsatRNAs contain remnants of snoRNAs. A. Consensus secondary structure of the human α satellite consensus sequences extracted from Dfam database. Red-colored base pairs show no compensatory mutation. Ocher base pairs have one compensatory mutation. **B. Consensus structure of the marmoset alphoid sequences located in introns.** The same color code as in A was used. **C. Best-scored target predictions for the marmoset alphoid consensus sequence as predicted by RNAsnoop** (Tafer *et al*, 2010)**.**

Figure 1

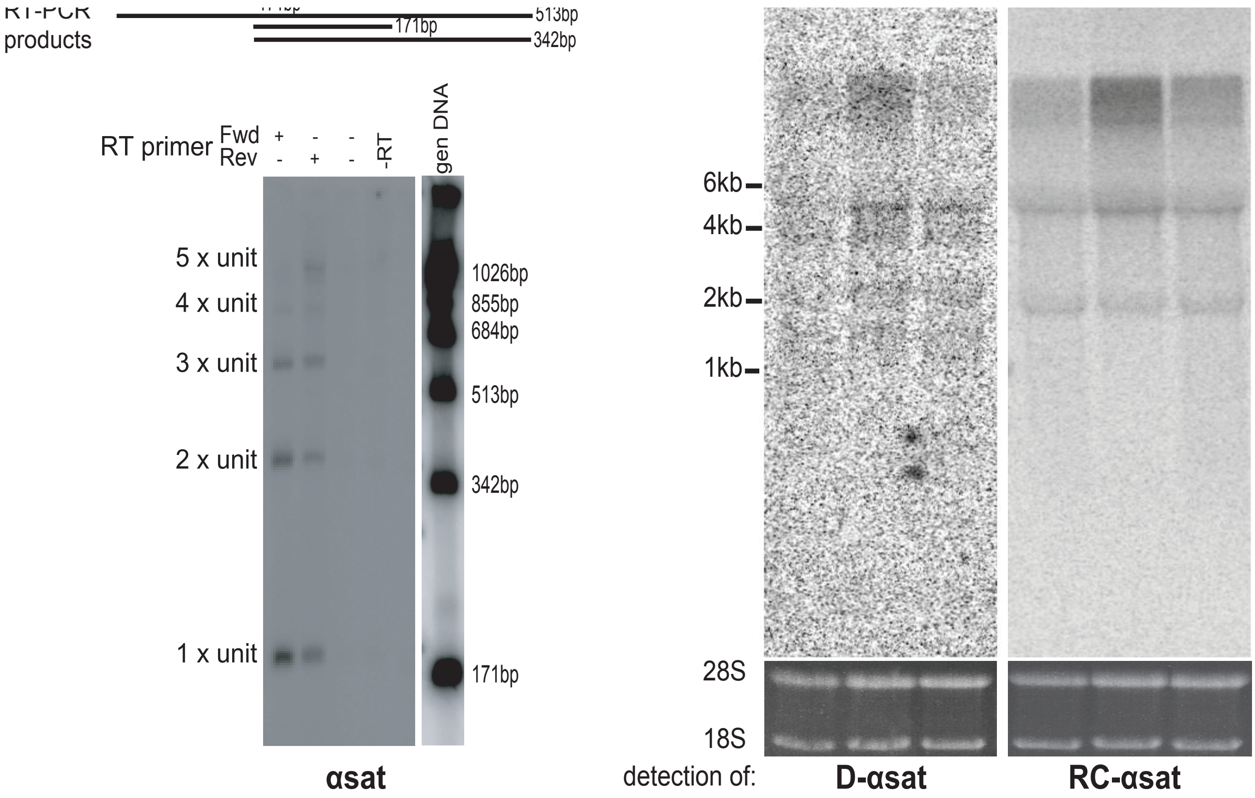

RT-PCR
products
171bp
513bp
171bp
342bp
RT primer
Fwd + - -
Rev - + -
-RT
gen DNA
5 x unit
4 x unit
3 x unit
2 x unit
1 x unit
1026bp
855bp
684bp
513bp
342bp
171bp
αsat
6kb
4kb
2kb
1kb
28S
18S
detection of:
D-αsat
RC-αsat

Figure 2

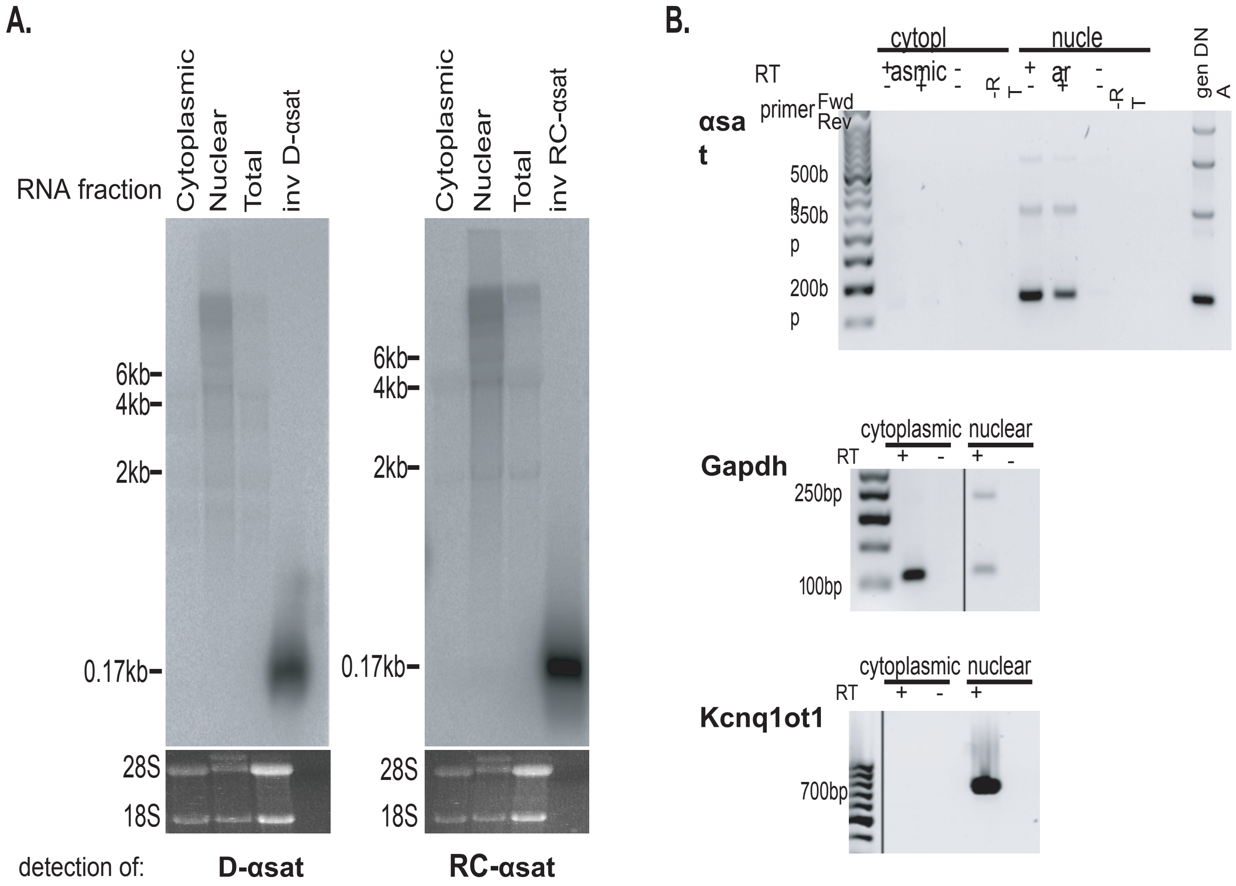

A.
RNA fraction
Cytoplasmic
Nuclear
Total
inv D-αsat
Cytoplasmic
Nuclear
Total
inv RC-αsat
6kb
4kb
2kb
0.17kb
6kb
4kb
2kb
0.17kb
28S
18S
28S
18S
detection of:
D-αsat
RC-αsat
B.
cytoplasmic
nuclear
RT
+ - + - -
-RT
+ - + - -
-RT
primer
Fwd
Rev
gen DNA
αsat
500bp
350bp
200bp
Gapdh
cytoplasmic
nuclear
RT
+ - + -
250bp
100bp
Kcnq1ot1
cytoplasmic
nuclear
RT
+ - +
700bp

Figure 3

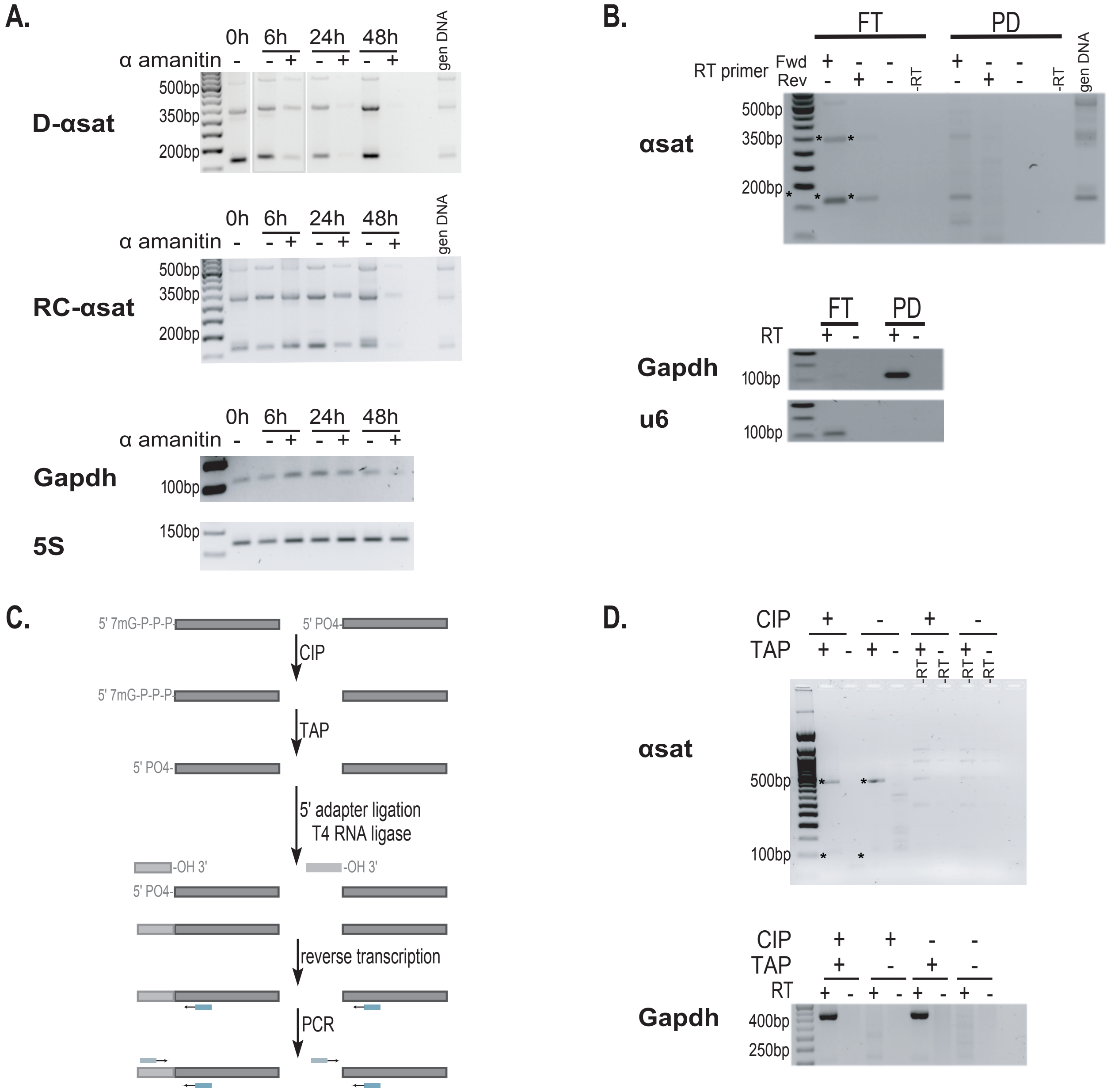

A.
0h
6h
24h
48h
α amanitin
- - + - + - +
gen DNA
D-αsat
500bp
350bp
200bp
RC-αsat
500bp
350bp
200bp
Gapdh
100bp
5S
150bp
B.
FT
PD
RT primer
Fwd
Rev
+ - - -RT + - - -RT
- + - - + -
gen DNA
αsat
500bp
350bp
200bp
*
Gapdh
u6
RT
+ - + -
100bp
100bp
C.
5' 7mG-P-P-P
5' PO4-
CIP
TAP
5' adapter ligation
T4 RNA ligase
-OH 3'
reverse transcription
PCR
D.
CIP
TAP
+ - + -
-RT
αsat
500bp
100bp
*
CIP
TAP
RT
+ + - -
+ - + -
Gapdh
400bp
250bp

Figure 4

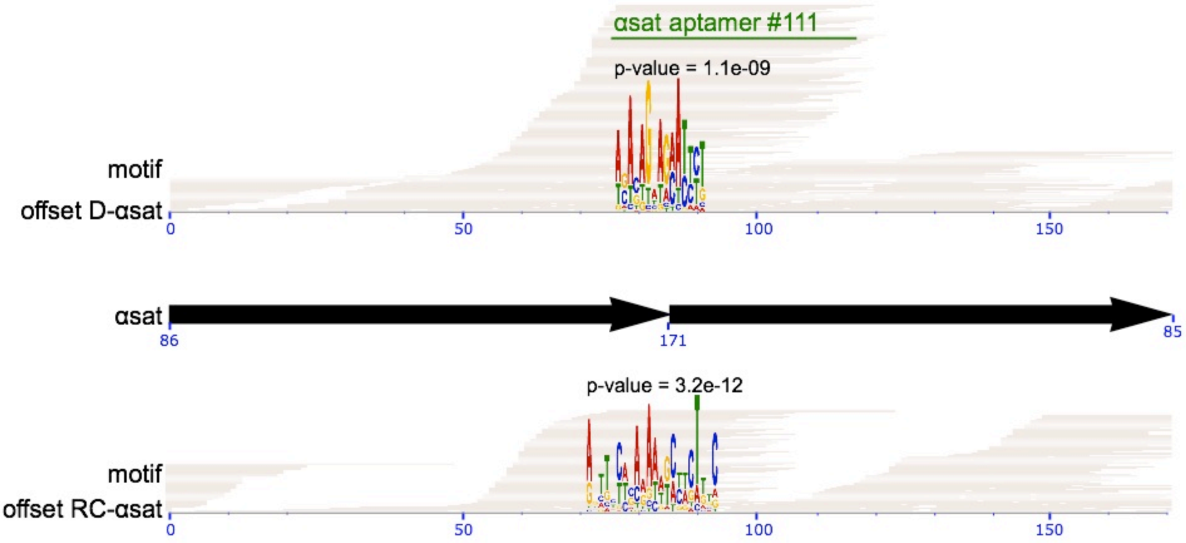

αsat aptamer #111
p-value = 1.1e-09
motif
offset D-αsat
0
50
100
150
αsat
86
171
85
p-value = 3.2e-12
motif
offset RC-αsat
0
50
100
150

Figure 5

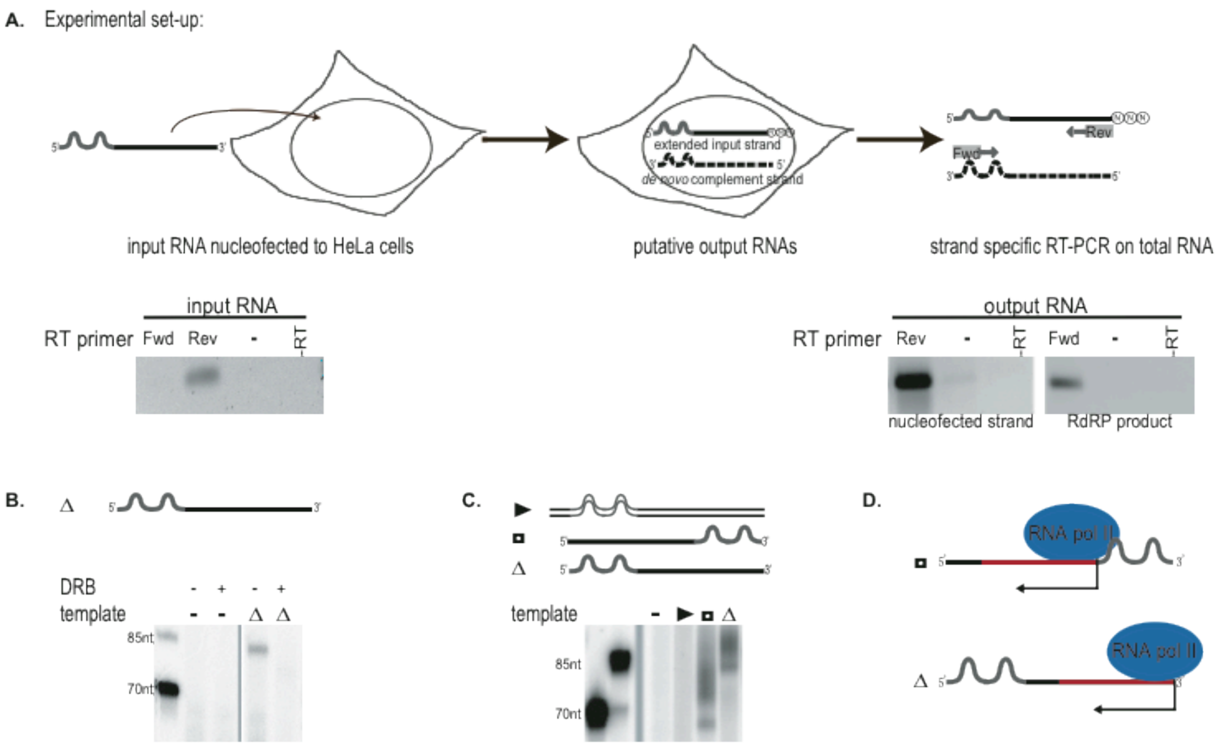

A. Experimental set-up:
5'
3'
extended input strand
de novo complement strand
Rev
Fwd
input RNA nucleofected to HeLa cells
putative output RNAs
strand specific RT-PCR on total RNA
input RNA
RT primer Fwd Rev - -RT
output RNA
RT primer Rev - -RT Fwd - -RT
nucleofected strand
RdRP product
B.
Δ
DRB - + - +
template - - Δ Δ
85nt
70nt
C.
template - ▶ ◘ Δ
85nt
70nt
D.
RNA pol II
◘
Δ

Figure 6

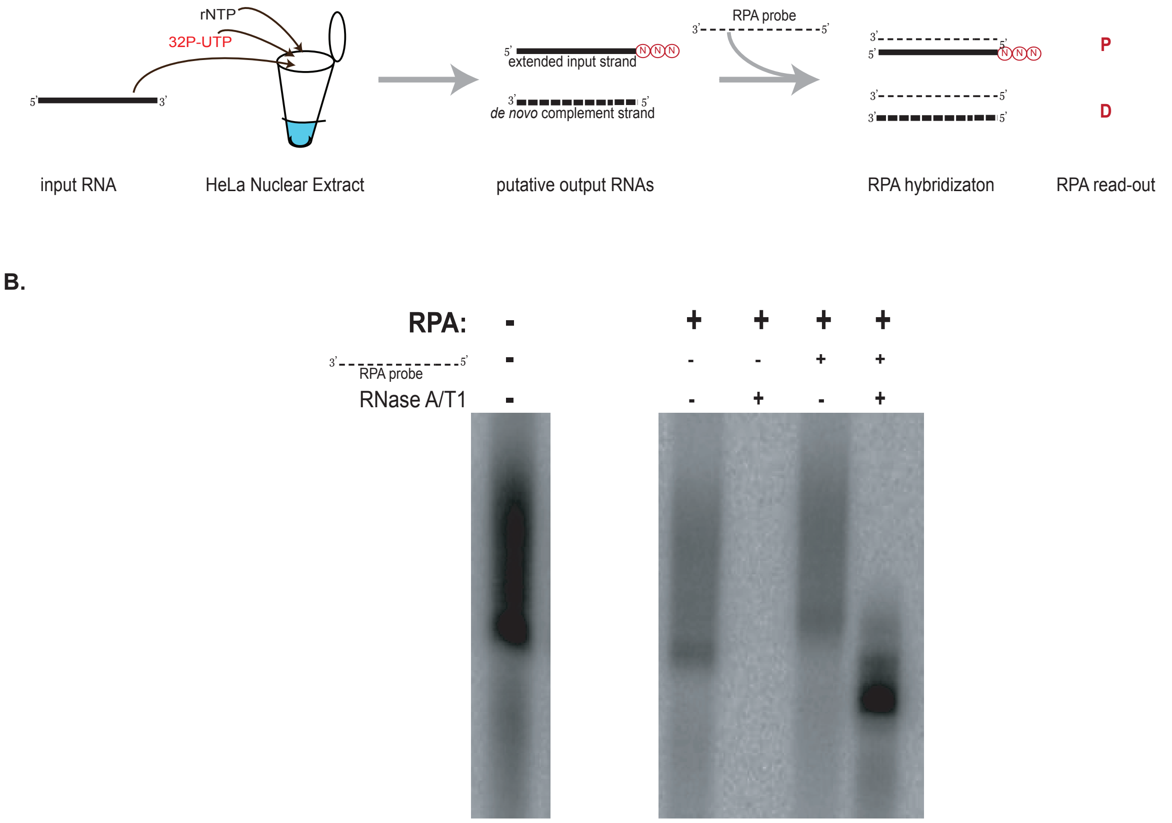
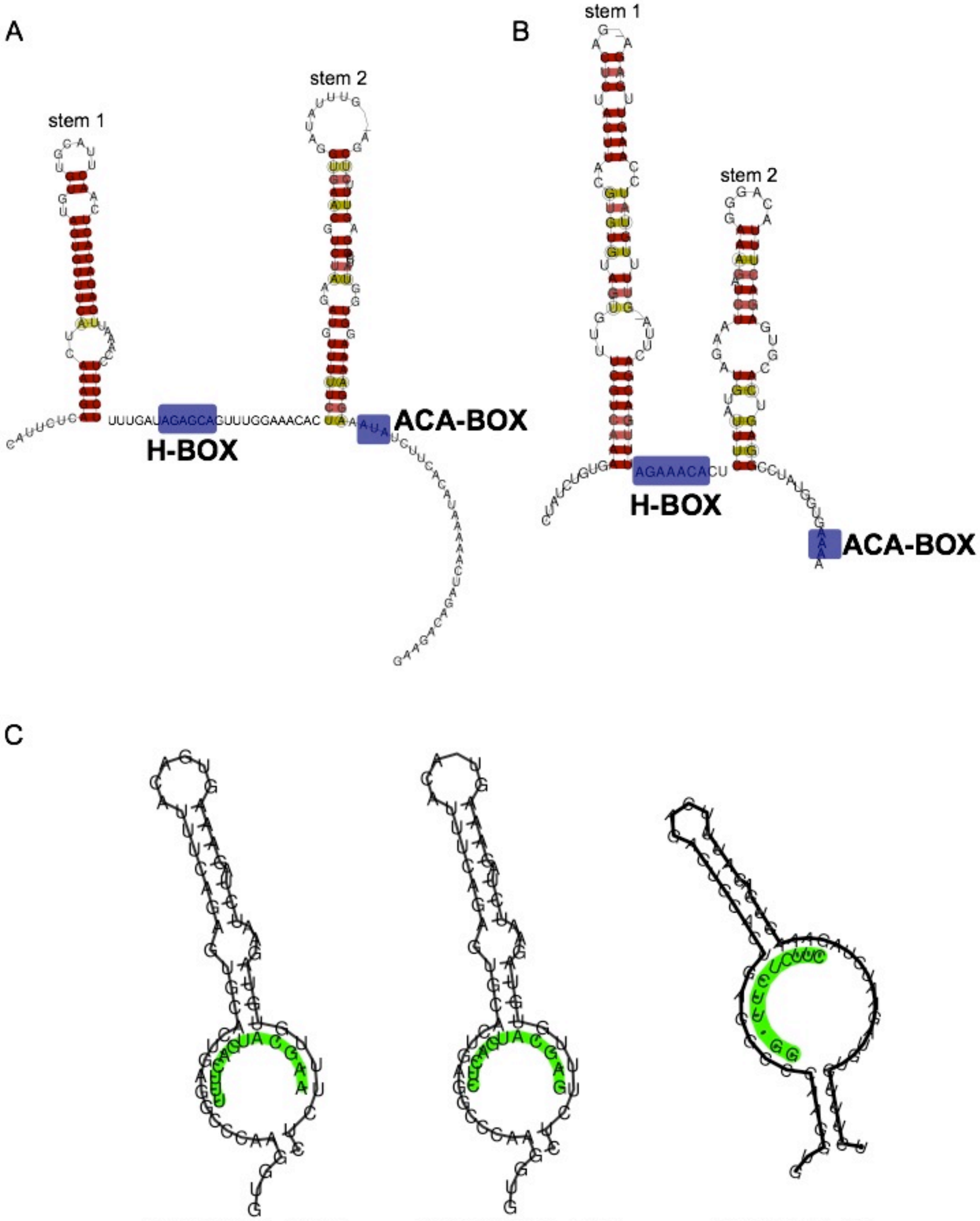

rNTP
32P-UTP
5'
3'
input RNA
HeLa Nuclear Extract
5'
extended input strand
N N N
3'
5'
de novo complement strand
putative output RNAs
RPA probe
3'
5'
RPA hybridizaton
P
D
RPA read-out
B.
RPA: - + + + +
3' RPA probe 5' - - - + +
RNase A/T1 - - + - +

Figure 7



stem 2: 28S_4276

stem 2: 28S_4571

stem 2: U5_14

RNA polymerase II-binding aptamers in human ACRO1 satellites disrupt transcription in *cis*

Jennifer L. Boots[1,†], Frederike von Pelchrzim[1,†], Adam Weiss[1,†], Bob Zimmermann[1,†], Theres Friesacher[4], Maximilian Radtke[1], Marek Żywicki[2], Doris Chen[1], Katarzyna Matylla-Kulinska[1], Florian Brueckner[3], Patrick Cramer [3], Bojan Zagrovic[4] and Renée Schroeder[1,*]

[1] Department of Biochemistry and Molecular Cell Biology, Max F. Perutz Laboratories, University of Vienna, Dr. Bohrgasse 9/5; A-1030 Vienna, Austria.

[2] Laboratory of Computational Biology, A. Mickiewicz University, Poznan, Poland.

[3] Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany.

[4] Department of Structural and Computational Biology, Max F. Perutz Laboratories, University of Vienna, Campus Vienna Biocenter 5, A-1030 Vienna, Austria.

[†] These authors contributed equally to this work.

[*] To whom correspondence should be addressed. Tel: +43-1-4277-54690, Email: renee.schroeder@univie.ac.at

Classification: Biological Sciences, Genetics

ABSTRACT

Transcription elongation is not a smooth process. The elongation rate depends on the underlying DNA sequence and varies on a gene-by-gene basis. The direct interplay between the nascent RNA and the transcribing RNA polymerase is a poorly explored field. Here we screened the human genome for RNAs that regulate transcription elongation by direct binding to RNA polymerase II (Pol II). We performed Genomic SELEX with a human genome-derived RNA library as prey and a highly purified Pol II as bait. We identified a variety of RNA polymerase II-binding APtamers (RAPs), which are prominent in repeat elements such as ACRO1 satellites, LINE1 retrotransposons and CA simple repeats, and also enriched in several protein-coding genes. When translated into protein in *silico*, human RAPs are highly enriched in the amino acids proline, serine and threonine, which are found in the CTD heptapeptide repeat of Pol II. In particular, ACRO1 satellites exhibit a strong similarity to Pol II CTD on the protein level. These observations have implications concerning the recently proposed mRNA/protein complementarity hypothesis and the origin of the genetic code, but also suggest a mechanism for RAP binding. Finally, using a reporter construct, we show that a subset of RAPs potently inhibit Pol II elongation *in cis*. We propose a novel mode of transcriptional regulation, in which the nascent RNA binds Pol II to silence its own expression, and hypothesize that this mechanism is employed by repetitive DNA elements.

Keywords: Transcription, RNA polymerase II, RNA aptamers, regulatory RNAs,

Significance Statement max 120 words PNAs

*We demonstrate the power of genomic SELEX in combination with deep sequencing as an approach to find silencing RNAs in complex genomes. We identified RNA Polymerase II-binding RNA aptamers (RAPs) as a novel class of RNA signals that control transcription in *cis*. RAPs can be found in repeat elements like ACRO1 satellites and LINE1 retrotransposons. A most intriguing observation is that ACRO1 satellites translate into a protein sequence with high similarity to the Pol II CTD heptapeptide repeat, which has implications for their evolution and the origin of the genetic code.

Introduction

Control of gene expression is essential for all living organisms to coordinate growth and development. Transcription, as the first step, is tightly regulated, and Pol II progression along the gene is not smooth. Pol II pauses at the promoter-proximal region and also during elongation (1–3). The dynamics of the elongating polymerase vary on a gene-by-gene basis suggesting that the underlying gene sequence is a significant factor for transcription efficiency (1, 4). A large number of protein factors regulate transcription in various ways and, recently, several RNAs have been identified that interfere with transcription via diverse mechanisms. For example, long non-coding RNAs affect transcription by changing chromatin structure and function (5). A few RNAs have been demonstrated to be *trans*-acting regulators of transcription (6–8). To date, only three naturally occurring trans-acting RNAs have been reported to directly bind to RNA polymerase and inhibit transcription: 6S RNA (*E. coli*), B2 RNA (*M. musculus*) and Alu RNA (*H. sapiens*) (6–8). In addition, an *in vitro* selected RNA, the FC aptamer, is able to inhibit transcription of yeast Pol II *in vitro* by binding to the active center cleft of Pol II (9).

The bacterial 6S RNA is the best-studied example of a *trans*-acting RNA that regulates the activity of bacterial RNA polymerase. Upon entry into stationary phase, 6S RNA binds the active center of the σ^{70}-containing holoenzyme and inhibits housekeeping transcription (6, 10). In order to recycle the polymerase, 6S RNA is used as a template in an RNA-dependent RNA polymerase reaction which disrupts the RNA-protein interactions and allows 6S RNA to slide out of the active center (10). In eukaryotes, small RNAs have also been suggested to inhibit housekeeping transcription *in trans* by direct binding to Pol II. Mouse B2 and human Alu RNAs are induced by stress (11) and downregulate initiation of Pol II transcription at promoters (8, 12). Certain RNAs are also able to serve as a template for an ancient RNA-dependent RNA polymerase activity of Pol II (13–15). These examples show that RNA polymerases are very versatile machines capable of accommodating many different RNAs and of adapting to changing demands.

A less-explored field is the impact of *cis*-acting nascent RNA-borne signals on transcription. Bacterial riboswitches, located in the 5’ untranslated regions of mRNAs, can dynamically refold in response to ligand binding or temperature shift and promote transcription elongation or termination (12, 16). Similarly, eukaryotic Pol II activity has been shown to be affected by secondary structure in the nascent RNA. By inhibition of backtracking stable structural elements prevent pausing and thereby increase the rate of transcription (17, 18). Furthermore, nascent RNAs can bind and trap transcription factors to the site of transcription contributing to transcription factor association with their cognate DNA elements (19). Alternatively, nascent transcripts can recruit proteins that cause transcription attenuation (16). For example, the recognition motifs of Nrd1 and Nab3, components of a yeast transcription terminator complex, are enriched in ncRNAs but depleted from mRNAs (16, 20).

In this work, we tested the hypothesis that the direct interaction of the nascent RNA with the transcription machinery is able to regulate transcription. This phenomenon is known from the bacterial world, in which the phage *putL* and *putR* RNA structures interact with the exit channel of RNA Polymerase and block termination (21). We aimed to identify human RNA sequences that have a high affinity to RNA polymerase II and hypothesized that they could interact with and reprogram the transcription machinery. We performed a genomic SELEX experiment using the human genomic DNA as template for RNA sequences and

RNA Pol II as bait. We obtained a collection of RNA polymerase binding APtamers termed RAPs and analyzed their capacity to interfere with transcription in *cis*. We focused on one of the most highly enriched SELEX sequences derived from ACRO1 satellites and showed that ACRO1-derived RAPs are potent self-silencing elements.

Results

Genomic SELEX identifies Pol II-binding aptamers encoded in the human genome

We constructed an RNA library (18) representing the human genome in short (30-400 nt) transcripts and screened it for high-affinity binding to a purified complete Pol II 12-subunit complex from *Saccharomyces cerevisiae* since human Pol II could not be obtained in sufficient purity and quantity. Due to the high degree of conservation of the enzyme (22) and the fact that murine B2 RNA is able to bind to the *S. cerevisiae* Pol II core (23), we assumed that the binding sites for other RNAs might also be conserved. In the course of the SELEX procedure (Fig. 1A), Pol II-binding RNAs started to enrich in the 4th cycle (Fig. 1B) showing that the vast majority of RNAs in the starting pool do not bind to Pol II. We enforced higher stringency in the 6th and 7th cycles by lowering the protein concentration, thereby increasing the RNA-to-protein ratio in order to select sequences that bind in the low nanomolar range. We selected 200 clones from the 7th cycle for Sanger-sequencing which resulted in 74 individual RNAs. We validated the selection by showing that a set of exemplary RNAs from the 7th SELEX cycle are expressed (SI Fig. S1A), bind human Pol II *in vitro* (SI Fig. S1B) and can be co-immunoprecipitated with Pol II from HeLa lysates (Fig, 1C). The predominant RNA species among the 200 individually cloned aptamers were derived from repeat regions, such as LINEs, SINEs and satellites. These findings show that the successfully selected RNAs bind to Pol II in their natural context. Binding of total RNA from the 7th cycle pool to purified human Pol II can be partially outcompeted by B2 RNA, thus, a fraction of RAPs presumably interacts with the Pol II active site (SI Fig. S1C).

RAPs are found throughout the human genome, most notably in repeat regions

Although the selection procedure resulted in the successful isolation of RNAs binding to Pol II, no significantly enriched sequence was observed in the small sample of 200 clones, suggesting that the pool from the 7th cycle contained many diverse sequences. Therefore, we subjected this enriched pool to deep sequencing and computational analysis (**Figure 2**). A database was established to better access the outcome of the selection (http://alu.abc.univie.ac.at/pbe; username: renee; pw: pbepbepbe), which links all sequences to their genomic regions displayed in a GBROWSE instance (24). Enriched RNAs were mapped uniquely or multiple times to the genome. The unique hits were enriched in genic and intergenic regions, in sense as well as antisense orientation relative to the coding strand. The most prominently enriched RAP 5765 maps to the sense strand of intron 13 of the *MARK4* gene on chromosome 19 (Table 1). The majority of sequences, however, mapped to repeat regions and their enrichment was normalized according to their frequency in the human genome (Table 2). The enriched RNAs did not contain one single dominant sequence or structural motif, suggesting that Pol II can bind a variety of diverse RNA molecules. Generally, RAPs were more CA-rich than expected by chance (SI Fig. S2) and the highest enrichment score among the repeats was reached by $(\text{CAC}^{A}/_{T}{}^{C}/_{A})_n$ simple repeats and the ACRO1 family of satellites.

ACRO1 satellites

The ACRO1 consensus repeat unit is 147 bp long and occurs as 1.3-2.4 kb and 256 bp long arrays within a 6 kb higher-order repeat structure containing portions of LINEs, LTRs and DNA transposons. We termed these higher-order repeats “ACREs” for ACRO1-containing repeat elements (Fig. 3A and SI Fig. S3A). While ACREs are partially or fully conserved among all sequenced primates (SI Fig. S3B), no non-primate organism was found to carry a homologue of the ACRO1 repeat. ACRO1 satellites are moderately abundant tandem paralogue repeat elements clustered in the pericentromeric region of chromosome 4 and dispersed on chromosomes 1, 2, 19 and 21 (Fig. 3B and C). Many ACRO1 satellites have been mapped by FISH to chromosome 3 and to the acrocentric chromosomes 13, 14, 15, and 22 (25). However, these regions have not been annotated yet, indicating that many, if not most, ACREs are not represented in the current build of the human genome. Figure 3D shows SELEX read stacks mapping to the ACRO1 consensus unit defining the Pol II-binding region. These read stacks cover the ACRO1 RAPs, which are not individual *bona fide* transcripts, but rather domains within longer RNAs with Pol II-binding potential. We were unable to detect stable transcripts derived from ACRO1 satellites in HeLa cells. Nevertheless, ACRO1 satellites have been reported to be expressed at very low levels in several epithelial cancers (26).

“We noticed that ACRO1 satellites are over-represented in codons for amino acids also present in the Pol II subunit 1 CTD, especially proline, serine and threonine.”?. When translating the ACRO1 consensus sequence in *silico* into protein and aligning it with no gaps with the Pol II CTD, 23 out of 49 amino acids are identical and most convincingly also reflect the repetitive nature of the heptapetptide repeat (Fig. 3E). Furthermore, the ACRO1 RAPs harbor part of the sequence previously identified in a random SELEX experiment that binds to the Pol II CTD with an estimated K_D of 600 nM (38) This is reminiscent of the stereochemical hypothesis of genetic code origin that suggests that the code evolved in part from direct binding preferences between amino acids and their codons (27–29). Recently, we have extended this hypothesis to suggest that proteins, especially if unstructured, might in general bind specifically to RNAs that share codon composition with their mRNAs (30–32). We therefore also translated all the enriched human RAPs into amino acids in all three 5’→3’ reading frames and surprisingly found a strong bias for amino acids proline, serine and threonine, which are present in the Pol II CTD heptapeptide repeat YSPTSPS (Fig. 3F).

LINE1 retrotransposons are rich in RAPs

Another class of repeats prominent in our selection were the LINE elements, which was especially interesting because they had previously been reported to disrupt their own expression (33). There are multiple RAPs located within the 4 kb LINE1 ORF2 sequence (Fig. 4A). LINEs were shown to inhibit transcription when introduced into a reporter construct

(Fig. 4B) and transfected into HeLa cells (33). In this context, it had not been possible to narrow down the sequences responsible for disruption of transcription, though the effect was clearly dependent on the length of the LINE sequence. These results possibly indicate that LINEs contain sequences reducing their expression to avoid active invasion and damage of the genome caused by retrotransposition. The fact that RAPs were especially enriched in active full-length LINEs supports this hypothesis (SI Fig. S4). We repeated the above-mentioned experiments and analyzed the role of RAPs in LINE silencing. As can be seen in Figure 4C, the presence of ORF2 abrogated transcription of the reporter gene (L1), and elimination of the flanking RAPs led to a partial recovery (L1BS). These results corroborate the notion that sequences within the LINE1 ORF2 interfere with transcription. The fact that these sequences were enriched in the SELEX experiment suggests that the silencing is mediated by RNA-Pol II interaction.

RAPs disrupt transcription in *cis*

Encouraged by this observation, we used the same system to test whether highly enriched RAPs, such as ACRO1 repeats and RAP 5765, could also lead to transcriptional disruption. Single RAPs inserted into the *GFP-LacZ* reporter system had no or only a minor effect on steady-state RNA levels (Fig. 4D). However, insertion of multiple ACRO1 repeat units into the reporter resulted in a strong transcriptional disruption. A short insert of 0.3 kb containing two ACRO1-derived RAPs already had a visible effect, and ACRO1 insertion of 1.1 and 1.4 kb almost completely eliminated the RNA product (Fig. 4F and SI Fig. S5). When multiple RAPs of the highly enriched genic 5765 aptamer were cloned in tandem, they severely disrupted transcription of the GFP reporter and the number of RAPs correlated with the extent of transcriptional repression (Fig. 4E and SI Fig. S5). This down-regulating effect of the RAPs was alleviated when reverse complement sequences were used as controls confirming sequence and/or structural specificity and ruling out the possibility that a *trans*-acting DNA-binding factor constitutes a roadblock to transcription.

We further focused our analysis on ACRO1 satellites and asked whether the promoter has an impact on the transcriptional downregulation mediated by RAPs. Replacing the CMV with the alpha-globin promoter in the *GFP-LacZ* reporter resulted in similarly depleted GFP expression levels (Fig. 4G). In addition, RAPs did not have an effect on the cognate locus *in trans* (Fig. 4H). It is thus possible that RAPs either regulate their expression co-transcriptionally or affect the stability of the mature RNA.

To distinguish between post- and co-transcriptional regulation, we monitored transcript levels upstream and downstream of the ACRO1 insertion by RT-qPCR (Fig. 5A). Total RNA was isolated and RT-qPCR primer pairs were designed to flank the ACRO1 sequence. We compared transcript levels at three loci upstream and three loci downstream of the ACRO1 insert. The decrease of the downstream RNA levels in ACRO1-containing construct, but not in reverse complement (ORCA) or no-insert (-ins) controls, indicates that RNA production was compromised at the ACRO1 locus. We repeated the experiment with separated Poly(A)+ and Poly(A)- fractions of total RNA (SI Fig. S12). We reasoned that the Poly(A)- fraction contained incomplete products of ongoing transcription and could thus uncover true co-transcriptional regulatory events, whereas the Poly(A)+ fraction contained full-length RNAs that escaped the regulation (SI Fig. S13). Indeed, the RNA profile in the Poly(A)+ fraction was comparable between ACRO1 construct and controls, but the downstream RNA did strongly decrease in the Poly(A)- fraction indicating that RAPs have no impact on the fate of the mature full-length transcript (Fig. 5B and C). These results show that the RAP-mediated inhibition is co-transcriptional, spatially restricted to the vicinity of the RAP template and that RAP-containing RNAs are stable once fully transcribed.

To test whether individual RAPs exert transcriptional repression in their endogenous context, we took the same approach to quantify transcript levels upstream and downstream of the most highly enriched genic RAP 5765 within the *MARK4* gene intron 13 (Fig. 5D). The results show a moderate decrease of downstream RNA indicating that even a single RAP can modulate transcriptional output in its endogenous context.

Discussion

Genomic SELEX is a powerful tool to extract silencing information from genomes

Transcription is a central process in cellular life, and its regulation occurs at multiple levels. The number of proteins known to interfere with this process is large. However, we reasoned that RNA might also be a potent regulator of transcription, and that especially the nascent RNA might contain signals that communicate with the transcription machinery. Genomic SELEX using the complete human DNA as source of RNAs and purified Pol II as bait is a powerful approach in this context because this procedure is unbiased and also includes DNA sequences that are expressed at a very low level or not at all in *vivo* (34). Because the human genome contains a high percentage of non-transcribed repetitive elements, we hypothesized that their silencing could be linked to their underlying RNAs of origin. Many retroelements in the human genome are derived from retrotransposed functional non-coding RNAs (35, 36). LINE retrotransposons, for example, are known to disrupt transcription in *cis* (33). Using this unbiased approach, we identified a large number of human RNA Pol II aptamers (RAPs) encoded throughout the human genome both in unique and, most prominently, in repetitive elements. RAPs do not constitute a single RNA family with one common motif or structure, although they are generally CA-rich. RAPs are very diverse suggesting that there are many different ways that RNAs can interact with Pol II, perhaps not surprisingly, as the Pol II complex is very large and contains many potential interaction sites on its surface and in its active site. The yeast Pol II active center has been shown to be very flexible and able to accommodate quite large RNAs (23), and a recent cryo-EM analysis of the mammalian Pol II showed high degree of similarity between the two enzymes (37).

ACRO1 satellites are derived from Pol II CTD and might shed light on the origin of the genetic code

Recently, we have demonstrated that nucleobase-density profiles of typical mRNA coding sequences match closely the nucleobase-affinity profiles of their cognate proteins, with anti-matching seen only in the case of adenine profiles (30–32). This finding generalized the stereochemical hypothesis of the origin of the genetic code (27–29), but also suggested that proteins, especially if unstructured, may bind in a co-aligned, complementary fashion to their cognate mRNAs, but also other RNAs that share features with their mRNAs (30–32). In direct support of this proposal, here we could show that, remarkably, ACRO1 satellites encode a protein sequence similar to the Pol II CTD and that, in addition, the RAPs are enriched in codons for the amino acids proline, serine and threonine, which feature heavily in the Pol II CTD sequence. This, in turn, allows us now to propose that the mechanism of RAP binding to Pol II may in part involve direct interactions between the codons contained in RAPs with their corresponding amino acids in Pol II and, especially, its CTD. Further analysis of these exciting possibilities is a topic of our current work and will be published elsewhere.

Furthermore, it is possible that ACRO1 repeats are evolutionarily derived from the Pol II CTD to introduce an additional level of transcription regulation close to centromeres. ACRO1 elements are moderately abundant in the human genome and are mainly located in pericentromeric regions which are transcriptionally inactive. Their mobility could have been provided by the mobile elements contained within the ACREs (SI Fig. S3)**.** This must be a very recent acquisition as they can only be found in primates.

RAPs represent a novel type of regulatory RNA signals

In this work, we identified a novel level of transcription regulation by showing that RNA signals on the nascent RNA can interfere with the transcribing Pol II in *cis*, abrogating transcription. It had already elegantly been shown that the secondary structure of the nascent RNA affects the rate of Pol II transcription in *vitro* by inhibiting backtracking and thus preventing the polymerase to escape from pausing (17). Interaction between Pol II CTD with mRNA has also been reported to suppress transcription-coupled 3'-end processing (38). RAPs are RNA sequences that were enriched in a SELEX procedure due to their virtue of binding to Pol II. They are not *bona fide* transcripts but rather domains within potentially expressed RNAs that convey Pol II binding capacity to their host transcripts. In the context of our experiments, RAPs are part of the nascent transcript interacting with Pol II in *cis* during transcription. We observed that their effect on transcription is additive and that the more RAPs are present on the nascent RNA the stronger the inhibitory effect. Most importantly, the inhibitory effect is co-transcriptional. Once the RNA is fully transcribed, RAPs have no impact either on transcription or on the stability of the transcript. Based on these observations, we hypothesize that the nascent RNA can cross-talk to Pol II via many potential interaction sites on its surface, or via the CTD, and thereby disrupt transcription (Fig. 5E).

Recently, circular intronic long noncoding RNAs were shown to accumulate at the site of transcription, associate with the elongating RNA polymerase and act as positive regulators of transcription (39). Here we add another layer of transcriptional regulation that involves *cis*-acting sequences within the nascent transcript that affect transcription elongation. This might be an essential self-regulatory strategy for repeat elements to stay silent, enabling their survival in the genome during evolution. In addition, we hypothesize that RAP-mediated control of transcription might play a role in gene-regulatory processes, which depend on the rate of Pol II progression, such as alternative splicing and termination (40). Indeed, several RAPs map downstream of alternative splice sites and alternative polyadenylation sites.

RAP-mediated transcription termination is a conserved phenomenon from bacteria to yeast to humans.

In this work, we have presented evidence that Pol II can "sense" the nature of transcripts by means of direct interaction and that some RNA sequences encoded in the human genome have the potential to interfere with their own transcription *in cis*. We propose a novel mode of transcriptional control in human cells, wherein the nascent RNA binds to the transcribing Pol II making it elongation-incompetent (Fig. 5E). Similar screens have been performed for the *E. coli* genome and the bacterial RNA polymerase and the yeast *S. cerevisiae* genome and yeast Pol II (Sedlyarova et al, in preparation; Klopf et al, accompanying manuscript). *E. coli* RAPs cause Rho-dependent premature transcription termination by uncoupling translation and transcription. Like human RAPs, yeast RAPs induce premature transcription termination demonstrating that RAP-mediated transcription interference is a conserved phenomenon. A cross-talk between the nascent RNA and the transcription machinery could provide the primary signal that determines the fate of transcripts.

Materials & Methods

Library construction and Genomic SELEX

The genomic library was created as described previously [27,28], with human genomic DNA purchased from Sigma (CAS number 9007-49-2) as template. After transcribing the genomic library into RNA, the RNA pool was bound to Pol II of *S. cerevisiae* in an *in vitro* binding reaction as described in ref (17). For the 1st-5th cycles, RNA was added at 1 µM and protein at 100 nM. To increase stringency and competition, RNA was added at 1 µM and protein at 10 nM for the 6th and 7th cycles. The binding buffer contained 10 mM HEPES pH 7.25, 40 mM NH_4SO_4, 10 µM $ZnCl_2$, 1 mM KCl, 10 mM DTT, 5 % glycerol and 10 mM $MgCl_2$.

Co-immunoprecipitation

HeLa cells grown in 10 cm dishes were harvested at 80 % confluence with 1 ml lysis buffer (10 mM HEPES pH 7.0, 100 mM KCl, 5 mM $MgCl_2$, 0.5 % Nonidet P-40, 1 mM DTT, 100 U/ml RNAse inhibitor (Promega), 2 mM vanadyl ribonucleoside complexes solution, 25 µl/ml protease inhibitor cocktail for mammalian tissues) per 10 cm^{-1} and removed from the dish with a cell scraper. After 10 min on ice cells were centrifuged at 4 °C, 1000 × *g*. Whole cell extracts were prepared for co-IP as described[28]. RNA purified from the immunoprecipitates and input RNA were analysed by RT-PCR with the Qiagen RT-PCR kit using primers specific for the different RNAs.

Antibodies

Pol II and DNA polymerase antibodies were purchased from Abcam (ab817/ab5408 and ab3181, respectively). Pol II-antibody recognizes the phosphorylated as well as the unphosphorylated form of Pol II. The concentration of antibodies used for immunoprecipitations was 2 µl/ml.

Transfection, microscopy and RNA preparation

HeLa cells were grown to 70-90 % confluence and transfected with 0.4 mg of plasmid per cm^2 of culture dish using Lipofectamine 2000 (Invitrogen) according to manufacturer's instructions. After 24 h, fluorescence was monitored with AxioObserver Z1 microscope coupled to AxioCam MRm (Carl Zeiss MicroImaging) and RNA was extracted with TRI Reagent (Sigma).

Northern blot

Total RNA was separated on a 0.8 % agarose gel containing 6.7 % formaldehyde, capillary-blotted onto a Hybond-XL membrane (GE Healthcare) and UV-crosslinked. ^{32}P-labeled DNA probe was hybridized in ULTRAhyb-Oligo Buffer (Ambion) at 42 °C overnight. The probe was 5'-labeled with T4 PNK (NEB).

Flow cytometry

GFP-positive cells were quantified by FACSCalibur (BD Biosciences) and data were analyzed in Cyflogic (CyFlo Ltd, Finland) and SPSS (IBM) software. From each sample, fluorescence of 10,000 cells was measured and only GFP-positive events, as determined by mock-transfected cell fluorescence, were taken into account.

Poly(A) fractionation

150 pmol biotinylated Oligo(dT) (Promega) was bound for 10 min at room temperature to 0.6 ml MagneSphere® magnetic beads (Promega) prepared according to manufacturer's instructions. 80 mg of total RNA was denatured at 65 °C, 10 min, chilled on ice for 5 min and mixed with Oligo(dT)-beads solution. After 10 min incubation at room temperature the beads were washed six times and Poly(A)+ RNA was eluted according to manufacturer's instructions. Before washing of the beads, the first supernatant was taken as Poly(A)- RNA. Both fractions were ethanol-precipitated.

RT-PCR and RT-qPCR

2 mg of total RNA or 200 ng of Poly(A)-fractionated RNA was denatured with 200 pmol of random nonamers (Sigma) at 70 °C for 10 min. The reaction was split in two, one without reverse transcriptase as a control. RT was performed at 45 °C for 90 min using OmniScript (Qiagen). 1/40 of the total reaction was used for PCR and approximately 1/30 was used per qPCR well. qPCR was performed in Mastercycler® realplex (Eppendorf) with HOT FIREPol® qPCR Mix (Medibena) and primers specified in Supplementary Table S1. Transfection was controlled for by normalizing expression values to neo and subsequently all amplicons were normalized to GFP 1.

Accession numbers: The ACRO1 sequence used in the reporter assay has been deposited in the Genbank with the number GenBank KF726396. The sequences presented in this work are filed in a specific portal containing the RAP database and the link to the human genome (http://alu.abc.univie.ac.at/pbe.

Supplementary Information **is contained in the attached word file.**

Acknowledgments We thank Arndt von Haeseler and Heiko Schmidt for bioinformatical advice and access to the CIBIV computer cluster. We thank Jef Boeke for the gift of the reporter plasmid. We thank the Goodrich and Kugel labs for advice and help with the initial characterization of RAPs. We thank Eva Klopf for helping with the preparation of the manuscript. We thank the members of the Schroeder lab for discussions and for critically commenting the manuscript. Supported by the Austrian Science Fund (FWF grant SFB F4308) and by the Austrian GENAU Project "RNomics" D1042. Patrick Cramer was supported by the Deutsche Forschungsgemeinschaft (SFB646, SFB960, SFB1064, CIPSM, NIM, QBM, GRK1721), an ERC Advanced Grant, the Jung-Stiftung and the Vallee Foundation. Bojan Zagrovic was supported by an ERC Starting Grant (Nr. 279408).

Author Contributions RS designed the experiments and supervised the study. JB, FvP, AW and MR performed the experiments. JS and KMK assisted with the experiments. B. Zimmermann, DC, MZ, TF and B. Zagrovic performed the bioinformatic analysis. FB and PC purified Pol II. FvP, AW, B. Zimmermann, MR, B. Zagrovic and RS wrote the paper. JLB, FvP, AW and B. Zimmermann contributed equally to this work.

The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to renee.schroeder@univie.ac.at.

References:

1. Adelman K, Lis JT (2012) Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet* 13(10):720–31.
2. Churchman LS, Weissman JS (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* 469(7330):368–373.
3. Zhou Q, Li T, Price DH (2012) RNA Polymerase II Elongation Control. *Annu Rev Biochem* 81:119–143.
4. Jonkers I, Lis JT (2015) Getting up to speed with transcription elongation by RNA polymerase II. *Nat Rev Mol Cell Biol* 16(3):167–177.
5. Rutenberg-Schoenberg M, Sexton AN, Simon MD (2016) The Properties of Long Noncoding RNAs That Regulate Chromatin. *Annu Rev Genom Hum Genet* 17(9):1–926.
6. Wassarman KM, Storz G (2000) 6S RNA regulates E. coli RNA polymerase activity. *Cell* 101(6):613–623.
7. Espinoza C a, Allen T a, Hieb AR, Kugel JF, Goodrich J a (2004) B2 RNA binds directly to RNA polymerase II to repress transcript synthesis. *Nat Struct Mol Biol* 11(9):822–829.
8. Mariner PD, et al. (2008) Human Alu RNA Is a Modular Transacting Repressor of mRNA Transcription during Heat Shock. *Mol Cell* 29(4):499–509.
9. Thomas M, et al. (1997) Selective targeting and inhibition of yeast RNA polymerase II by RNA aptamers. *J Biol Chem* 272(44):27980–6.
10. Wassarman KM (2012) 6S RNA: A regulator of transcription. *Regulatory RNAs in Prokaryotes*, pp 109–130.
11. Liu W man, Chu W ming, Choudary P V., Schmid CW (1995) Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. *Nucleic Acids Res* 23(10):1758–1765.
12. Serganov A, Nudler E (2013) A decade of riboswitches. *Cell* 152(1–2):17–24.

13. Cramer P, et al. (2008) Structure of eukaryotic RNA polymerases. *Annu Rev Biophys* 37:337–352.

14. Windbichler N, von Pelchrzim F, Mayer O, Csaszar E, Schroeder R (2008) Isolation of small RNA-binding proteins from E. coli: evidence for frequent interaction of RNAs with RNA polymerase. *RNA Biol* 5(1):30–40.

15. Wettich A, Biebricher CK (2001) RNA species that replicate with DNA-dependent RNA polymerase from Escherichia coli. *Biochemistry* 40(11):3308–3315.

16. Schulz D, et al. (2013) Transcriptome surveillance by selective termination of noncoding RNA synthesis. TL - 155. *Cell* 155 VN-(5):1075–1087.

17. Zamft B, Bintu L, Ishibashi T, Bustamante C (2012) Nascent RNA structure modulates the transcriptional dynamics of RNA polymerases. *Proc Natl Acad Sci* 109(23):8948–8953.

18. Singer BS, Shtatland T, Brown D, Gold L (1997) Libraries for genomic SELEX. *Nucleic Acids Res* 25(4):781–786.

19. Sigova AA, et al. (2015) Transcription factor trapping by RNA in gene regulatory elements. *Science (80-)* 350(6263):978–981.

20. Barrandon C, Spiluttini B, Bensaude O (2008) Non-coding RNAs regulating the transcriptional machinery. *Biol Cell* 100(2):83–95.

21. Santangelo TJ, Artsimovitch I (2011) Termination and antitermination: RNA polymerase runs a stop sign. *Nat Rev Microbiol* 9(5):319–329.

22. Cramer P, et al. (2000) Architecture of RNA polymerase II and implications for the transcription mechanism. *Sci (New York, NY)* 288(5466):640–649.

23. Kettenberger H, et al. (2006) Structure of an RNA polymerase II-RNA inhibitor complex elucidates transcription regulation by noncoding RNAs. *Nat Struct Mol Biol* 13(1):44–48.

24. Stein LD, et al. (2002) The generic genome browser: A building block for a model organism system database. *Genome Res* 12(10):1599–1610.

25. Warburton PE, et al. (2008) Analysis of the largest tandemly repeated DNA families in

the human genome. *BMC Genomics* 9:533.

26. Ting DT, et al. (2011) *Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers.* doi:10.1126/science.1200801.

27. Woese CR, Dugre DH, Saxinger WC, Dugre SA (1966) The molecular basis for the genetic code. *Proc Natl Acad Sci U S A* 55(4):966–74.

28. Koonin E V, Novozhilov AS (2009) Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* 61(2):99–111.

29. Yarus M, Widmann JJ, Knight R (2009) RNA-amino acid binding: A stereochemical era for the genetic code. *J Mol Evol* 69(5):406–429.

30. Hlevnjak M, Polyansky AA, Zagrovic B (2012) Sequence signatures of direct complementarity between mRNAs and cognate proteins on multiple levels. *Nucleic Acids Res* 40(18):8874–8882.

31. Polyansky AA, Zagrovic B (2013) Evidence of direct complementary interactions between messenger RNAs and their cognate proteins. *Nucleic Acids Res* 41(18):8434–8443.

32. Polyansky AA, Hlevnjak M, Zagrovic B (2013) Proteome-wide analysis reveals clues of complementary interactions between mRNAs and their cognate proteins as the physicochemical foundation of the genetic code. *RNA Biol* 10(8):1248–1254.

33. Han JS, Szak ST, Boeke JD (2004) Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* 429(6989):268–274.

34. Lorenz C, von Pelchrzim F, Schroeder R (2006) Genomic systematic evolution of ligands by exponential enrichment (Genomic SELEX) for the identification of protein-binding RNAs independent of their expression levels. *Nat Protoc* 1(5):2204–2212.

35. Brosius J (2003) How significant is 98.5% “junk” in mammalian genomes? *Bioinformatics* 19 Suppl 2:II35.

36. Matylla-Kulinska K, Tafer H, Weiss A, Schroeder R (2014) Functional repeat-derived RNAs often originate from retrotransposon-propagated ncRNAs. *Wiley Interdiscip Rev*

RNA 5(5):591–600.

37. Bernecky C, Herzog F, Baumeister W, Plitzko JM, Cramer P (2016) Structure of transcribing mammalian RNA polymerase II. *Nature* 529(7587):551–4.

38. Kaneko S, Manley JL (2005) The mammalian RNA polymerase II C-terminal domain interacts with RNA to suppress transcription-coupled 3??? end formation. *Mol Cell* 20(1):91–103.

39. Zhang Y, et al. (2013) Circular Intronic Long Noncoding RNAs. *Mol Cell* 51(6):792–806.

40. De La Mata M, et al. (2003) A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell* 12(2):525–532.

Figure Legends

Figure 1. Genomic SELEX for RNA polymerase II-binding elements (RAPs). (A) The initial human DNA library was *in vitro* transcribed and the resulting RNA pool was bound to the highly purified yeast Pol II. Protein-bound RNAs were retained on the filter and non-binding RNAs were discarded. Selected RNAs were eluted from the filter and reverse transcribed into DNA. After PCR amplification, the resulting cDNA pool was subjected to another cycle of SELEX. After sufficient enrichment the pool can be either cloned and individually sequenced or subjected to parallel sequencing (17). **(B)** Enrichment of Pol II-bound human RNAs is shown for each SELEX cycle. The percentage of the recovered RNA was calculated in relation to the input RNA (red bars). In cycles 1-5 a 10:1 molar excess of RNA over protein was used, whereas in cycle 6 and 7, the RNA to protein ratio is increased to 100:1. BSA was used as a negative control (black bars). **(C)** To validate binding of selected RNAs to human Pol II *in vivo*, lysate of heat-shocked HeLa cells was co-immunoprecipitated with RNA Pol II- or DNA polymerase-specific antibodies and subjected to RT-PCR. 5S and Hsf1 are abundant cellular RNAs used as control that were not enriched by SELEX.

Figure 2. Schematic of the workflow for the selection and the analysis of RAPs. (A) A human RNA library was constructed and selected for RNAs binding to Pol II. The enriched pool from the 7th cycle was subjected to 454 sequencing and later the pool from the 6th cycle was Solexa sequenced. The obtained reads were filtered, mapped to the human genome (hg18 and hg19) and annotated to contigs of 400 nt in length. **(B)** Top enriched RAP 5765. Typical read stacks mapped and annotated to the human genome and displayed as custom track in the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway). **(C)** Enrichment of RAPs in human ACRO1 satellites was weighed and normalized to their frequency in the genome (blue bars). Arrows represent individual sequenced reads. Many of the ACRO1-associated RAPs map to the ACRO-rich centromeric region of chromosome IV. Each arrow corresponds to one read, its direction indicating the sequence orientation compared to the reference genome (plus strand).

Figure 3. The structure and distribution of ACRO1 satellites. (A) ACRE (ACRO-containing repeat element) is a higher order repeat structure of 6 kb harboring the ACRO satellite array. **(B)** Organization of the ACRE cluster in the pericentromeric region of chromosome 4, the densest region of sequenced ACREs. Note that (A) shows consensus ACRE, not specifically ACRE 12. **(C)** ACREs were found on chromosomes 1, 2, 4, 19 and 21. **(D)** Sequence of ACRO1 consensus repeat unit and its SELEX enrichment profile. (E) Alignment of a translation of the consensus ACRO1 sequence with the human Pol II CTD (residue range given) with identical residues outlined in red. (F) Frequency of different codons in RAP sequences in all 5'->3' reading frames.

Figure 4. RAPs induce transcriptional silencing. (A) The LINE1 retrotransposon is illustrated here with the restriction sites “B” and “S” indicated (26). LINE1-associated RAPs from the 7th SELEX cycle were mapped to the consensus with at least 80 % identity. **(B)** Vector used to monitor *in vivo* expression of the reporter cassette (adapted from (26)). RAPs or control sequences were cloned between the *GFP* and the *LacZ* sequences or in case of L1 and L1BS in place of LacZ gene. **(C)-(H)** Northern blot analyses of total RNA extracted from HeLa cells transfected with various RAP-containing reporters show RNA levels of the reporter gene (gfp) and a transfection control (neo). The minor bands visible especially in (E), lanes 5-8, probably derive from unspecific hybridization to 28S rRNA (C) The cassettes contained empty GFP-LacZ fusion (-ins), LINE1 ORF2 (L1), its shortened version trimmed to the region between the “B” and “S” sites (L1BS), (D) Diverse RAP sequences cloned into the reporter system. Apart from the single ACRO unit (9258), which had about a 3-fold decrease compared to no-insert plasmid (-ins), none of the other RAPs had an effect on the transcript levels. B2 RNA, which interferes with transcription *in trans*, was tested as a control. (E) RAP 5765 cloned in tandem one to six times and six times in reverse complement (inv) (F) ACRO1 as 0.3 kb, 1.1 kb and full 1.4 kb elements and its reverse complement (ORCA). (G) Expression from the alpha-globin promoter. CMV promoter

that drives expression of the reporter cassette was replaced with alpha-globin promoter and transcription was monitored by Northern blot 24 hours after transfection. (H) To test whether presence of ACRO element affects reporter expression on a different plasmid, cassettes with empty GFP-LacZ fusion (-ins) and with full-length ACRO1 element (ACRO) were co-transfected into HeLa and expression was assessed by Northern blot.

Figure 5. Autoregulation of RAPs is co-transcriptional. (A-C) RT-qPCR quantification of six different amplicons along the reporter transcript. Total RNA was isolated from HeLa cells 24 h after transfection with vectors carrying no insert (blue lines), ACRO (green lines) or its reverse complement ORCA (orange lines) inserts. In (B) and (C) RNA was further fractionated according to the presence (+) or absence (-) of the Poly(A) tail. All values are plotted on a log scale relative to GFP 1, the 5'-most amplicon. Note different scale in (C). Error bars represent SEM of five (total RNA) and four (fractionated RNA) experiments. The positions of the amplicons are indicated by red bars below the panel. The reporter gene is a part of the vector from Figure 3B. **(D)** RT-qPCR quantification of four amplicons surrounding the endogenous RAP 5765. Distance of the amplicon from the RAP (in bp) is indicated. Values are plotted on a linear scale relative to amplicon -356. Error bars represent SEM of three experiments. **(E)** Model of transcriptional inhibition by RAPs. Pol II initiates at transcription start site (TSS) and continues into productive elongation. When RAPs are present on the nascent transcript, the RNA binds Pol II, either in the active site or elsewhere, rendering it elongation-incompetent. Presumably, the transcript then lacks a polyA signal and is eliminated from the cell. Note that the combined action of several RAPs might be needed for efficient regulation.

Figure 1

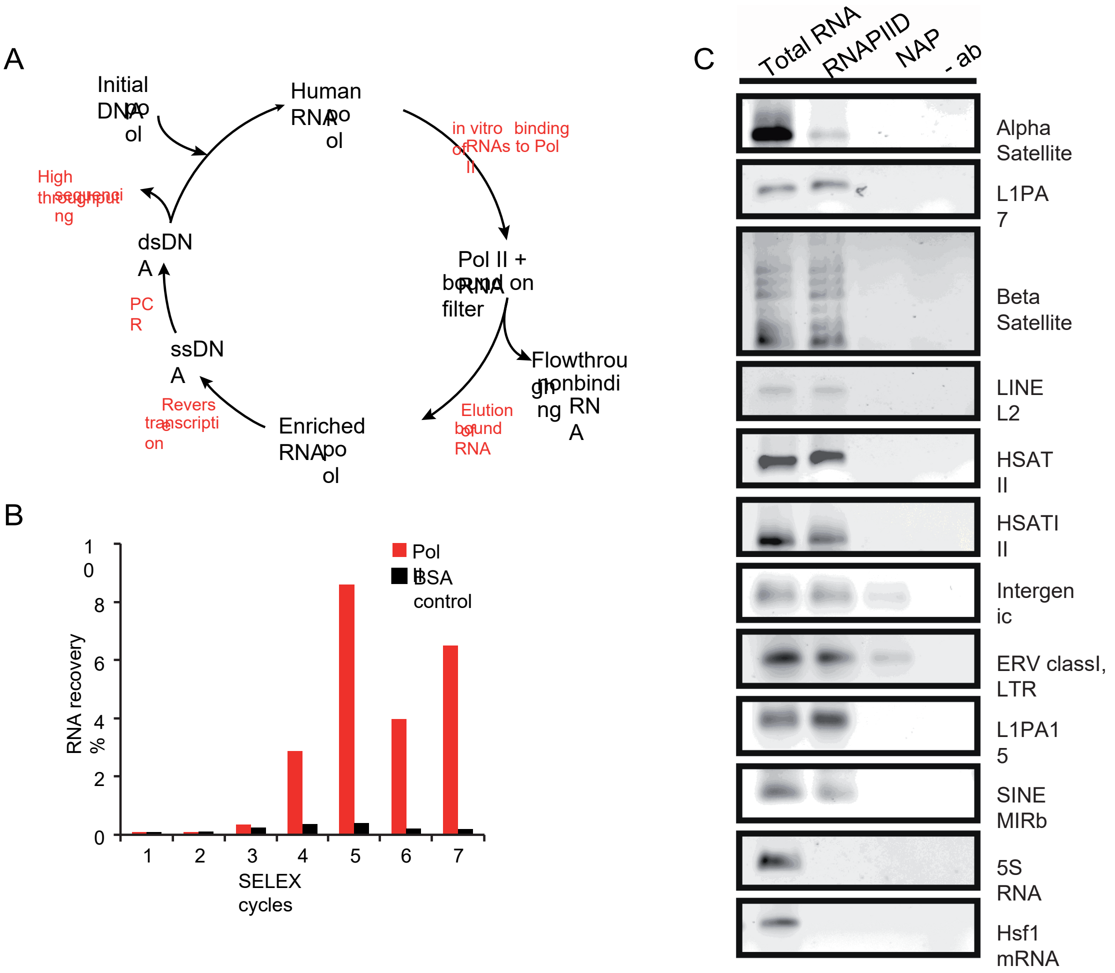

A
Initial DNA pool
Human RNA pool
in vitro binding of RNAs to Pol II
Pol II + RNA bound on filter
Flowthrough nonbinding RNA
Elution of bound RNA
Enriched RNA pool
Reverse transcription
ssDNA
PCR
dsDNA
High throughput sequencing
B
10
8
6
4
2
0
% RNA recovery
1
2
3
4
5
6
7
SELEX cycles
🟥 Pol II
■ BSA control
C
Total RNA
RNAPIID
NAP
- ab
Alpha Satellite
L1PA7
Beta Satellite
LINE L2
HSAT II
HSATIII
Intergenic
ERV classI, LTR
L1PA15
SINE MIRb
5S RNA
Hsf1 mRNA

Figure 2

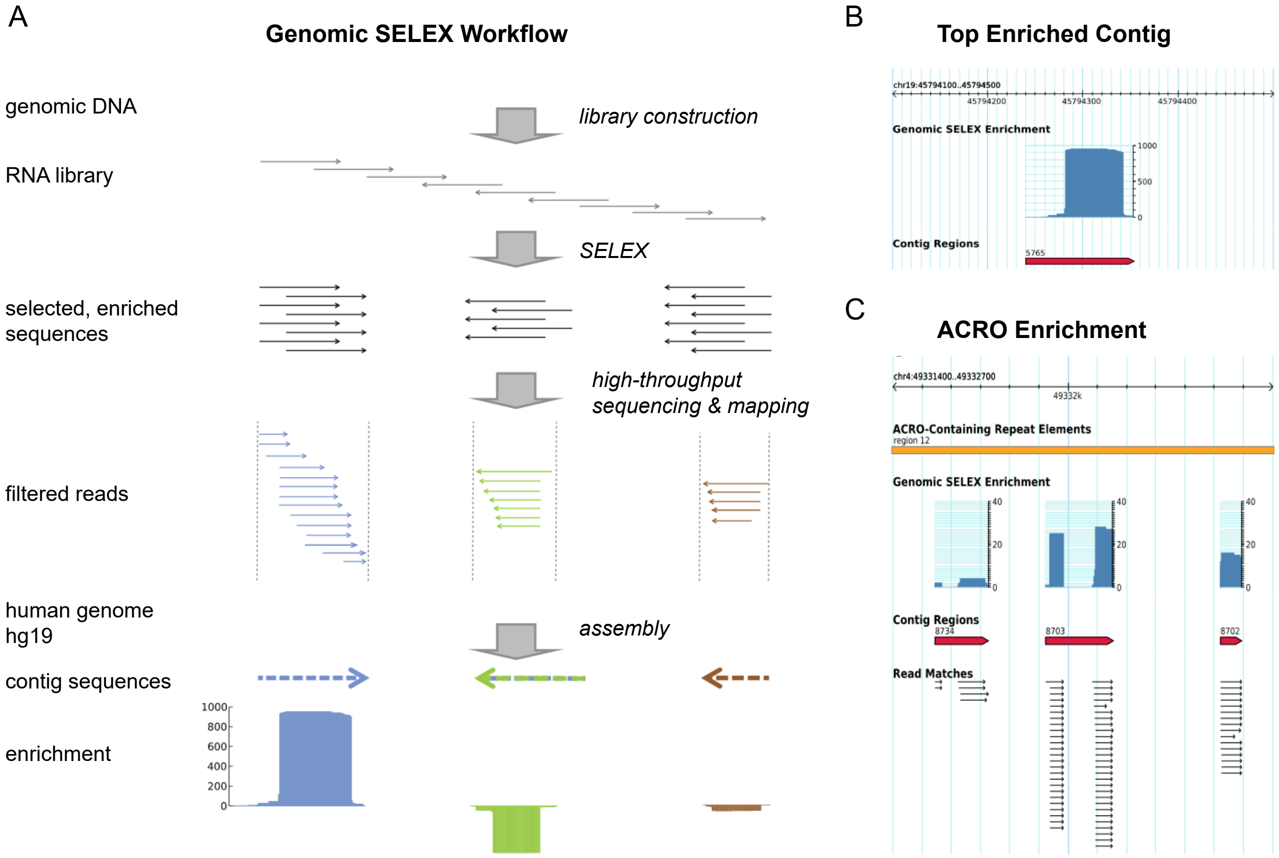

A
Genomic SELEX Workflow
genomic DNA
library construction
RNA library
SELEX
selected, enriched sequences
high-throughput sequencing & mapping
filtered reads
human genome hg19
assembly
contig sequences
enrichment
1000
800
600
400
200
0
B
Top Enriched Contig
chr19:45794100..45794500
45794200
45794300
45794400
Genomic SELEX Enrichment
1000
500
0
Contig Regions
5765
C
ACRO Enrichment
chr4:49331400..49332700
49332k
ACRO-Containing Repeat Elements
region 12
Genomic SELEX Enrichment
40
20
0
Contig Regions
8734
8703
8702
Read Matches

Figure 3

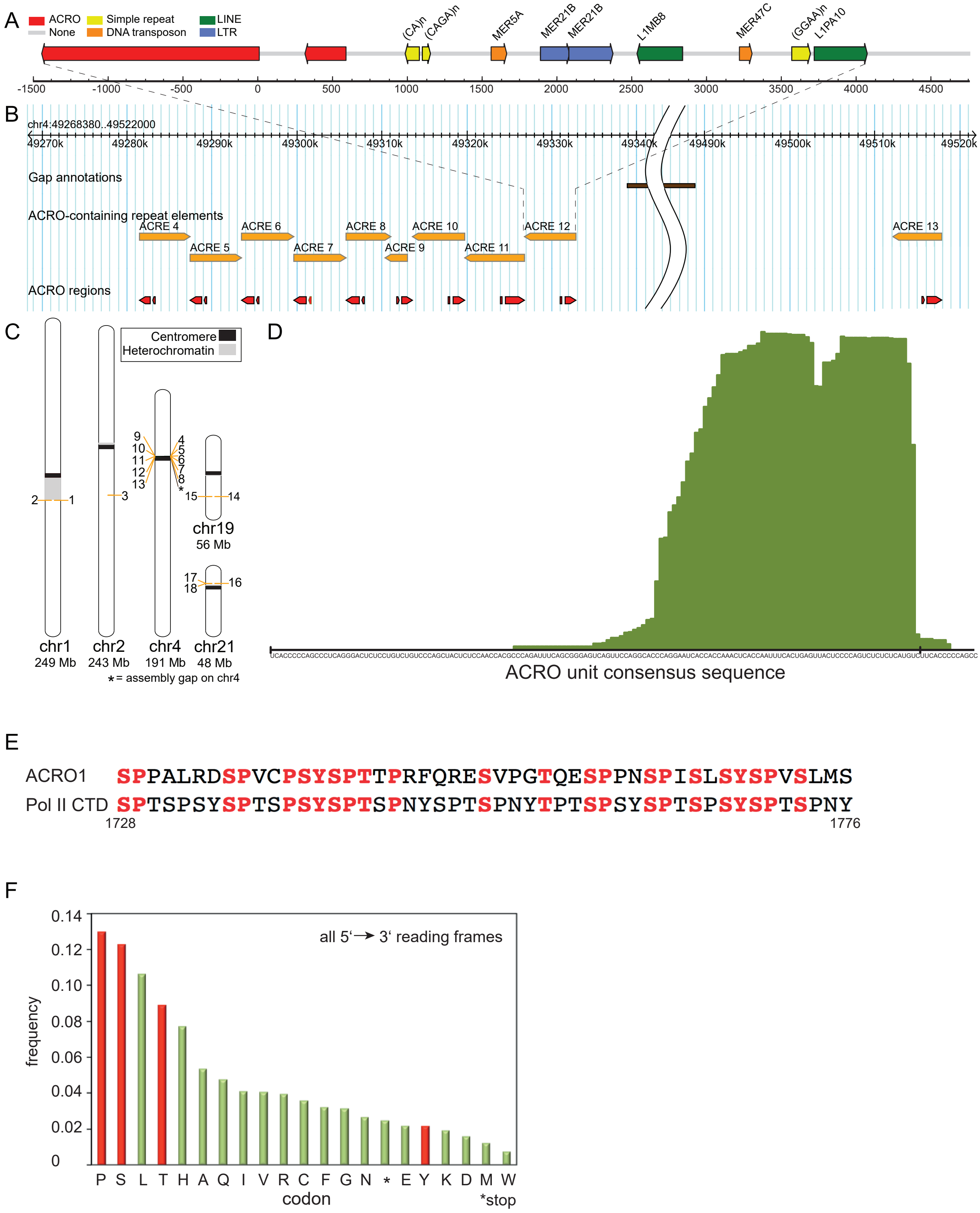

A
ACRO
None
Simple repeat
DNA transposon
LINE
LTR
(CA)n
(CAGA)n
MER5A
MER21B
MER21B
L1MB8
MER47C
(GGAA)n
L1PA10
-1500
-1000
-500
0
500
1000
1500
2000
2500
3000
3500
4000
4500
B
chr4:49268380..49522000
49270k
49280k
49290k
49300k
49310k
49320k
49330k
49340k
49490k
49500k
49510k
49520k
Gap annotations
ACRO-containing repeat elements
ACRE 4
ACRE 5
ACRE 6
ACRE 7
ACRE 8
ACRE 9
ACRE 10
ACRE 11
ACRE 12
ACRE 13
ACRO regions
C
Centromere
Heterochromatin
2
1
3
9
10
11
12
13
4
5
6
7
8
*
15
14
17
18
16
chr1
249 Mb
chr2
243 Mb
chr4
191 Mb
chr19
56 Mb
chr21
48 Mb
* = assembly gap on chr4
D
UCACCCCCAGCCCUCAGGGACUCUCCUGUCUGUCCCAGCUACUCUCCAACCACGCCCAGAUUUCAGCGGAGUCAGUUCCAGGCACCCAGGAAUCACCACCAAACUCACCAAUUUCACUGAGUUACUCCCCAGUCUCUCUCAUGUCUUCACCCCCAGCC
ACRO unit consensus sequence
E
ACRO1 SPPALRDSPVCPSYSPTTPRFQRESVPGTQESPPNSPISLSYSPVSLMS
Pol II CTD SPTSPSYSPTSPSYSPTSPNYSPTSPNYTPTSPSYSPTSPSYSPTSPNY
1728
1776
F
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0
frequency
all 5'→ 3' reading frames
P S L T H A Q I V R C F G N * E Y K D M W
codon
*stop

Figure 4

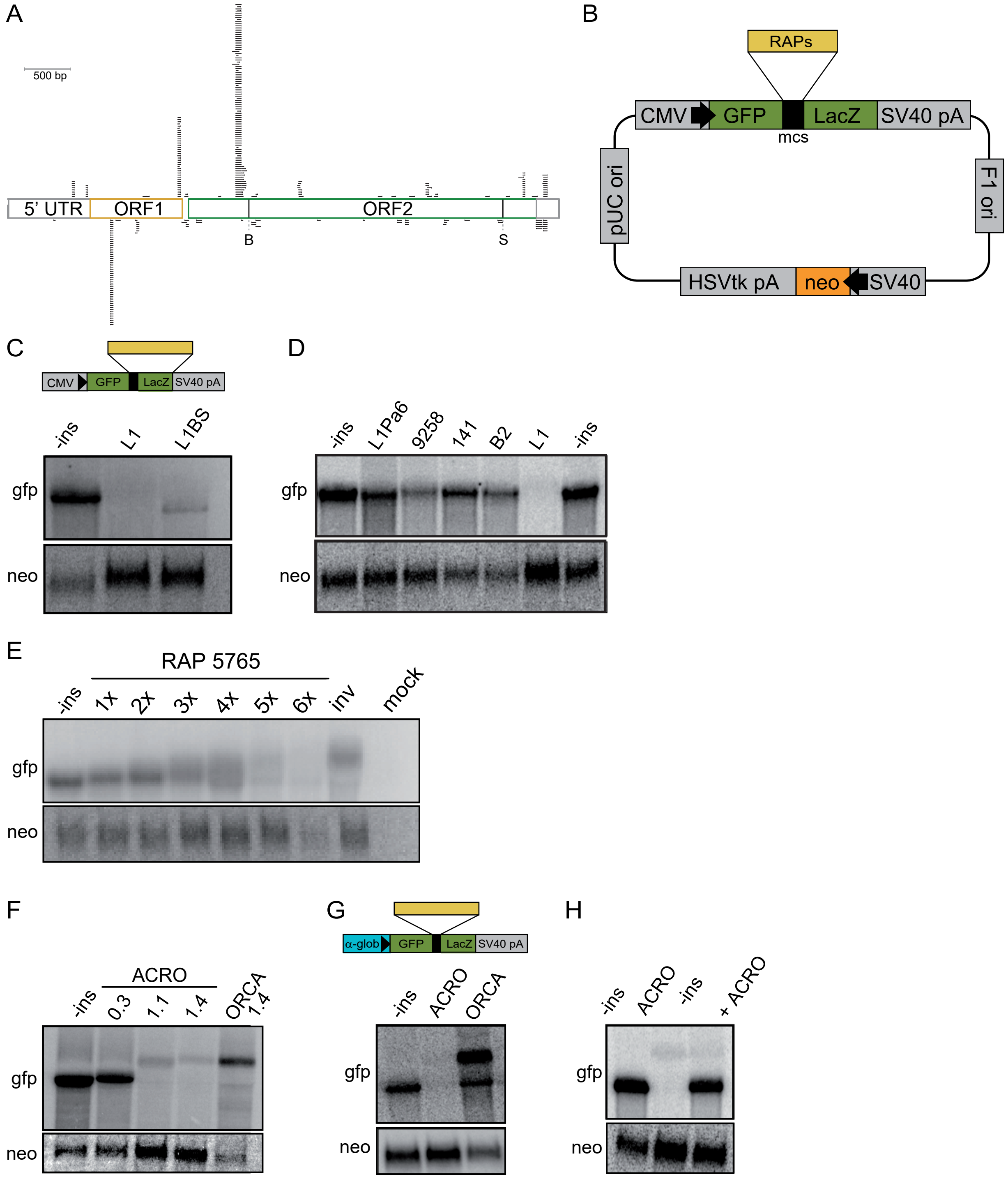
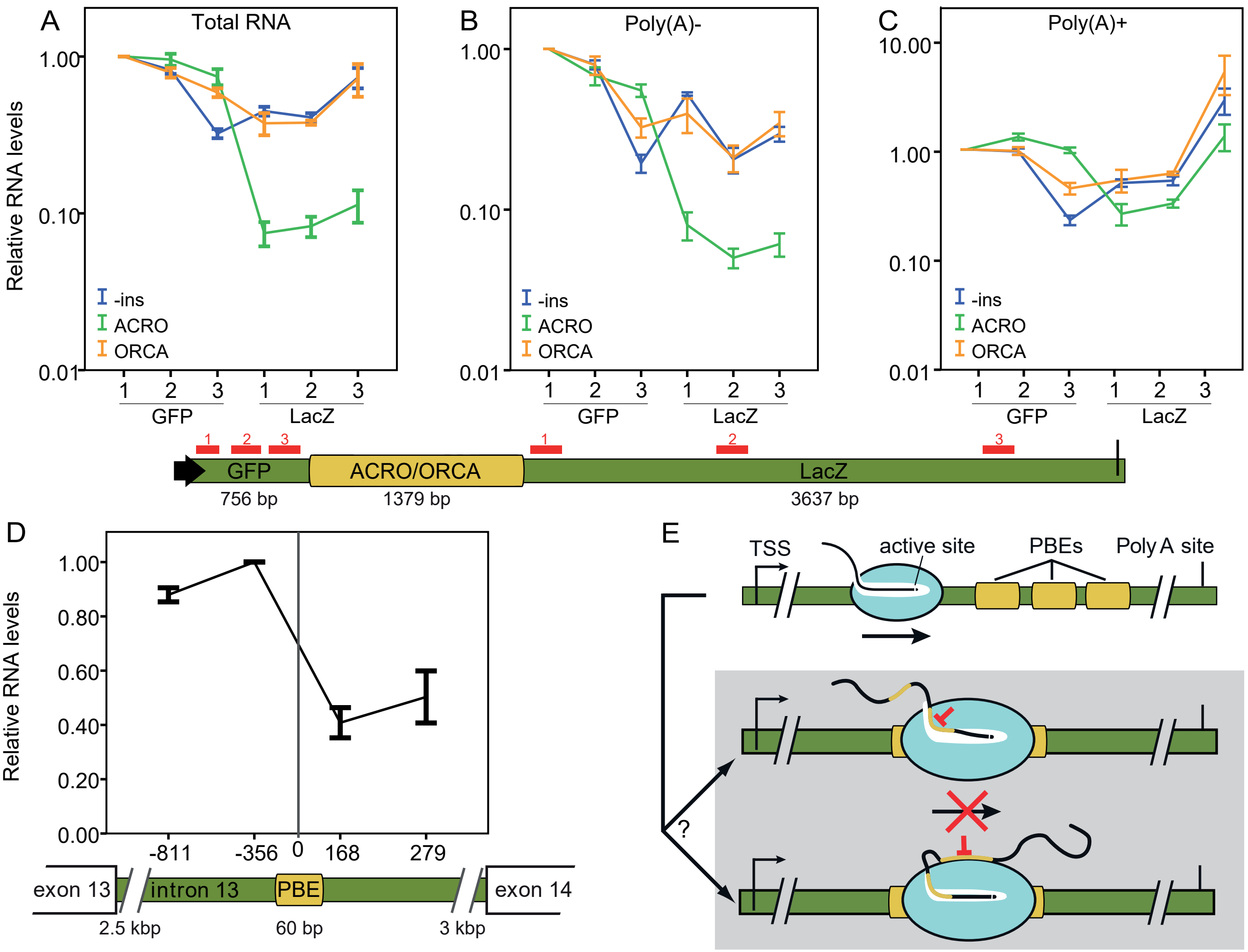

A
500 bp
5' UTR
ORF1
ORF2
B
S
B
RAPs
CMV
GFP
LacZ
SV40 pA
mcs
pUC ori
F1 ori
HSVtk pA
neo
SV40
C
CMV
GFP
LacZ
SV40 pA
-ins
L1
L1BS
gfp
neo
D
-ins
L1Pa6
9258
141
B2
L1
-ins
gfp
neo
E
RAP 5765
-ins
1X
2X
3X
4X
5X
6X
inv
mock
gfp
neo
F
ACRO
-ins
0.3
1.1
1.4
ORCA 1.4
gfp
neo
G
α-glob
GFP
LacZ
SV40 pA
-ins
ACRO
ORCA
gfp
neo
H
-ins
ACRO
-ins
+ ACRO
gfp
neo

Figure 5


A
Total RNA
Relative RNA levels
1.00
0.10
0.01
-ins
ACRO
ORCA
1 2 3 1 2 3
GFP
LacZ
B
Poly(A)-
1.00
0.10
0.01
-ins
ACRO
ORCA
1 2 3 1 2 3
GFP
LacZ
C
Poly(A)+
10.00
1.00
0.10
0.01
-ins
ACRO
ORCA
1 2 3 1 2 3
GFP
LacZ
1 2 3
1
2
3
GFP
ACRO/ORCA
LacZ
756 bp
1379 bp
3637 bp
D
Relative RNA levels
1.00
0.80
0.60
0.40
0.20
0.00
-811
-356
0
168
279
exon 13
intron 13
PBE
exon 14
2.5 kbp
60 bp
3 kbp
E
TSS
active site
PBEs
Poly A site
?

Published online 17 December 2015 *Nucleic Acids Research, 2016, Vol. 44, No. 4* ***1703–1717***
doi: 10.1093/nar/gkv1473

The spliceosome-associated protein Nrl1 suppresses homologous recombination-dependent R-loop formation in fission yeast

Lucia Aronica[1,2,*], Torben Kasparek[3], David Ruchman[1], Yamile Marquez[4], Lubos Cipak[5], Ingrid Cipakova[5], Dorothea Anrather[6], Barbora Mikolaskova[5], Maximilian Radtke[1], Sovan Sarkar[3], Chen-Chun Pai[3], Elizabeth Blaikley[3], Carol Walker[3], Kuo-Fang Shen[7], Renee Schroeder[1], Andrea Barta[4], Susan L. Forsburg[7] and Timothy C. Humphrey[3,*]

[1]Department of Biochemistry and Cell Biology, Max F. Perutz Laboratories, Vienna A-1030, Austria, [2]Department of Oncology, Stanford University, Stanford 94305, USA, [3]CRUK/MRC Oxford Institute for Radiation Oncology, Oxford OX37DQ , UK, [4]Department of Medical Biochemistry, Max F. Perutz Laboratories,Medical University of Vienna, Vienna A-1030, Austria, [5]Cancer Research Institute, Slovak Academy of Sciences, Bratislava 81438, Slovakia, [6]Max F. Perutz Laboratories, Mass Spectrometry Facility, Vienna A-1030, Austria and [7]University of Southern California, Los Angeles 90089-0911, USA

Received March 25, 2015; Revised November 17, 2015; Accepted December 03, 2015

ABSTRACT

The formation of RNA–DNA hybrids, referred to as R-loops, can promote genome instability and cancer development. Yet the mechanisms by which R-loops compromise genome instability are poorly understood. Here, we establish roles for the evolutionarily conserved Nrl1 protein in pre-mRNA splicing regulation, R-loop suppression and in maintaining genome stability. *nrl1Δ* mutants exhibit endogenous DNA damage, are sensitive to exogenous DNA damage, and have defects in homologous recombination (HR) repair. Concomitantly, *nrl1Δ* cells display significant changes in gene expression, similar to those induced by DNA damage in wild-type cells. Further, we find that *nrl1Δ* cells accumulate high levels of R-loops, which co-localize with HR repair factors and require Rad51 and Rad52 for their formation. Together, our findings support a model in which R-loop accumulation and subsequent DNA damage sequesters HR factors, thereby compromising HR repair at endogenously or exogenously induced DNA damage sites, leading to genome instability.

INTRODUCTION

Genome instability in the form of increased rates of mutations or chromosomal aberrations is a hallmark of most tumor cells and a key factor in cancer development, progression and prognosis (1). While dysfunctional DNA repair is a well recognized cause of genome instability (2), it is becoming increasingly appreciated that defects in mRNA biogenesis may also destabilize genomes through the formation of mutagenic structures referred to as R-loops (3). R-loops are three-stranded structures, which form during transcription when the nascent mRNA hybridizes to the complementary DNA template, forming an RNA/DNA hybrid and a displaced DNA strand (4). Through direct promotion of DNA damage (5) and indirect effects on gene expression (6–10), R-loops lead to different forms of genome instability. The genome-threatening effects of R-loops also play a role in tumor development (11–15), but the underlying mechanism is poorly understood.

Pre-mRNA splicing is a key process in genome maintenance (16,17), as reflected by its disruption in various cancer types (18). Increasing evidence suggests that splicing factors, R-loop suppression and DNA repair interface with each other in a coordinated manner to safeguard genome stability. Splicing factors can not only prevent R-loop formation (19), but also promote homologous recombination (HR) repair (20,21). Conversely, HR factors can both repress and promote R-loop levels in the cell (12,22), and interact both physically and functionally with the splicing machinery (23,24). These findings suggest that perturbation of splicing may lead to genome instability by inducing both accumulation of R-loops and defects in DNA repair in the cell.

In this study we show that the evolutionarily conserved protein Nrl1 associates with the spliceosome, affects pre-

[*]To whom correspondence should be addressed. Tel: +44 1865 617327; Fax: +44 1865 617318; Email: timothy.humphrey@oncology.ox.ac.uk
Correspondence may also be addressed to Lucia Aronica. Tel: +1 650 724 3439; Email: laronica@stanford.edu

mRNA splicing of a subset of genes and non-coding RNAs, and contributes to genome stability by both suppressing R-loops and promoting HR repair in the fission yeast *Schizosaccharomyces pombe*. Our findings suggest a model in which R-loop formation acts to sequester the HR machinery, thus leading to inhibition of HR repair. As the human ortholog of Nrl1 is down-regulated and associated with Copy Number Loss (CNL) in cancer (25,26), this mechanism may have important implications for the emerging yet still elusive role of R-loops in cancer.

MATERIALS AND METHODS

Strains, media and growth conditions

The genotypes of the strains used in this study are listed in the Supplementary Table S6. Strains carrying a deletion or a TAP-tagged version of *nrl1* have been constructed as described in (27) and (28), respectively. The strains were grown at 30°C in standard yeast extract with supplements (YE6S), minimal medium (EMM), or Pombe Minimal Glutamate (PMG), and crosses were performed at 25°C. For spot assays exponential phase cultures were serially diluted 5-fold, spotted onto the indicated media in the presence or absence of genotoxic drugs and incubated at 32°C for 3 days before analysis. For irradiation experiments, cells were grown to mid-log phase and exposed to 100 Gy of gamma irradiation (3.3 Gy/min, for 30 min). After irradiation cells were recovered for 30 min at 30°C. For plasmid rescue experiments PlRT3-based plasmids containing a *LEU2* gene were transformed into *leu1–32* strains and selected on leucine-deficient media before being spotted onto YE6S plates as indicated.

Colony sectoring and DSB assay

The sectoring assay was performed as previously described (29). The minichromosome Ch^{16}-LMYAU was crossed into wild-type and *nrl1Δ* strains from a donor strain. Cells were grown on selective media with thiamine (2μM) to repress HO expression from rep81X-*nmt*-HO (Leu+) integrated into SPCC1795.09 on the left arm of Ch^{16}-LMYAU. Cells were then diluted in MQ water and ~100 cells plated onto sectoring plates (containing either EMM+ arginine (15 mg/l), histidine (15 mg/l), uracil (15 mg/l), leucine (15 mg/l), and adenine (5 mg/l) with and without thiamine (break off/on) to identify break-dependent, and independent Ch^{16} loss. To detect break-induced LOH, cells were treated as above but grown in the absence of leucine to select for the left arm of the minichromosome. Plates were incubated for 56 h at 32°C and stored for 48 h at 4°C before being scored for the presence of sectored colonies. Results were confirmed by repeating the assay two further times.

Site-specific DSB assay

The DSB assay was performed as previously described (29). Wild-type and *nrl1Δ* strains containing the minichromosome Ch^{16}-RMYAH and either p28 (rep81X-HO) or p40 (rep81X) were grown exponentially in EMM liquid culture (with appropriate supplements to select for the plasmid while allowing for loss of Ch^{16}-RMYAH) for 48 h in the absence of thiamine to induce expression of HO endonuclease The percentage of colonies undergoing NHEJ/SCR (R^+ Y^R A^+ H^+), GC (R^+ Y^S A^+ H^+), Ch^{16} loss (R^- Y^S A^- H^-) and extensive break-induced LOH (R^+ Y^S A^- H^-) were calculated. To determine the levels of break-induced GC, Ch^{16} loss and LOH; background events at 48 h in a blank vector assay were subtracted from break-induced events in cells transformed with rep81X-HO. Each experiment was performed three times using three independently derived strains. A minimum of 1000 colonies were scored for each strain.

Protein purification and LC-MS/MS analysis

Isolation of Nrl1-TAP associated proteins, proteolytic digest (trypsin) and chromatographic separation of the peptides were performed as previously described (30) (Supplemental Methods). Raw data were searched with MaxQuant 1.5.1.2 (31) against the *S. pombe* database (http://www.pombase.org/) with tryptic specificity, 5 ppm precursor tolerance, 20 ppm fragment ion tolerance, filtered for 1% FDR on peptide and protein level.

Yeast two-hybrid assay

All constructs were made using vectors supplied in the Matchmaker GAL4 2-hybrid system (Clontech). Two-hybrid DNA-binding domain (BD) constructs were made in the pAS2–1 vector containing the *TRP1* gene for selection on tryptophan-deficient media and activation domain (AD) constructs were made in the pGADT7 vector containing the *LEU2* gene for selection on leucine-deficient media. *Saccharomyces cerevisiae* strain PJ69–4A was cotransformed simultaneously with both AD and BD constructs by the lithium acetate method as described in the Yeast Protocols Handbook of the Matchmaker system (Clontech). Cotransformants growing on both –Ade and –His selective media were assayed for β-galactosidase activity.

RNAseq library preparation and bioinformatic analysis

WT, *nrl1Δ*, WT+IR and *nrl1Δ*+IR strand specific cDNA libraries were prepared with lexogen SENSE protocol using $poly(A)^+$ RNA as previously described (32). Two biological replicates were used for each sample and libraries were sequenced using the Illumina platform. The resulting paired end sequencing reads (100-bp long) of each sample and biological replicate were aligned independently using Tophat v2.0.11. The following (not default) parameters were used for performing the alignment: -i 30, -I 2000, -p 16, -a 15 –library-type fr-firststrand, –b2-very-sensitive, –microexon-search and -G (gene annotation *S. pombe* ASM294v2).

Splicing analysis of WT and *nrl1Δ* was performed using the splice junctions predicted by Tophat. Only those introns that present at least two unique reads in both biological replicates were used for further analysis. Introns were classified as new if they were not included in the gene annotation (ASM294v2). To determine differences in intron splicing, the PSI (percentage of spliced in) was calculated by using uniquely mapped splice junction and exonic reads.

Only those changes over 15% (ΔPSI > 15) and a *P*-value $\leq$ 0.05 between WT and *nrl1Δ* are illustrated in Figure 5 and listed in Supplementary Table S3. For obtaining differentially expressed genes between a pair of samples (Supplementary Tables S5.1–S5.5) Cuffquant and Cuffdiff from the Cufflinks v2.2.1 package were used.

Nuclear Spreading and Indirect Immunofluorescence

Chromosome spreads were performed as previously described (33). For R-loop detection, slides were incubated with the mouse monoclonal antibody S9.6—kind gift of N. Proudfoot (Sir William Dunn School of Pathology, UK) and L. Székvölgyi (University of Debrecen, Hungary)—as previously described (34). For RNase H controls, slides were incubated with 2 U of RNase H (Roche) and 1 μg RNase A (Roche) in PBS buffer for 2 h prior to antibody treatment. For co-localization analysis cells were grown in YE6S medium at 25°C in a shaking incubator for ~20 h to reach mid-log phase and then were treated with or without 3 μM Bleomycin at 25°C for 4 h before harvesting. Primary antibodies (S9.6, KeraFAST; rabbit anti-GFP antibody ab290, Abcam Inc) at 1:500 dilution and secondary antibodies (approciate Cy2 conjugated Donkey anti-Mouse IgG2 antibody, Jackson; Texas Red conjugated Goat anti Rabbit IgG antibody; Jackson) were used in these experiments. After a final wash step, the slides were mounted with ~10 μl of mounting medium containing DAPI (Invitrogen).

Image collection

Images were acquired with a DeltaVision Core widefield deconvolution microscope (Applied Precision, Issaquah, WA, USA) using an Olympus 603/1.40, PlanApo, NA = 1.40 objective lens and a 12-bit Photometrics CoolSnap HQII CCD, deep- cooled, Sony ICX-285 chip. The system x-y pixel size is 0.1092 mm x-y. softWoRx v4.1 (Applied Precision) software was used at acquisition electronic gain = 1.0 and pixel binning 1 3 1. Excitation illumination was from a solid-state illuminator (seven-color version); Cy2 was detected with a 0.1-s exposure; Texas Red was detected with a 0.1-s exposure; DAPI was detected with a 0.2-s exposure. Suitable polychroic mirror Semrock DAPI/FITC/A594/Cy5 API#52–852112–000 bs generally: 433/55–522/34–593/64–655LPish was used. Twelve z sections at 0.4 mm were acquired. Three-dimensional stacks were deconvolved with manufacturer-provided optical transfer function using a constrained iterative algorithm and images were maximum- intensity projected. Images were contrast adjusted using a histogram stretch with an equivalent scale and gamma for comparability.

RESULTS

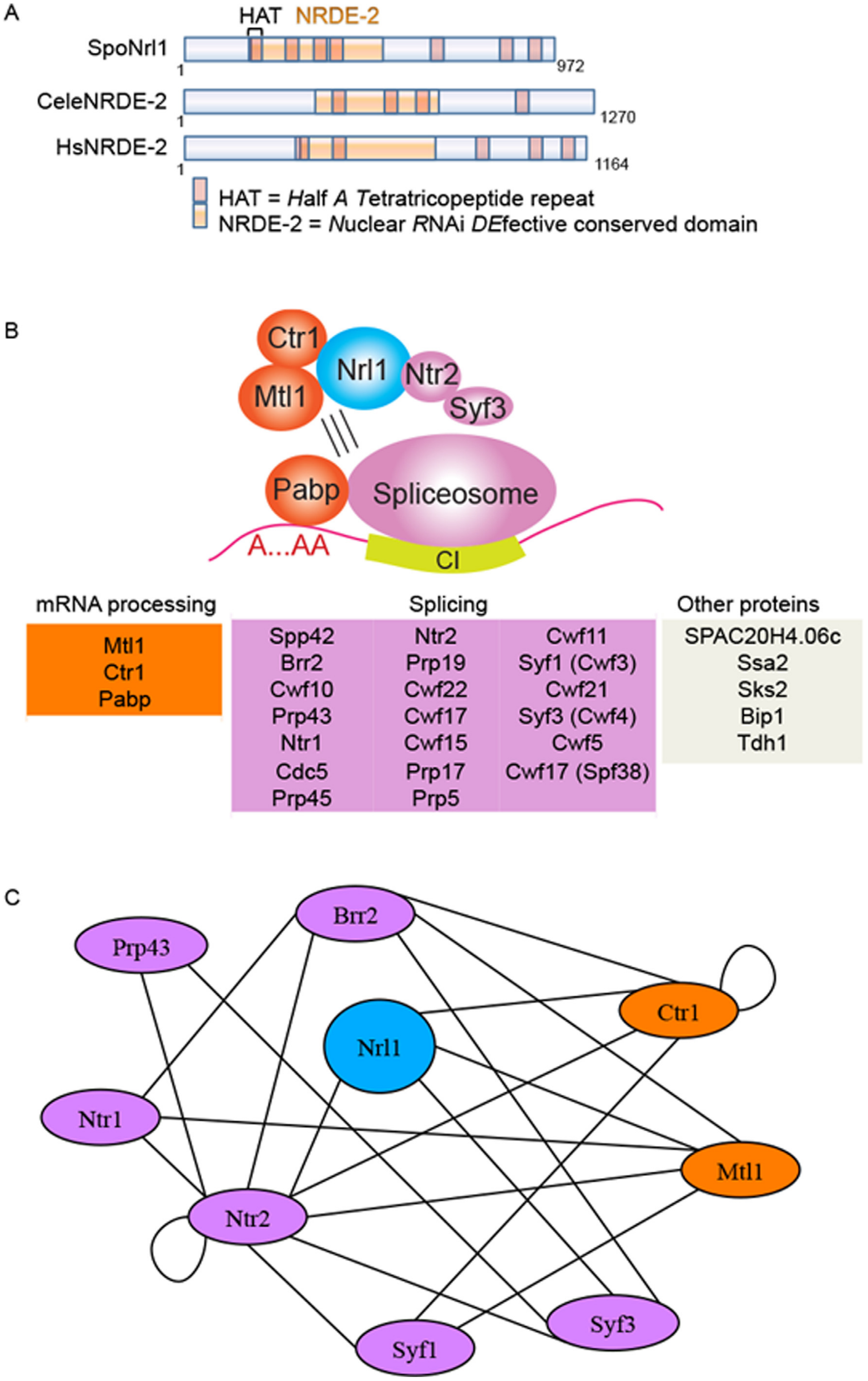
Nrl1 associates with the spliceosome and affects pre-mRNA splicing

In the worm *Caenorabditis elegans* the protein NRDE-2 is required for nuclear RNAi silencing by a mechanism based on inhibition of transcription elongation (35). We performed a computational analysis of NRDE-2, and identified homologous proteins in other eukaryotes, including human (Supplementary Figure S1). Surprisingly, these NRDE-2 like factors were unrelated to any known RNAi factors but were structurally similar to splicing proteins (Supplementary Figure S2).

To explore the function of NRDE-2 factors further, we studied the NRDE-2 like gene SPBC20F10.05, which we named *nrl1* (*N*RDE-2 *l*ike 1), in the tractable model organism *S. pombe* (Figure 1A). Nrl1 and its worm and human orthologs share a conserved domain of unknown function and a common three-dimensional structure containing Half-A-Tetratricopeptide (HAT) motifs, which closely resembles the domain architecture of pre-mRNA processing and splicing factors such as Syf1 and Syf3 (36) (Figure 1A, Supplementary Figure S2).

To analyze Nrl1 in its cellular context, we isolated Nrl1-associated factors by tandem affinity purification (TAP) of Nrl1-TAP tagged strains, and identified the purified proteins by mass spectrometry (MS) (28,37). We performed TAP purification both in presence and absence of RNase A to distinguish between core complex proteins and factors indirectly binding to Nrl1 through RNA-mediated interaction (Figure 1B, Supplementary Figure S3, Table S1). Nrl1 copurified with an RNA-resistant core complex consisting of the pre-mRNA processing factors Mtl1 and Ctr1 and an RNA-sensitive sub-complex including components of the U2·U5·U6 spliceosome and Prp19 complexes (38). However, we also performed yeast-two-hybrid analysis and found that Nrl1 interacts directly not only with Mtl1 and Ctr1 but also with the splicing proteins Ntr2 and Syf3, which is structurally related to Nrl1 (Figure 1C, Supplementary Figure S2, Table S2). This suggests that the RNA-binding proteins Ntr2 and Syf3 represent bridge binders between an RNA-independent core complex (Nrl1-Mtl1-Ctr1) and an RNA-dependent sub-complex consisting of other splicing and RNA-processing factors. In contrast, no RNAi factor copurified with Nrl1 (Supplementary Table S4). These findings are consistent with a recent publication in which Nrl1 was shown to interact with spliceosome components (39), and further identify splicing proteins Ntr2 and Syf3 as proteins attracting Nrl1 into the spliceosome.

To gain functional insights into the role of Nrl1, we created an *nrl1* deletion mutant (*nrl1Δ*) by replacing the *nrl1*$^+$gene with a *natMX4* drug resistance cassette (strains 16594–5), and analyzed its phenotype. To explore a possible role for Nrl1 in pre-mRNA splicing we performed high-throughput paired-end sequencing of polyA+-selected mRNA (RNA-Seq) from wild-type and *nrl1Δ* cells. To determine differences in intron splicing, the PSI was calculated by dividing the number of uniquely mapped exonic reads by the sum of uniquely mapped exonic reads and uniquely mapped splice junction reads spanning exon-exon borders. An intron was considered to be differentially spliced if there was a difference in its retention of more than 15% (ΔPSI > 15 with a *P*-value $\leq$ 0.05) between wild-type and *nrl1Δ* cells. All introns obtained in our RNA-seq experiments were evaluated, and 43 introns in protein-coding genes and non-coding RNAs met the criteria mentioned above (Supplementary Figure S4, Table S3). The affected genes were involved in a variety of processes including cellular transport and metabolism, transcription regulation and pre-mRNA processing. Together, these findings indicate that Nrl1 physically and functionally associates with the splicing machin-



mRNA processing	Splicing			Other proteins
Mtl1	Spp42	Ntr2	Cwf11	SPAC20H4.06c
Ctr1	Brr2	Prp19	Syf1 (Cwf3)	Ssa2
Pabp	Cwf10	Cwf22	Cwf21	Sks2
	Prp43	Cwf17	Syf3 (Cwf4)	Bip1
	Ntr1	Cwf15	Cwf5	Tdh1
	Cdc5	Prp17	Cwf17 (Spf38)	
	Prp45	Prp5		

Figure 1. Nrl1 associates with spliceosome proteins. (**A**) Comparison of *S. pombe* (Spo) Nrl1, *C. elegans* (Cele) and *Homo sapiens* (Hs) NRDE-2 like proteins. HAT = halfa-tetratricopeptide domain. (**B**) Nrl1-associated proteins were isolated from exponentially growing WT cells harboring a TAP-tagged *nrl1* allele (17106) in the presence or absence of RNase A by tandem affinity purification and identified by mass spectrometry (MS) analysis. A core complex consisting of Nrl1, Mtl1, Ctr1, Ntr2 and Syf3 associates through RNA-dependent interactions with the spliceosome. Blue: Nrl1; pink: splicing factors; orange: mRNA processing factors. Only the top 30 proteins based on spectral counting are shown. (**C**) Yeast-two-hybrid interaction map of Nrl1 interactome. All constructs were made using vectors supplied in the Matchmaker GAL4 2-hybrid system (Clontech). Two-hybrid DNA-binding domain (BD) constructs were made in the pAS2–1 vector containing the *TRP1* gene for selection on tryptophan-deficient media and activation domain (AD) constructs were made in the pGADT7 vector containing the *LEU2* gene for selection on leucine-deficient media.

ery and its loss results in changes in the splicing patterns of a subset of introns.

nrl1 deletion leads to accumulation of endogenous DNA damage

Unexpectedly, *nrl1Δ* cells exhibited several markers of endogenous DNA damage. *nrl1Δ* cells were elongated (average cell length = 16.4 μm) compared with wild-type cells (average length = 10.8 μm) with a subpopulation of giant cells (3–5% of cells) reaching 30–50 μm and exhibiting nuclear fragmentation together with chromatin hypercondensation (Figure 2A, left). This elongated phenotype is frequently observed in cells accumulating unrepaired DNA lesions, which activate DNA damage checkpoint pathways to delay the cell cycle and provide an opportunity for DNA repair (40). We therefore tested whether the depletion of the G2/M checkpoint kinase Chk1 or the intra-S phase checkpoint kinase Cds1 could attenuate the elongated cell phenotype of *nrl1Δ*. We found *nrl1Δ chk1Δ* double mutants but not *nrl1Δ cds1Δ* double mutants exhibited a cell length comparable to wild-type (average length of 9.8 μm and 15.8 μm, respectively) with no detectable giant cells. This finding indicates that loss of Nrl1 activates the DNA damage checkpoint resulting in Chk1-triggered G2/M arrest (Figure 2A, right). We next assayed Chk1 activation by Rad3-catalyzed phosphorylation, which is evident as a mobility shift on western blots (41,42) (Supplementary Figure S5). While Chk1 activity was detected in response to MMS in both wild-type and *nrl1Δ* cells, no Chk1 activity was detected in untreated cells. Thus, it is likely that Chk1 activation is triggered endogenously in a minority of cells reflecting the elongated G2-arrested subpopulation of *nrl1Δ*.

To investigate the nature of the endogenous DNA damage observed in *nrl1Δ*, we measured the accumulation of spontaneous Rad52 DNA repair foci in wild-type and *nrl1Δ* strains bearing a yellow fluorescent protein tagged Rad52 (Rad52-YFP) (Figure 2B). Rad52 is a repair protein, which accumulates at DNA lesions to facilitate repair of DNA double-strand breaks (DSBs) through the HR pathway (43–45). Strikingly, a significant increase in Rad52-YFP foci was observed in *nrl1Δ* (23%, $P < 0.05$) compared with wild-type (7%) under normal growth conditions. Together these observations indicate that loss of Nrl1 leads to endogenous DNA damage, accumulation of Rad52-bound DNA lesions and activation of the G2/M checkpoint.

Further, *nrl1Δ* cells were hypersensitive to exogenous DNA damage as shown by treatment with the genotoxic drugs bleomycin (Bleo), methylmethane sulphonate (MMS), and camptothecin (CPT) (Figure 2C). In contrast, *nrl1Δ* cells were not hyper-sensitive to hydroxyurea (not shown). Notably, *nrl1Δ rad3Δ* and *nrl1Δ rad52Δ* double mutants were synthetically sick and exhibited even greater sensitivity to DNA damage than single mutants (Supplementary Figure S6). These observations confirm that in the absence of Nrl1 cells accumulate unrepaired DNA damage, which further sensitizes HR and checkpoint mutants involved in the DNA damage response.

Nrl1 is required for efficient DSB repair by homologous recombination

The accumulation of Rad52 foci in response to endogenous DNA damage and *nrl1Δ* hypersensitivity to exogenous DNA insults suggested that Nrl1 might be required for efficient DNA DSB repair. To explore a possible role for Nrl1 in DSB repair, we employed a previously described colony-sectoring assay to allow rapid visualization of defects in repair of a broken nonessential minichromosome (Ch^{16}-LMYAU) following site-specific DSB induction by the HO endonuclease (29). Consistent with a role for Nrl1 in DSB repair, *nrl1Δ* cells exhibited elevated levels of break-induced minichromosome loss and break-induced chromosomal rearrangements resulting in loss of heterozygosity (LOH) as determined by the break-induced sectoring assay (Supplementary Figure S7A–B)

To quantify DSB repair in *nrl1Δ* cells DSB-induced marker loss was assessed using a DSB assay in which a previously described minichromosome (Ch^{16}-RMYAH) was cleaved uniquely at the *MAT*a target site following HO endonuclease derepression from a plasmid (pREP81X-HO) (46) (Figure 3A). Following break induction by thiamine depletion, cells were plated onto YE6S plates and colonies were replica plated to selective plates to determine the marker loss profile for each colony (Figure 3B). The DSB repair profile indicated that colonies exhibiting a gene conversion (GC) phenotype (arg$^+$ hyg^s, ade$^+$ his$^+$) were significantly reduced in an *nrl1Δ* background (49%) compared to wild-type levels (72%, $P < 0.05$). Compared with wild-type, *nrl1Δ* also showed a significant increase in failed repair events resulting in both Ch^{16} loss (arg$^-$ hyg^s, ade$^-$ his$^-$ colonies: *nrl1Δ* = 29%, wild-type = 18%; $P < 0.01$) and LOH phenotype (arg$^+$ hyg^S ade$^-$ his$^-$ colonies: *nrl1Δ* = 18%, wild type = 5%; $P < 0.01$). No significant difference in levels of arg$^+$ hyg^R ade$^+$ his$^+$ colonies, arising from Non Homologous End Joining (NHEJ)/sister chromatid recombination (SCR), was observed (*nrl1Δ* = 10%, wild-type 8%). The low levels of GC and high levels of both failed DSB repair (Ch^{16} loss) and misrepair (LOH) indicate that Nrl1 is required for efficient HR repair of DSBs and thereby for maintaining genome stability.

Loss of Nrl1 results in prolonged accumulation of RPA and Rad52 foci upon DNA damage

The first key step in HR repair is sensing the presence of DSBs to recruit the effectors of the DNA damage response (DDR). Rad11 (RPA1), is a component of the replication protein A (RPA) complex which binds to single-stranded DNA (ssDNA) at resected DSBs, thereby promoting DNA damage sensing (47). To analyze whether Nrl1 might be required for this early step of the HR pathway, we measured the kinetics of recruitment and unloading of Rad11 by fluorescence microscopy in wild-type and *nrl1Δ* cells bearing a green fluorescent protein tagged Rad11 (Rad11-GFP), following exposure to 50 Gy of ionizing radiation (IR) (Figure 4A). Remarkably, Rad11 foci, which tailed off 90 min after irradiation in wild-type cells, persisted at very high levels up to 5 h after irradiation (Rad11 positive cells: WT = 19%, *nrl1Δ* = 50%). The dramatic increase and prolonged persistence of Rad11 foci in *nrl1Δ* cells suggested that the down-

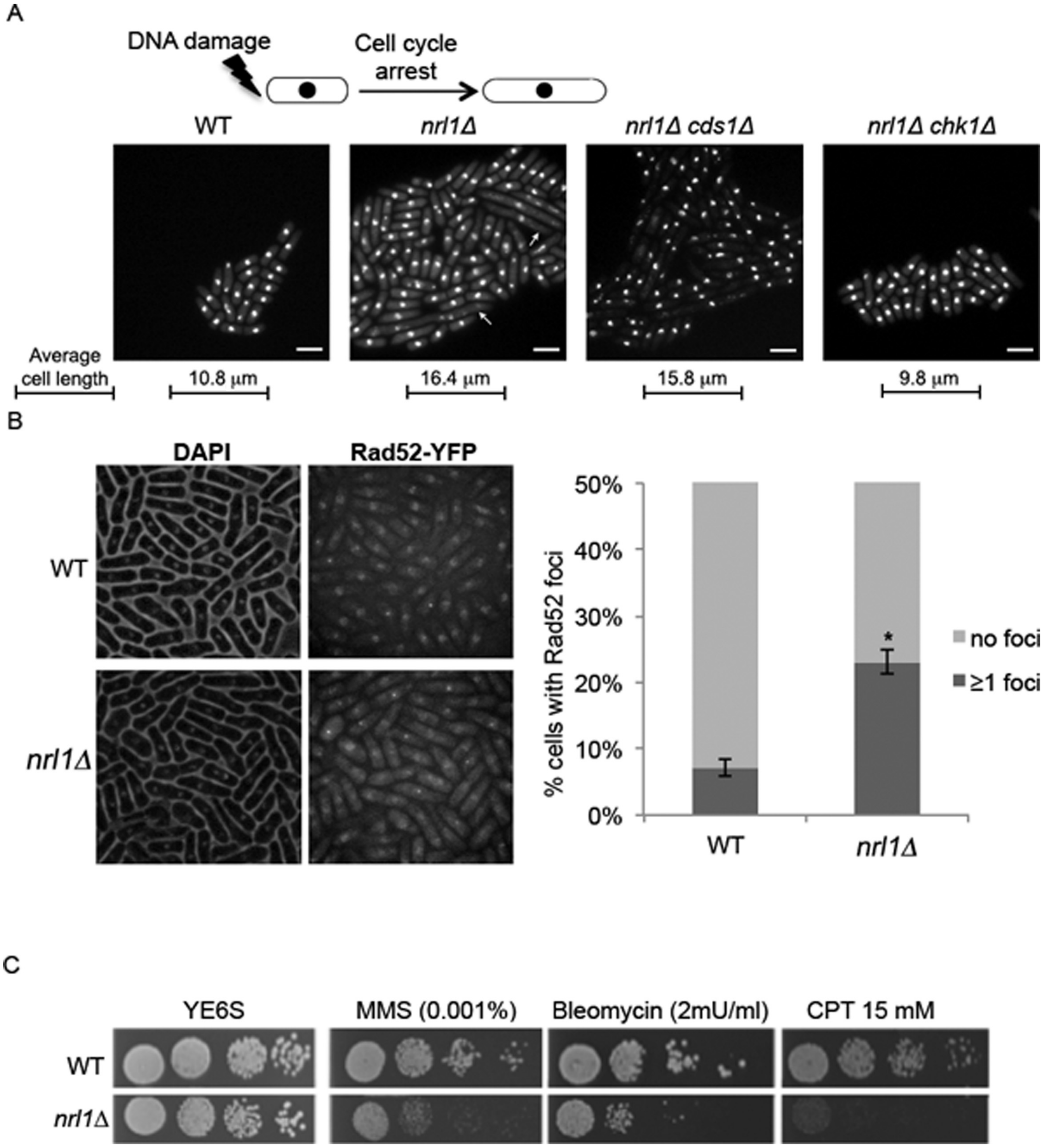


Figure 2. *nrl1*$^{+}$ deletion leads to accumulation of endogenous DNA damage and sensitivity to genotoxic agents. (**A**) Top: Depiction of DNA-damage induced cell cycle arrest and elongated cell morphology. Bottom: Fluorescent microscopy analysis of DAPI-stained wild-type (WT) (TH8342), *nrl1Δ* (TH8341), *nrl1Δ chk1* (5972) and *nrl1Δ cds1* (16594). Arrowheads indicate abnormally elongated cells displaying chromatin fragmentation. Scale bar = 10 μm. (**B**) Left: Fluorescent microscopy analysis of DAPI-stained Rad52-YFP harboring WT and *nrl1Δ* cells. Right: Quantification of cells with one or more Rad52-YFP foci. Mean and standard deviation were scored from five experiments, n > 200. The asterisk (*) indicates a significant difference compared with WT as determined by Wilcoxon signed rank test ($P = 0.0313$). (**C**) *nrl1Δ* strains are hypersensitive to genotoxic agents. Five-fold serial dilutions of WT (TH2094) and *nrl1Δ* (TH8103) on YE6S, methyl methanesulfonate (MMS), bleomycin (Bleo) and camptothecin (CPT) at indicated concentrations.

stream HR repair response might also be impaired in *nrl1Δ* cells. We therefore analyzed foci accumulation of yellow fluorescent protein tagged Rad52 (Rad52-YFP) and cyan fluorescent protein tagged Rad51 (Rad51-CFP) in wild-type and *nrl1Δ* cells following exposure to IR (50 Gy). Rad52 facilitates the displacement of Rad11 from ssDNA and its replacement with Rad51, the central recombinase of the HR pathway, which catalyzes strand invasion into homologous sequences during GC events (48). Consistent with the increase of Rad11 foci, Rad52 foci persisted at very high levels up to 5 h after irradiation (Rad52 positive cells: WT = 17%, *nrl1Δ* = 49%) (Figure 4B). In contrast, despite peaking at 30 min after irradiation in both wild-type and *nrl1Δ* cells, overall Rad51-CFP foci formation was considerably lower in *nrl1Δ* cells (35%) compared with wild-type cells (65%; Figure 4C). Importantly, this phenotype was not due to a decrease of Rad51 protein, as levels of Rad51-eCFP were comparable between wild-type and *nrl1Δ* cells (Figure 4D). This finding indicates that Nrl1 is required for efficient loading of Rad51 at DSBs, which is consistent with the sig-

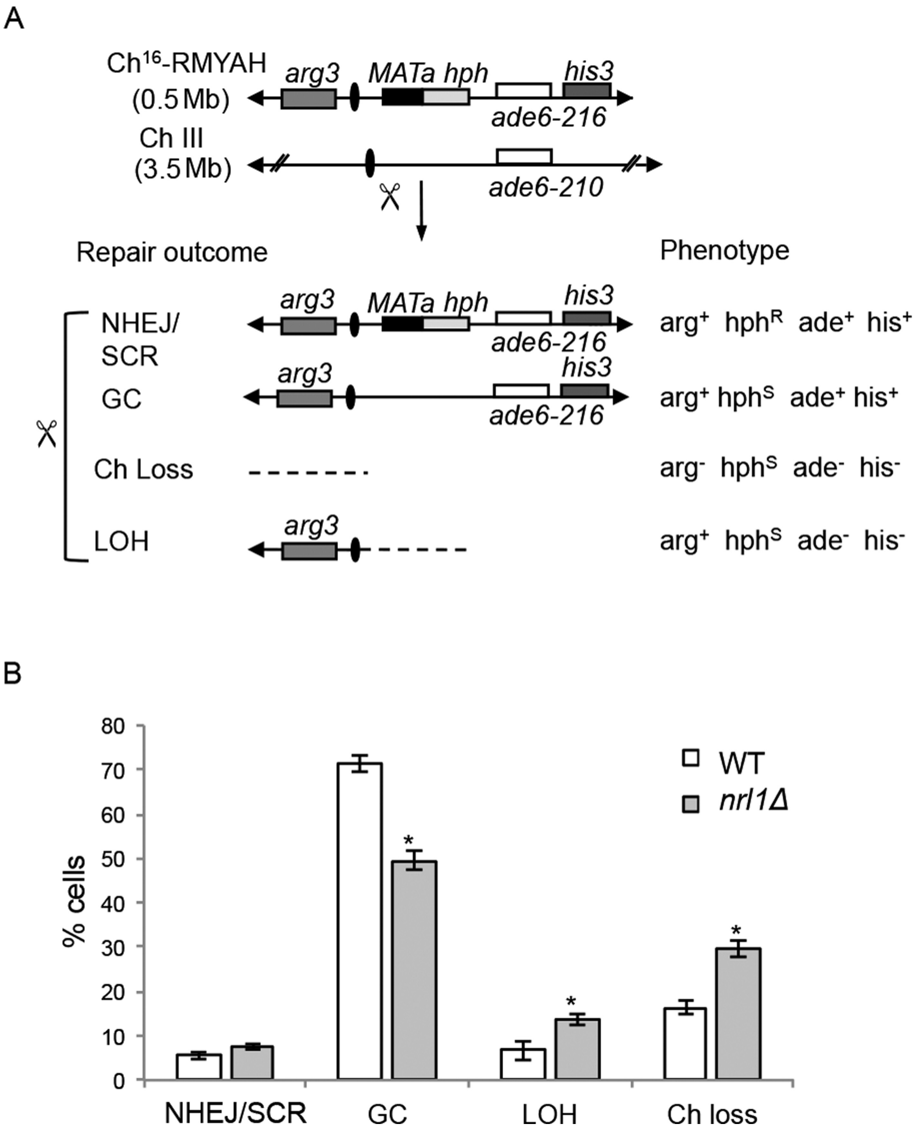


Figure 3. Nrl1 is required for efficient DSB repair by Homologous Recombination. (**A**) Schematic of the minichromosome Ch^{16}-RMYAH and possible outcomes following DSB induction at *MAT*a target site (scissors). (**B**) Quantitative DSB assay. Percentage of DSB-induced marker loss in WT Ch^{16}-RMYAH transformed with pREP81X-HO (TH4104, TH4121–2) or pREP81X (TH4125) and *nrl1Δ* Ch^{16}-RMYAH transformed with pREP81X-HO (TH8913–5) or pREP81X (TH8916–8) backgrounds. Means ± standard errors of three experiments are shown. The asterisk (*) represents significant difference compared with WT ($P < 0.01$). NHEJ: Non Homologous End Joining; SCR: sister chromatid recombination; Ch loss: chromosome loss; LOH: loss of heterozygosity.

nificantly reduced levels of Rad51-mediated GC observed in *nrl1Δ* compared to wild-type cells following DSB induction.

We therefore tested whether Nrl1 could directly recruit Rad51 or other DDR factors to DSBs. To this end, we examined Nrl1-associated proteins from Nrl1-TAP strains exposed to IR (100 Gy). However, no new protein or DNA repair factor was found to interact with Nrl1 upon exposure to IR compared with untreated cells (Supplementary Table S4). This indicates that Nrl1 is not required to directly recruit DDR factors to sites of DNA damage and that other mechanisms underlie its requirement for efficient Rad51 loading at DSBs.

Nrl1 depletion and DNA damage result in similar transcriptional changes

Given its association with spliceosomal factors, we examined whether Nrl1 might indirectly affect HR repair by affecting the splicing or expression of DDR genes. We therefore tested the splicing of intron-containing HR transcripts whose proteins act upstream of Rad51 loading, including *rad11*$^{+}$*, rad55*$^{+}$*, rad57*$^{+}$ and *swi5*$^{+}$*,* in untreated and irradiated (IR; 100 Gy) wild-type and *nrl1*Δ cells by RT-PCR. None of these genes showed differential splicing in *nrl1*Δ (Supplementary Figure S8). Next, we performed RNA-Seq of polyA+ selected mRNA from irradiated cells (wild-type+IR, *nrl1*Δ+IR; 100 Gy) and compared the sequencing data with those obtained from untreated cells (wild-type,

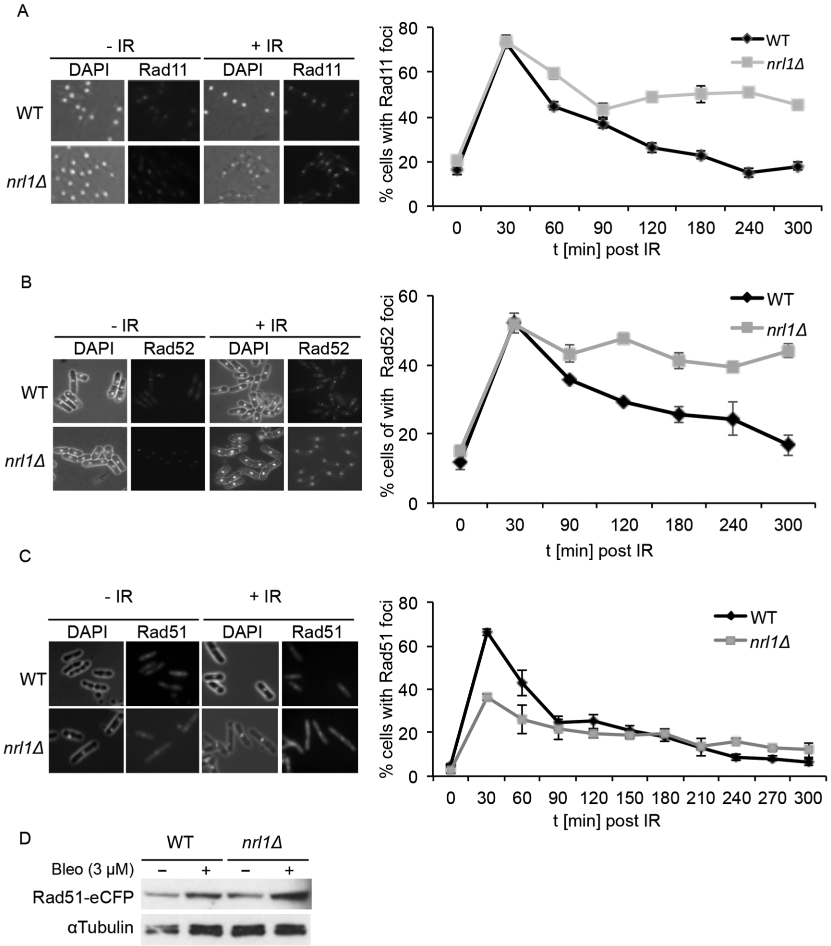


Figure 4. Loss of Nrl1 results in prolonged accumulation of RPA and Rad52 foci upon DNA damage. (**A**) *nrl1Δ* cells show persistently high levels of Rad11 foci upon DNA damage. Left: Rad11-GFP-tagged WT (TH2151) and *nrl1Δ* (TH8125) cells were grown in YE6S until exponential phase, treated with 50 Gy of IR and analyzed by fluorescence microscopy (100x; Phase/DAPI and GFP). Right: Quantitation of cells with Rad11-GFP foci at indicated time points. (**B**) *nrl1Δ* shows persistently high levels of Rad52 foci upon DNA damage. Left: Rad52-YFP tagged WT (TH8906) and *nrl1Δ* (TH8907) cells were treated as in (A) and analyzed by fluorescent microscopy (100x; Phase/DAPI and eCFP). Right: Quantitation of cells with Rad52-YFP foci at indicated time points. (**C**) *nrl1Δ* show decreased level of Rad51 foci upon DNA damage. Left: Rad51-eCFP tagged WT (TH3607) and *nrl1Δ* (TH8123) cells were treated as in (A) and analyzed by fluorescent microscopy (100x; Phase/DAPI and eCFP). Right: Quantitation of cells with Rad51-eCFP foci at indicated time points. (**D**) Protein levels of Rad51 are not affected in *nrl1Δ* cells. Western blot of Rad51-eCFP tagged WT (TH3607) and *nrl1Δ* (TH8123) cells in the presence (+) or absence (−) of bleomycin (bleo, 3 μM).

*nrl1*Δ). In both irradiated and untreated cells, *nrl1* deletion did not affect the expression of any known DNA repair genes (Supplementary Table S5.1, Table S5.4). In contrast, in untreated cells, *nrl1Δ* induced profound transcriptional changes, which were remarkably similar to those induced by IR in wild-type cells. A total of 231 genes were differentially expressed between wild-type and *nrl1Δ,* whereby 85 genes displayed similar transcriptional changes to *nrl1Δ* compared to wild-type +IR cells (85 out of 153, 55%; r = 0.86, *P*-value ≤ 2.2e-16; Figure 5, Supplementary Table S5.1, Table S5.3). Notably, this common fraction of genes affected in both *nrl1Δ* and wild type +IR is significantly higher than the previously reported fraction of genes affected in both wild-type cells exposed to IR and wild-type cells treated with the alkylating drug MMS (30%, *P*-value ≤ 4.5e-11) (49). Consistent with these findings, only 42 genes were differentially expressed in *nrl1Δ*+IR compared with *nrl1Δ* (Supplementary Table S5.4)*,* while no genes showed significant expression changes in wild-type +IR compared with *nrl1Δ*+IR (Supplementary Table S5.5). These findings indicate that the absence of Nrl1 results in transcriptional changes similar to those induced by IR.

Nrl1 prevents HR-dependent R-loop accumulation

The DNA damage-like transcriptional profile of *nrl1Δ* and the association of Nrl1 with the splicing machinery suggested that *nrl1Δ* might accumulate endogenous DNA lesions in the form of R-loops. R-loops are genome-threatening structures consisting of an RNA:DNA hybrid and a displaced ssDNA, which can arise from defects in splicing (19). We therefore sought to analyze R-loop formation in *nrl1Δ*. To this end we performed immunostaining on chromosome spreads from wild-type and *nrl1Δ* using the mouse monoclonal S9.6 antibody, which recognizes RNA/DNA duplexes (50). *nrl1Δ* cells showed a dramatic

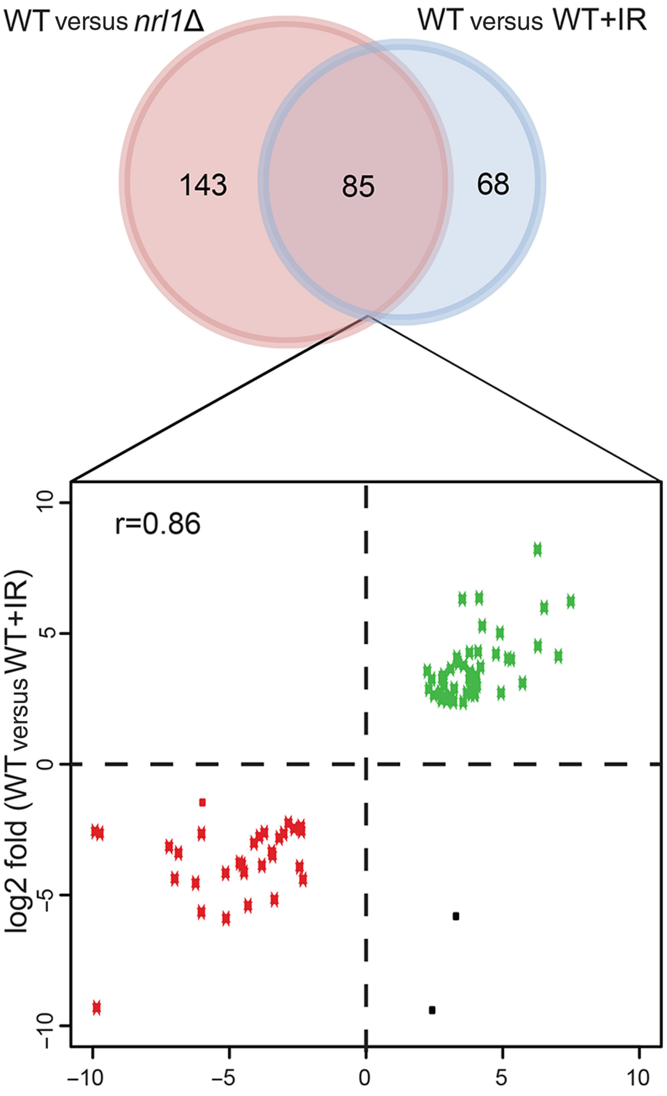


Figure 5. *nrl1Δ* displays DNA damage-associated transcriptional changes. Venn diagram of differentially expressed genes between WT versus *nrl1Δ* (red circle) and WT versus WT+IR (blue circle). Eighty five differentially expressed genes were shared between both comparisons. The log2 fold changes of those 85 genes are shown in the scatter plot below (Pearson correlation coefficient, r = 0.86). Down-regulated genes are depicted in red dots while up-regulated genes are depicted in green dots. Black dots are genes that are differentially expressed but are not congruent in both comparisons.

increase in R-loop accumulation (53%) when compared with wild-type cells (7%; $P = 0.01$; Figure 6A–B, Supplementary Figure S6). Furthermore, this increase was even higher upon exposure to IR (wild-type = 10%, *nrl1Δ* = 71%; $P = 0.003$). The observed immunostaining signals were sensitive to pre-treatment with RNase H, which specifically degrades RNA/DNA hybrids, thus confirming these foci as R-loops (Figure 6A). These findings indicate that *nrl1Δ* accumulates R-loops.

We next tested whether R-loops might be a source of endogenous DNA damage in *nrl1Δ*. A prediction from this was that removal of the R-loop degrading enzymes RNase H1 and RNase H2 should result in synthetic growth defects and enhanced sensitivity to DNA damage in an *nrl1Δ* background due to increased accumulation of RNA/DNA hybrids. Consistent with this prediction, we found that crossing *nrl1Δ* with *rnh1Δ* gave a reduced number of viable progeny compared to that expected, and the triple mutant *nrl1Δ rnh1Δ rnh201Δ* was obtained much less frequently. Accordingly, we found that *nrl1Δ rnh1Δ rnh2Δ* exhibited acute sensitivity to bleocin (Figure 6C). Thus R-loops are the likely source of DNA damage in *nrl1Δ*.

These findings raised the intriguing hypothesis that HR proteins may bind R-loops in *nrl1Δ* thus explaining the increased levels of endogenous Rad52 foci and long-term persistence of Rad11 and Rad52 foci following IR. We therefore analyzed the localization of HR proteins in relation to R-loops by immunostaining in WT and *nrl1Δ* strains harboring a Rad11-GFP, Rad52-YFP or Rad51-CFP protein by co-immunostaining both in the absence and presence of DNA damage (Figure 7A–C). We observed approximately 25% colocalization between RNA:DNA hybrids and each of these HR factors in the absence of exogenous DNA damage in both wild-type and *nrl1Δ*. Strikingly, in the presence of bleomycin, the percentage of RNA:DNA hybrids associating with HR proteins increased dramatically–in both wild-type and *nrl1Δ* for Rad11 (wt = 75%, *nrl1Δ* = 94%, Figure 7A) and Rad51 (wt = 82%, *nrl1Δ* = 91%, Figure 7B), but only in *nrl1Δ* in the case of Rad52 (wt = 36%, *nrl1Δ* = 92%, Figure 7C). These findings indicate that, despite their different pattern of foci formation, Rad11, Rad52 and Rad51 similarly associate with R-loops in *nrl1Δ* upon DNA damage.

We next analyzed whether HR factors might associate with R-loops to facilitate their formation. Rad51 and Rad52 have recently been shown to catalyze R-loop formation in pre-mRNA processing mutants in *S. cerevisiae* (22). We therefore tested whether HR factors might similarly mediate R-loop formation in the absence of Nrl1 in *S. pombe*. Strikingly, R-loop formation was significantly decreased in both *nrl1Δ rad51Δ* (3%; $P = 0.01$) and *nrl1Δ rad52Δ* (6%; $P = 0.02$) mutants compared to *nrl1Δ* (53%; Figure 7D). The increased IR-induced R-loop formation in an *nrl1Δ* background also required both Rad51 and Rad52 (Figure 6C; Supplementary Figures S9 and S10). These findings indicate that HR proteins promote R-loop formation in *nrl1Δ*.

The accumulation of HR-dependent R-loops in *nrl1Δ* suggested that R-loops or their associated DNA damage might sequester HR factors from exogenous DNA damage lesions thus explaining the HR repair defect in *nrl1Δ* cells. A prediction from this sequestration model was that overexpression of Rad51 would rescue *nrl1Δ* sensitivity to DNA damage. Consistent with this, we found that Rad51 overexpression (Rad51OP) following transformation of *nrl1Δ* cells with pIRT-Rad51 plasmid, but not with empty vector, partially suppressed the bleocin sensitivity of *nrl1Δ* (Figure 7E). These findings support a model, in which the HR repair defects in *nrl1Δ* are due to sequestration of HR factors at sites of R-loop formation.

DISCUSSION

We have investigated the function of the evolutionarily conserved Nrl1 protein in fission yeast, and identified its role in

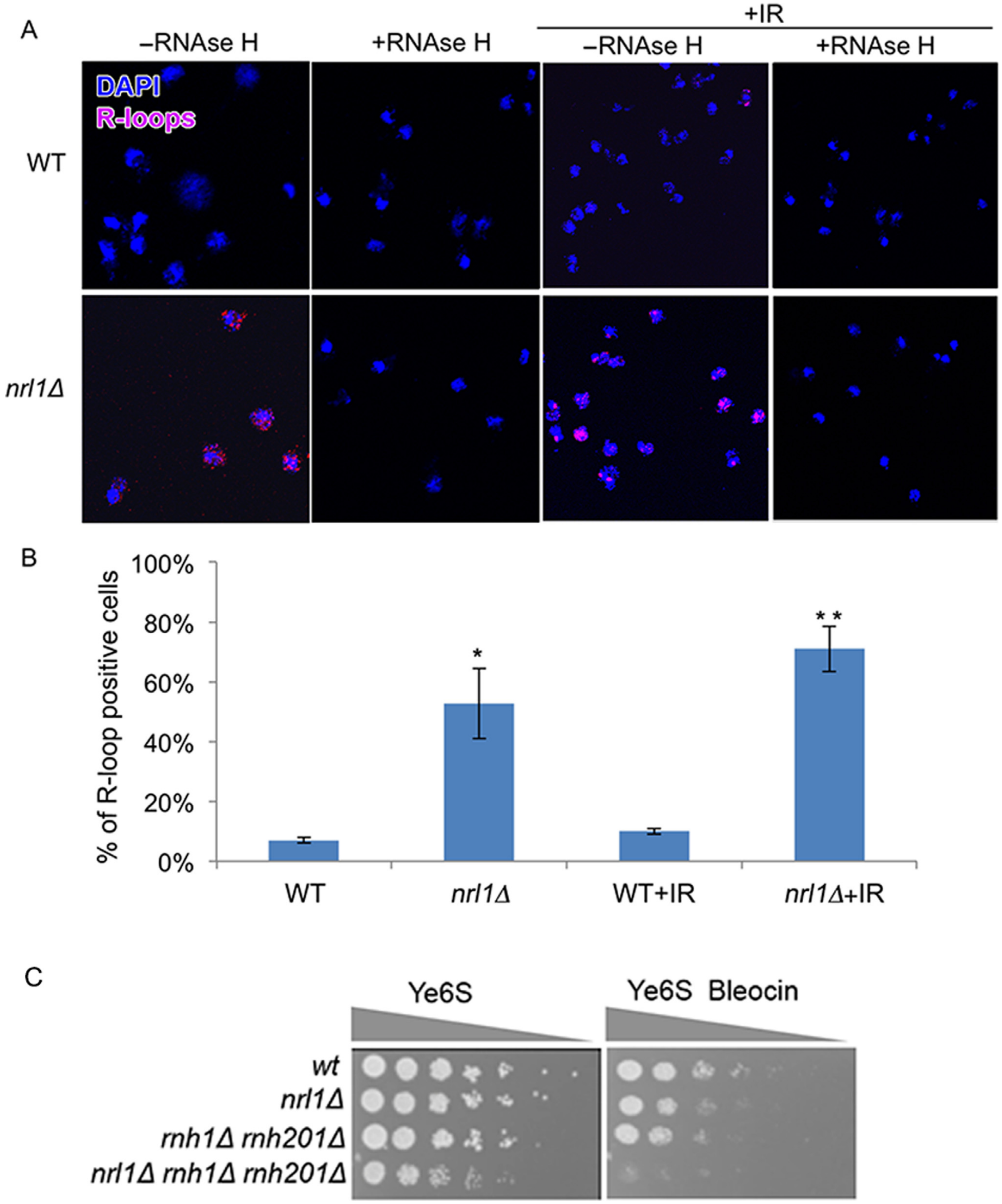


Figure 6. Nrl1 prevents R-loop accumulation. (**A**) Immunofluorescence analysis of RNA–DNA hybrids in chromosome spreads from WT (TH8342) and *nrl1Δ* (16581) using the mouse monoclonal S6.9 antibody. As negative control, the spreads were pre-treated with RNase H (+RNase H) before immunostaining as previously described (34). +IR: The cells were exposed to 100 Gy of IR before immunostaining. (B) Quantification of the R-loop positive nuclei in A. Mean and standard deviation were scored from triplicate experiments, n > 200. The asterisks (*) indicate significant differences compared with WT as determined by paired T-test (*$P = 0.01$, **$P = 0.003$). (**C**) *nrl1Δ* becomes hypersensitive to bleocin in the absence of Rnh1 and Rnh201. Fivefold serial dilution of *nrl1Δ* (TH8341) *rnh1Δ rnh201Δ* (TH8743) *and nrl1Δ rnh1Δ rnh201Δ* (TH8904) in the absence and presence of bleocin (0.2 μg/ml).

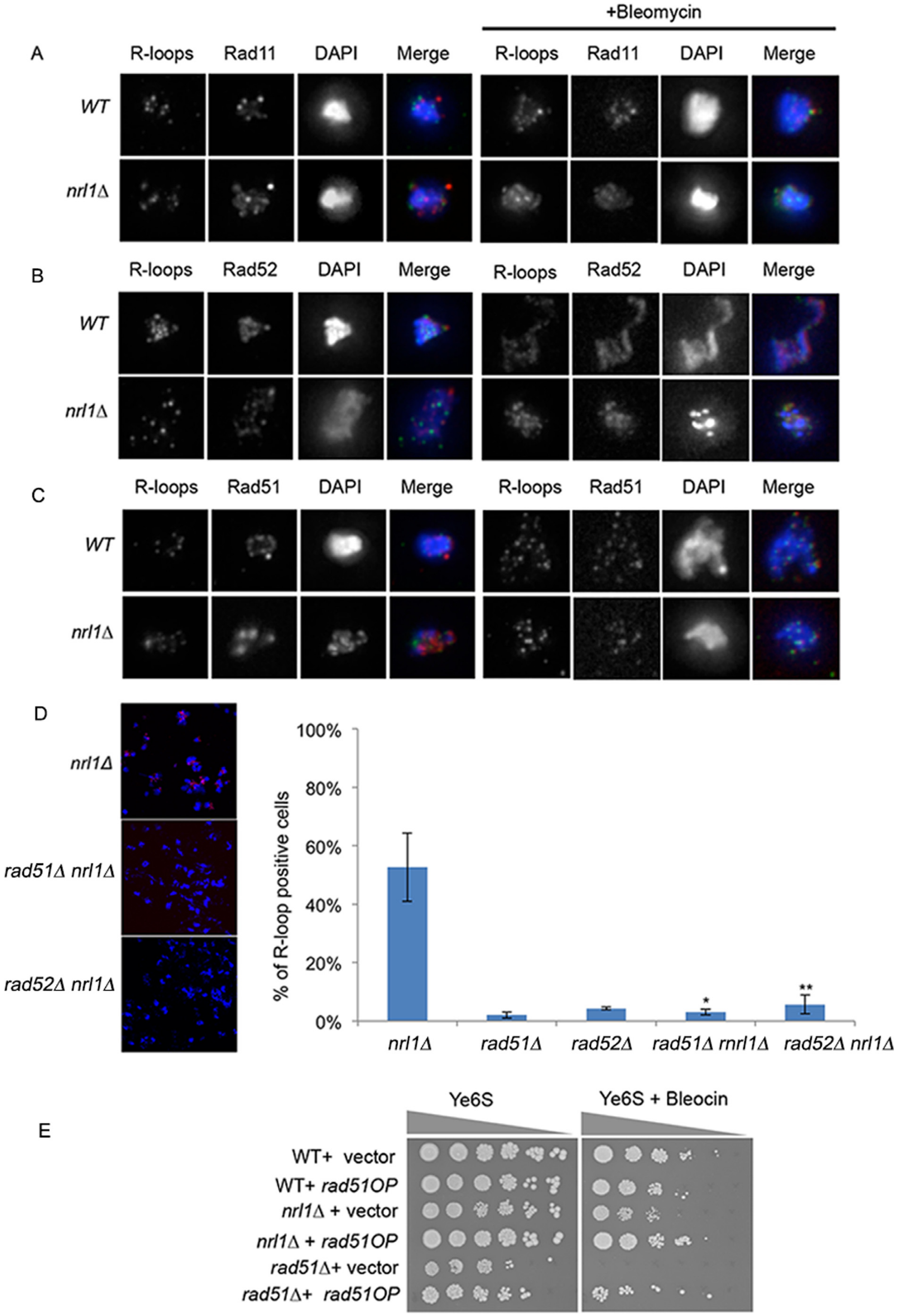


Figure 7. R-loops associate with and are dependent on HR factors in *nrl1Δ*. (**A**) Immunostaining images of RNA/DNA hybrids in relation to rad11-GFP foci in a wild-type (TH2151) and *nrl1Δ* (TH8125) in the absence (left panels) and presence (right panels) of bleomycin (3 μM) for 4 h at 25 °C. (**B**) Immunostaining images of RNA/DNA hybrids in relation to Rad52-YFP foci in a wild-type (TH8096) and *nrl1Δ* (TH8097) in the absence (left panels) and presence (right panels) of bleomycin (**C**) Immunostaining images of RNA/DNA hybrids in relation to Rad51-eCFP foci in a wild-type (TH3607) and *nrl1Δ* (TH8123) in the absence (left panels) and presence (right panels) of bleomycin. Bars, 5 μm. (**D**) Quantification of R-loop positive nuclei in *nrl1Δ*, *rad51Δ, rad52Δ, rad51Δ nrl1Δ* and *rad52Δ nrl1Δ*. Mean and standard deviation were scored from triplicate experiments, $n > 200$. The asterisks (*) indicate significant differences compared with *nrl1Δ* as determined by paired T-test ($*P = 0.01$, $**P = 0.02$). (**E**) Overexpression of Rad51 (Rad51OP) suppresses the bleocin sensitivity of *nrl1Δ*. Five-fold serial dilution of wild-type (TH351), *nrl1Δ* (TH8341) and *rad51Δ* (TH2801) strains transformed with either pIRT3 (vector) or pIRT3-Rad51 (Rad51OP) as indicated, and spotted onto YE6S or YE6S in the presence of 0.3 μg/ml bleocin.

both efficient pre-mRNA splicing and maintaining genome stability through the suppression of R-loops and the promotion of efficient HR repair.

We found that Nrl1 forms a core complex with the mRNA processing factors Mtl1 and Ctr1 and the splicing factors Ntr2 and Syf3, which mediate the interaction between the core complex and the spliceosome. Moreover, deletion of *nrl1*$^{+}$ led to significant changes in splicing patterns at several genomic loci, thus identifying a link between Nrl1 and pre-mRNA splicing. A possible role of Nrl1 in pre-mRNA splicing is also supported by a study from Lee and colleagues, in which Nrl1 was shown to interact with splicing factors and regulate the splicing of non-annotated introns of several developmental genes and retrotransposons (39). While our findings broadly concur with those of Lee *et al.*, who reported that *nrl1* deletion affected the splicing of 135 non-annotated introns, we detected a lower number of differentially spliced introns in our study, with significant changes in splicing at 10 newly identified non-annotated introns out of 43 introns differentially spliced in *nrl1Δ* compared with wild-type. These differences may reflect differences in growth media, bioinformatic selection criteria (see Methods) and the fact that Lee *et al.* used a different genetic background (*nrl1Δ rrp6Δ*) for their study. These findings raise important questions as to how Nrl1 functionally interacts with the splicing machinery to influence the splicing of this subset of introns. Whether Nrl1 binds directly to the affected pre-mRNAs, and how its loss interferes with splicing dynamics will be the subject of future studies.

Splicing mutants have previously been shown to promote R-loop formation (19). In line with a role for Nrl1 in pre-mRNA splicing, we additionally found that loss of Nrl1 resulted in a remarkably high degree of R-loop formation. In this respect, Nrl1 may suppress R-loops by ensuring timely processing of pre-mRNAs, thus reducing their ability to rehybridize with the DNA template, as has been shown for the splicing factor ASF/SF2 (19). Concerning the nature of R-loop-induced DNA damage in *nrl1Δ*, our data suggest that the displaced ssDNA at R-loops may facilitate DNA damage checkpoint activation thus leading to cell elongation, accumulation of endogenous Rad52 foci and a DNA damage-like transcriptional response. It would be interesting to determine the genomic distribution of R-loops accumulating in *nrl1Δ,* as R-loops form at different loci in different RNA processing mutants. While wild-type cells accumulate R-loops on actively transcribed protein-coding genes and ribosomal DNA regions, RNase H mutants display high levels of R-loops at tRNA genes, retrotransposons and mitochondrial genes *in S. cerevisiae* (51). In contrast, defects in the R-loop helicase Sen1 induces R-loop formation on short and actively transcribed genes, in line with the transcription termination function of Sen1 at these loci (52,53). To provide further functional insights into how Nrl1 suppresses R-loop formation it would be interesting to determine whether R-loops in *nrl1Δ* cells form at splice sites, such as those detected in this study and reported by Lee *et al.* (39) or at highly transcribed genes in the presence or absence of DNA damage.

In addition to its role in R-loop suppression, we found that Nrl1 is required for genome mainteinance. *nrl1Δ* displayed sensitivity to the DNA damaging agents MMS, bleomycin and CPT and defective DSB repair by HR- with significantly reduced gene conversion, increased chromosome loss and extensive chromosomal rearrangements leading to loss of heterozygosity, as determined by the DSB assay. Moreover, *nrl1Δ* displayed prolonged accumulation of Rad11 and Rad52 foci, and reduced formation of Rad51 foci following IR compared with wild-type. Concomitantly, we found that exposure to DNA damage results in significantly increased levels of both R-loops and their colocalization with Rad11, Rad52 and Rad51 in *nrl1Δ*. Finally, consistent with these findings, we identified a role for the HR machinery in facilitating R-loop formation in *nrl1Δ* cells, with loss of either Rad51 or Rad52 abrogating R-loop formation.

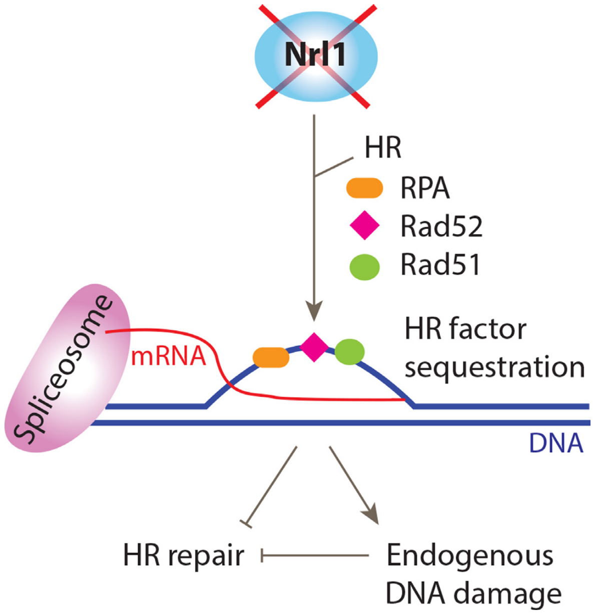


Figure 8. Model depicting the impact of Nrl1 loss on HR repair and genome stability. Loss of Nrl1 results in inefficient splicing, leading to HR-dependent R-loop formation and endogenous DNA damage. HR factors are sequestered at R-loops or associated sites of endogenous DNA damage resulting in compromised HR repair of exogenous DNA damage and genome instability. See text for details.

From our findings we propose the following model to explain the role of Nrl1 in R-loop suppression, genome stability and HR repair: changes in pre-mRNA processing, arising either directly or indirectly through loss of Nrl1, result in increased R-loop accumulation. HR factors facilitate this process, and are hence sequestered to sites of R-loop formation and/or R-loop induced DNA damage. This, in turn, leads to both elevated levels of endogenous DNA damage and defects in HR-repair of exogenous DNA insults (Figure 8). In support of this model, *rad51* overexpression alleviates the sensitivity of *nrl1Δ* to genotoxic agents, suggesting that an excess of Rad51 may increase the pool of free proteins, thus rescuing the function of the sequestered form. This sequestration model explains why Nrl1 is required for efficient HR repair despite not interacting directly with HR factors nor affecting their expression or splicing. This model also provides a unifying mechanism to explain the composite genome instability phenotype of *nrl1Δ*. In this respect, the finding that R-loop formation increases after IR exposure in *nrl1Δ* may reflect an increased HR activity following DNA damage, with subsequent promotion of R-loop formation and/or stability. Consistent with this is the observation that the levels of colocalization between HR factors and R-loops significantly increase upon DNA damage in *nrl1Δ*. Alternatively, the increased levels of R-loops following IR may result directly from DNA damage-induced blocking of RNA polymerase progression, which may potentially increase R-loop formation in the presence of defective splicing in *nrl1Δ*. In agreement with a role for HR factors in R-loop formation is also the aberrant pattern of HR protein foci observed in *nrl1Δ* upon IR exposure- with prolonged accumulation of Rad11 and Rad52 foci, and reduced formation of Rad51 foci. We speculate that while Rad11, Rad52 and Rad51 are similarly recruited to nascent R-loops to facilitate their formation, only Rad11 and Rad52 may bind to secondary lesions arising from R-loops. These regions are likely to consist of ssDNA stretches, and may no longer associate with Rad51 either because they are not substrates, or because they are subject to repair through a Rad52-dependent and Rad51-independent pathway such as single strand annealing (SSA). In favor of this hypothesis, Rad52 co-localizes with R-loops upon DNA damage only in *nrl1Δ*, which is consistent with the increased levels of endogenous Rad52 foci and possibly indicates persistent binding at R-loop associated DNA lesions. In addition, *nrl1Δ rad52Δ* double mutants exhibit increased sensitivity to bleomycin compared to the parental strains. As Rad52 is required for R-loop formation in *nrl1Δ*, this increased sensitivity may arise from the formation of a genotoxic R-loop precursor in *nrl1Δ rad52Δ* cells, which cannot be repaired in the absence of Rad52. Notably, R-loop induced DNA damage in *sen1 S. cerevisiae* mutants is also associated with Rad52 foci accumulation and repair through the HR pathway (53).

These observations are in line with previous findings that mutations in pre-mRNA processing factors can result in R-loop formation (10,11,19) and genome instability (5,11–15), and that this process is dependent on Rad51 and Rad52 in *S. cerevisiae* (22). Therefore, our data suggest an evolutionarily conserved role for HR in facilitating R-loop formation in pre-mRNA processing mutants. Further, our findings that deletion of the RNase H1 and H2 genes sensitized *nrl1Δ* cells to bleomycin closely mirror those of Lazzaro *et al.* (2012), which show that budding yeast strains lacking both RNase H1 and H2 as well as Rad51 (*rnh1Δ rnh201Δ rad51Δ*) are sensitive to genotoxins, and that loss of RAD52 is lethal in *rnh1Δ rnh201Δ* strains (54). Finally, a recent study by Keskin *et al.* identified a similar link between R-loops and the HR machinery by showing that RNase H not only degrades mutagenic R-loops but also inhibits RNA-templated HR repair (55). RNA–DNA hybrids may thus act as a double-edged sword in HR repair: they may promote HR repair through RNA-templated HR repair across a break-site as reported by Keskin *et al.*, but they may also compromise it when present at multiple different genomic loci by sequestering HR factors as proposed here.

Taken together, our data provide the first evidence that the spliceosome-associated factor Nrl1 can promote both HR repair and R-loop suppression, and suggest a mechanism of genome instability through R-loop mediated sequestration of HR factors. According to our model, the emerging yet elusive function of R-loops in tumor development may underlie both direct promotion of DNA damage, as shown previously (5), and indirect inhibition of HR repair. These findings therefore suggest how the human homolog of Nrl1 is implicated in cancer (25,26), and provide mechanistic insights into the oncogenic effects of R-loops (11–15).

ACCESSION NUMBERS

The mass spectrometry (MS) proteomics data have been deposited to the ProteomeXchange Consortium (56) via the PRIDE partner repository (http://www.pride-project.eu) with the data set identifier PXD001583. The RNA-Seq data are available in the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena/) under accession PRJEB7937.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank N. Proudfoot and L. Székvölgyi for antibodies (S9.6). We thank Margarita Schlackow and Monika Gullerova for helpful discussions, and Hannah Mischo for critical reading of the manuscript. The data deposition to the ProteomeXchange Consortium was supported by PRIDE Team, EBI.

FUNDING

Austrian Science Fund (FWF) [T527-B11 to L.A.; F43-P10 to A.B.]; Medical Research Council [MC_PC_12003 to T.H., S.S., C.C.P., E.B., C.W.]; Cancer Research UK [H3RPUX00 to T.K.; People Programme (Marie Curie Actions) European Union's Seventh Framework Programme [REA grant agreement No. 609427]; Slovak Academy of Sciences; Slovak Research and Development Agency [contract no. APVV-0111–12 and VEGA grant no. 2/0014/14 to L.C.]; National Institute of Health (NIH) [R01-GM059321 and R01-GM111040 to S.L.F.].

Conflict of interest statement. None declared.

REFERENCES

1. Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.
2. Malkova,A. and Haber,J.E. (2012) Mutations arising during repair of chromosome breaks. *Annu. Rev. Genet.*, **46**, 455–473.
3. Skourti-Stathaki,K. and Proudfoot,N.J. (2014) A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. *Genes Dev.*, **28**, 1384–1396.
4. Aguilera,A. and Garcia-Muse,T. (2012) R loops: from transcription byproducts to threats to genome stability. *Mol. Cell*, **46**, 115–124.
5. Hamperl,S. and Cimprich,K.A. (2014) The contribution of co-transcriptional RNA:DNA hybrid structures to DNA damage and genome instability. *DNA Repair (Amst)*, **19**, 84–94.
6. Ginno,P.A., Lim,Y.W., Lott,P.L., Korf,I. and Chedin,F. (2013) GC skew at the 5′ and 3′ ends of human genes links R-loop formation to epigenetic regulation and transcription termination. *Genome Res.*, **23**, 1590–1600.

7. Sun,Q., Csorba,T., Skourti-Stathaki,K., Proudfoot,N.J. and Dean,C. (2013) R-loop stabilization represses antisense transcription at the Arabidopsis FLC locus. *Science*, **340**, 619–621.
8. Chan,Y.A., Aristizabal,M.J., Lu,P.Y., Luo,Z., Hamza,A., Kobor,M.S., Stirling,P.C. and Hieter,P. (2014) Genome-wide profiling of yeast DNA:RNA hybrid prone sites with DRIP-chip. *PLoS Genet.*, **10**, e1004288.
9. Skourti-Stathaki,K., Proudfoot,N.J. and Gromak,N. (2011) Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. *Mol. Cell*, **42**, 794–805.
10. Chan,Y.A., Hieter,P. and Stirling,P.C. (2014) Mechanisms of genome instability induced by RNA-processing defects. *Trends Genet.*, **30**, 245–253.
11. Stirling,P.C., Chan,Y.A., Minaker,S.W., Aristizabal,M.J., Barrett,I., Sipahimalani,P., Kobor,M.S. and Hieter,P. (2012) R-loop-mediated genome instability in mRNA cleavage and polyadenylation mutants. *Genes Dev.*, **26**, 163–175.
12. Bhatia,V., Barroso,S.I., Garcia-Rubio,M.L., Tumini,E., Herrera-Moyano,E. and Aguilera,A. (2014) BRCA2 prevents R-loop accumulation and associates with TREX-2 mRNA export factor PCID2. *Nature*, **511**, 362–365.
13. Chernikova,S.B., Razorenova,O.V., Higgins,J.P., Sishc,B.J., Nicolau,M., Dorth,J.A., Chernikova,D.A., Kwok,S., Brooks,J.D., Bailey,S.M. *et al.* (2012) Deficiency in mammalian histone H2B ubiquitin ligase Bre1 (Rnf20/Rnf40) leads to replication stress and chromosomal instability. *Cancer Res.*, **72**, 2111–2119.
14. Ruiz,J.F., Gomez-Gonzalez,B. and Aguilera,A. (2011) AID induces double-strand breaks at immunoglobulin switch regions and c-MYC causing chromosomal translocations in yeast THO mutants. *PLoS Genet.*, **7**, e1002009.
15. Jackson,B.R., Noerenberg,M. and Whitehouse,A. (2014) A novel mechanism inducing genome instability in Kaposi's sarcoma-associated herpesvirus infected cells. *PLoS Pathog.*, **10**, e1004098.
16. Paulsen,R.D., Soni,D.V., Wollman,R., Hahn,A.T., Yee,M.C., Guan,A., Hesley,J.A., Miller,S.C., Cromwell,E.F., Solow-Cordero,D.E. *et al.* (2009) A genome-wide siRNA screen reveals diverse cellular processes and pathways that mediate genome stability. *Mol. Cell*, **35**, 228–239.
17. Lenzken,S.C., Loffreda,A. and Barabino,S.M. (2013) RNA splicing: a new player in the DNA damage response. *Int. J. Cell Biol.*, **2013**, 153634.
18. Sato,Y., Yoshizato,T., Shiraishi,Y., Maekawa,S., Okuno,Y., Kamura,T., Shimamura,T., Sato-Otsubo,A., Nagae,G., Suzuki,H. *et al.* (2013) Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat. Genet.*, **45**, 860–867.
19. Li,X. and Manley,J.L. (2005) Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. *Cell*, **122**, 365–378.
20. Rajesh,C., Baker,D.K., Pierce,A.J. and Pittman,D.L. (2011) The splicing-factor related protein SFPQ/PSF interacts with RAD51D and is necessary for homology-directed repair and sister chromatid cohesion. *Nucleic Acids Res.*, **39**, 132–145.
21. Adamson,B., Smogorzewska,A., Sigoillot,F.D., King,R.W. and Elledge,S.J. (2012) A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response. *Nat. Cell Biol.*, **14**, 318–328.
22. Wahba,L., Gore,S.K. and Koshland,D. (2013) The homologous recombination machinery modulates the formation of RNA-DNA hybrids and associated chromosome instability. *Elife*, **2**, e00505.
23. Savage,K.I., Gorski,J.J., Barros,E.M., Irwin,G.W., Manti,L., Powell,A.J., Pellagatti,A., Lukashchuk,N., McCance,D.J., McCluggage,W.G. *et al.* (2014) Identification of a BRCA1-mRNA splicing complex required for efficient DNA repair and maintenance of genomic stability. *Mol. Cell*, **54**, 445–459.
24. Marechal,A., Li,J.M., Ji,X.Y., Wu,C.S., Yazinski,S.A., Nguyen,H.D., Liu,S., Jimenez,A.E., Jin,J. and Zou,L. (2014) PRP19 transforms into a sensor of RPA-ssDNA after DNA damage and drives ATR activation via a ubiquitin-mediated circuitry. *Mol Cell*, **53**, 235–246.
25. Jones,K.B., Salah,Z., Del Mare,S., Galasso,M., Gaudio,E., Nuovo,G.J., Lovat,F., LeBlanc,K., Palatini,J., Randall,R.L. *et al.* (2012) miRNA signatures associate with pathogenesis and progression of osteosarcoma. *Cancer Res.*, **72**, 1865–1877.
26. Chiu,C.G., Nakamura,Y., Chong,K.K., Huang,S.K., Kawas,N.P., Triche,T., Elashoff,D., Kiyohara,E., Irie,R.F., Morton,D.L. *et al.* (2014) Genome-wide characterization of circulating tumor cells identifies novel prognostic genomic alterations in systemic melanoma metastasis. *Clin. Chem.*, **60**, 873–885.
27. Gregan,J., Rabitsch,P.K., Rumpf,C., Novatchkova,M., Schleiffer,A. and Nasmyth,K. (2006) High-throughput knockout screen in fission yeast. *Nat. Protoc.*, **1**, 2457–2464.
28. Cipak,L., Spirek,M., Novatchkova,M., Chen,Z., Rumpf,C., Lugmayr,W., Mechtler,K., Ammerer,G., Csaszar,E. and Gregan,J. (2009) An improved strategy for tandem affinity purification-tagging of *Schizosaccharomyces pombe* genes. *Proteomics*, **9**, 4825–4828.
29. Moss,J., Tinline-Purvis,H., Walker,C.A., Folkes,L.K., Stratford,M.R., Hayles,J., Hoe,K.L., Kim,D.U., Park,H.O., Kearsey,S.E. *et al.* (2010) Break-induced ATR and Ddb1-Cul4(Cdt)(2) ubiquitin ligase-dependent nucleotide synthesis promotes homologous recombination repair in fission yeast. *Genes Dev.*, **24**, 2705–2716.
30. Cipak,L., Gupta,S., Rajovic,I., Jin,Q.W., Anrather,D., Ammerer,G., McCollum,D. and Gregan,J. (2013) Crosstalk between casein kinase II and Ste20-related kinase Nak1. *Cell Cycle*, **12**, 884–888.
31. Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, **26**, 1367–1372.
32. Zhong,S., Joung,J.G., Zheng,Y., Chen,Y.R., Liu,B., Shao,Y., Xiang,J.Z., Fei,Z. and Giovannoni,J.J. (2011) High-throughput illumina strand-specific RNA sequencing library preparation. *Cold Spring Harb. Protoc.*,**8**, 940–949.
33. Loidl,J. and Lorenz,A. (2009) Analysis of *Schizosaccharomyces pombe* meiosis by nuclear spreading. *Methods Mol. Biol.*, **558**, 15–36.
34. Wahba,L., Amon,J.D., Koshland,D. and Vuica-Ross,M. (2011) RNase H and multiple RNA biogenesis factors cooperate to prevent RNA:DNA hybrids from generating genome instability. *Mol. Cell*, **44**, 978–988.
35. Guang,S., Bochner,A.F., Burkhart,K.B., Burton,N., Pavelec,D.M. and Kennedy,S. (2010) Small regulatory RNAs inhibit RNA polymerase II during the elongation phase of transcription. *Nature*, **465**, 1097–1101.
36. Ben-Yehuda,S., Dix,I., Russell,C.S., McGarvey,M., Beggs,J.D. and Kupiec,M. (2000) Genetic and physical interactions between factors involved in both cell cycle progression and pre-mRNA splicing in *Saccharomyces cerevisiae*. *Genetics*, **156**, 1503–1517.
37. Rigaut,G., Shevchenko,A., Rutz,B., Wilm,M., Mann,M. and Seraphin,B. (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.*, **17**, 1030–1032.
38. Chen,W., Shulha,H.P., Ashar-Patel,A., Yan,J., Green,K.M., Query,C.C., Rhind,N., Weng,Z. and Moore,M.J. (2014) Endogenous U2.U5.U6 snRNA complexes in *S. pombe* are intron lariat spliceosomes. *RNA*, **20**, 308–320.
39. Lee,N.N., Chalamcharla,V.R., Reyes-Turcu,F., Mehta,S., Zofall,M., Balachandran,V., Dhakshnamoorthy,J., Taneja,N., Yamanaka,S., Zhou,M. *et al.* (2013) Mtr4-like protein coordinates nuclear RNA processing for heterochromatin assembly and for telomere maintenance. *Cell*, **155**, 1061–1074.
40. Caspari,T. and Carr,A.M. (1999) DNA structure checkpoint pathways in *Schizosaccharomyces pombe*. *Biochimie*, **81**, 173–181.
41. Walworth,N.C. and Bernards,R. (1996) rad-dependent response of the chk1-encoded protein kinase at the DNA damage checkpoint. *Science*, **271**, 353–356.
42. Latif,C., den Elzen,N.R. and O'Connell,M.J. (2004) DNA damage checkpoint maintenance through sustained Chk1 activity. *J. Cell Sci.*, **117**, 3489–3498.
43. Kim,W.J., Lee,S., Park,M.S., Jang,Y.K., Kim,J.B. and Park,S.D. (2000) Rad22 protein, a rad52 homologue in *Schizosaccharomyces pombe*, binds to DNA double-strand breaks. *J. Biol. Chem.*, **275**, 35607–35611.
44. Lisby,M., Mortensen,U.H. and Rothstein,R. (2003) Colocalization of multiple DNA double-strand breaks at a single Rad52 repair centre. *Nat. Cell Biol.*, **5**, 572–577.
45. van den Bosch,M., Vreeken,K., Zonneveld,J.B., Brandsma,J.A., Lombaerts,M., Murray,J.M., Lohman,P.H. and Pastink,A. (2001)

Characterization of RAD52 homologs in the fission yeast *Schizosaccharomyces pombe*. *Mutat. Res.*, **461**, 311–323.

46. Tinline-Purvis,H., Savory,A.P., Cullen,J.K., Dave,A., Moss,J., Bridge,W.L., Marguerat,S., Bahler,J., Ragoussis,J., Mott,R. *et al.* (2009) Failed gene conversion leads to extensive end processing and chromosomal rearrangements in fission yeast. *EMBO J.*, **28**, 3400–3412.
47. Wold,M.S. (1997) Replication protein A: a heterotrimeric, single-stranded DNA-binding protein required for eukaryotic DNA metabolism. *Annu. Rev. Biochem.*, **66**, 61–92.
48. Sugawara,N., Wang,X. and Haber,J.E. (2003) In vivo roles of Rad52, Rad54, and Rad55 proteins in Rad51-mediated recombination. *Mol. Cell*, **12**, 209–219.
49. Watson,A., Mata,J., Bahler,J., Carr,A. and Humphrey,T. (2004) Global gene expression responses of fission yeast to ionizing radiation. *Mol. Biol. Cell*, **15**, 851–860.
50. Boguslawski,S.J., Smith,D.E., Michalak,M.A., Mickelson,K.E., Yehle,C.O., Patterson,W.L. and Carrico,R.J. (1986) Characterization of monoclonal antibody to DNA.RNA and its application to immunodetection of hybrids. *J. Immunol. Methods*, **89**, 123–130.
51. El Hage,A., Webb,S., Kerr,A. and Tollervey,D. (2014) Genome-wide distribution of RNA-DNA hybrids identifies RNase H targets in tRNA genes, retrotransposons and mitochondria. *PLoS Genet.*, **10**, e1004716.
52. Steinmetz,E.J., Warren,C.L., Kuehner,J.N., Panbehi,B., Ansari,A.Z. and Brow,D.A. (2006) Genome-wide distribution of yeast RNA polymerase II and its control by Sen1 helicase. *Mol. Cell*, **24**, 735–746.
53. Mischo,H.E., Gomez-Gonzalez,B., Grzechnik,P., Rondon,A.G., Wei,W., Steinmetz,L., Aguilera,A. and Proudfoot,N.J. (2011) Yeast Sen1 helicase protects the genome from transcription-associated instability. *Mol. Cell*, **41**, 21–32.
54. Lazzaro,F., Novarina,D., Amara,F., Watt,D.L., Stone,J.E., Costanzo,V., Burgers,P.M., Kunkel,T.A., Plevani,P. and Muzi-Falconi,M. (2012) RNase H and postreplication repair protect cells from ribonucleotides incorporated in DNA. *Mol Cell*, **45**, 99–110.
55. Keskin,H., Shen,Y., Huang,F., Patel,M., Yang,T., Ashley,K., Mazin,A.V. and Storici,F. (2014) Transcript-RNA-templated DNA recombination and repair. *Nature*, **515**, 436–439.
56. Vizcaino,J.A., Deutsch,E.W., Wang,R., Csordas,A., Reisinger,F., Rios,D., Dianes,J.A., Sun,Z., Farrah,T., Bandeira,N. *et al.* (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.*, **32**, 223–226.

Conclusions

As stated in the introduction, the basis for complex systems is not a complex starting material, but rather extensive interaction networks build on basic elements using simple rules. RNA transcription and processing are among the most potent cellular processes in respect to diversification and regulatory potential. The work presented here adds a previously unknown layer of regulatory interactions to this network.

The impact of intrasplicing on shaping the transcriptome has been shown with a set of examples, yet the extend of the obtained intra-splicing dataset implies that intrasplicing is a commonly induced (or occurring) event that has the potential of being a global regulator of gene expression. As this finding opens up new implications, it also poses more questions towards the regulation of these splicing events. How many of them are regulated and how many a random mis-splicing event? Is mis-splicing still mis-splicing when it is used to actively down-regulate expression of a certain gene? How many of the identified recursive splicing events are a side-product of high U1-occupancy in introns due to PCPA-prevention? And how many subsequent recursive splicing patterns are merely a fail-safe to rescue these transcripts, whose introns have been prematurely committed to splicing? Future studies will provide answers to these questions, leading to a more complete, yet more complex picture of the splicing processes and their impacts in cellular environments.

The second manuscript sheds light on the transcriptional processes at genomic centromeres and the interaction of the resulting α-satellite RNA and polII. This interaction results in RNA dependant RNA polymerase activity and a potential role of satellite RNA as a transcription regulator.

A similar interaction is observed in the ACRO-derived RNA, where small RNA aptamers were found to serve as a silencer of ACRO transcription. Besides this activity, these motifs have the potential to tune RNA splicing, as their positioning within introns could potentially lead to altered polymerase transcription speeds and thus influence splice site selection. The abundance and diversity of RNA polymerase binding aptamers identified in this study imply a novel and very diverse regulatory layer that sits directly at the core of transcription: the RNA itself.

The essential role of proteins in transcription and processing is showcased by Nrl1. Discovered by Lucia Aronica in yeast, it showed a close interaction with RNA processing and splicing factors. Its role in maintaining splicing patterns and genomic integrity, in complex interaction with other cellular factors, are exemplary for the diversity of regulatory networks in the cell.

REFERENCES

1. Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., Garc?a Gir?n, C., Hourlier, T., Howe, K., K?h?ri, A., Kokocinski, F., Martin, F. J., Murphy, D. N., Nag, R., Ruffier, M., Schuster, M., Tang, Y. A., *et al.* The Ensembl gene annotation system. *Database* **2016,** baw093 (2016).
2. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463,** 457–63 (2010).
3. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40,** 1413–1415 (2008).
4. Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P. & Burge, C. B. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–6 (2008).
5. Roeder, R. G. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* **21,** 327–35 (1996).
6. Coombes, C. E. & Boeke, J. D. An evaluation of detection methods for large lariat RNAs. *RNA* **11,** 323–31 (2005).
7. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* **17,** 487–500 (2016).
8. Espinoza, C. A., Allen, T. A., Hieb, A. R., Kugel, J. F. & Goodrich, J. A. B2 RNA binds directly to RNA polymerase II to repress transcript synthesis. *Nat. Struct. Mol. Biol.* **11,** 822–829 (2004).
9. Zimmermann, B., Bilusic, I., Lorenz, C. & Schroeder, R. Genomic SELEX: A discovery tool for genomic aptamers. *Methods* **52,** 125–132 (2010).
10. Wassarman, K. M. & Storz, G. 6S RNA regulates E. coli RNA polymerase activity. *Cell* **101,** 613–23 (2000).
11. Roy, D. & Lieber, M. R. G Clustering Is Important for the Initiation of Transcription-Induced R-Loops In Vitro, whereas High G Density without Clustering Is Sufficient Thereafter. *Mol. Cell. Biol.* **29,** 3124–3133 (2009).
12. Deem, A., Keszthelyi, A., Blackgrove, T., Vayl, A., Coffey, B., Mathur, R., Chabes, A. & Malkova, A. Break-induced replication is highly inaccurate. *PLoS Biol.* **9,** (2011).
13. Aronica, L., Kasparek, T., Ruchman, D., Marquez, Y., Cipak, L., Cipakova, I., Anrather, D., Mikolaskova, B., Radtke, M., Sarkar, S., Pai, C.-C., Blaikley, E., Walker, C., Shen, K.-F., Schroeder, R., Barta, A., Forsburg, S. L. & Humphrey, T. C. The spliceosome-associated protein Nrl1 suppresses homologous recombination-dependent R-loop formation in fission yeast.

Nucleic Acids Res. **44,** 1703–1717 (2016).

14. Cowling, V. H. Regulation of mRNA cap methylation. *Biochem. J.* **425,** 295–302 (2009).
15. Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 3171–5 (1977).
16. Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell* **12,** 1–8 (1977).
17. Roy, B., Haupt, L. M. & Griffiths, L. R. Review: Alternative Splicing (AS) of Genes As An Approach for Generating Protein Complexity. *Curr. Genomics* **14,** 182–94 (2013).
18. Mount, S. M., Pettersson, I., Hinterberger, M., Karmas, A. & Steitz, J. A. The U1 small nuclear RNA-protein complex selectively binds a 5??? splice site in vitro. *Cell* **33,** 509–518 (1983).
19. Matera, A. G. & Wang, Z. A day in the life of the spliceosome. *Nat. Rev. Mol. Cell Biol.* **15,** (2014).
20. Papasaikas, P. & Valcárcel, J. The Spliceosome: The Ultimate RNA Chaperone and Sculptor. *Trends Biochem. Sci.* **41,** 33–45 (2015).
21. De Conti, L., Baralle, M. & Buratti, E. Exon and intron definition in pre-mRNA splicing. *Wiley Interdisciplinary Reviews: RNA* **4,** (2013).
22. Wahl, M. C. & Lührmann, R. SnapShot: Spliceosome Dynamics I. *Cell* **161,** 1474–1474.e1 (2015).
23. Nguyen, T. H. D., Galej, W. P., Bai, X., Savva, C. G., Newman, A. J., Scheres, S. H. W. & Nagai, K. The architecture of the spliceosomal U4/U6.U5 tri-snRNP. *Nature* **523,** 47–52 (2015).
24. Nguyen, T. H. D., Galej, W. P., Fica, S. M., Lin, P.-C., Newman, A. J. & Nagai, K. CryoEM structures of two spliceosomal complexes: starter and dessert at the spliceosome feast. *Curr. Opin. Struct. Biol.* **36,** 48–57 (2016).
25. Golas, M. M., Sander, B., Bessonov, S., Grote, M., Wolf, E., Kastner, B., Stark, H. & Lührmann, R. 3D Cryo-EM Structure of an Active Step I Spliceosome and Localization of Its Catalytic Core. *Mol. Cell* **40,** 927–938 (2010).
26. Nguyen, T. H. D., Galej, W. P., Bai, X., Oubridge, C., Newman, A. J., Scheres, S. H. W. & Nagai, K. Cryo-EM structure of the yeast U4/U6.U5 tri-snRNP at 3.7 Å resolution. *Nature* **530,** 298–302 (2016).
27. Galej, W. P., Wilkinson, M. E., Fica, S. M., Oubridge, C., Newman, A. J. & Nagai, K. Cryo-EM structure of the spliceosome immediately after branching. *Nature* **537,** 197–201 (2016).
28. Azubel, M., Wolf, S. G., Sperling, J. & Sperling, R. Three-Dimensional Structure of the Native Spliceosome by Cryo-Electron Microscopy. *Mol. Cell* **15,** 833–839 (2004).
29. Woodward, L. A., Mabin, J. W., Gangras, P. & Singh, G. The exon junction complex: a lifelong guardian of mRNA fate. *Wiley Interdiscip. Rev. RNA* e1411 (2016). doi:10.1002/wrna.1411

30. Wahl, M. C. & Lührmann, R. SnapShot: Spliceosome Dynamics II. *Cell* **162,** 456–456.e1 (2015).

31. Wahl, M. C. & Lührmann, R. SnapShot: Spliceosome Dynamics III. *Cell* **162,** (2015).

32. Dujardin, G., Lafaille, C., Petrillo, E., Buggiano, V., Gómez Acuña, L. I., Fiszbein, A., Godoy Herz, M. a, Nieto Moreno, N., Muñoz, M. J., Alló, M., Schor, I. E. & Kornblihtt, A. R. Transcriptional elongation and alternative splicing. *Biochim. Biophys. Acta* **1829,** 134–40 (2013).

33. de la Mata, M., Alonso, C. R., Kadener, S., Fededa, J. P., Blaustein, M., Pelisch, F., Cramer, P., Bentley, D. & Kornblihtt, A. R. A slow RNA polymerase II affects alternative splicing in vivo. *Mol. Cell* **12,** 525–32 (2003).

34. McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblatt, J., Patterson, S. D., Wickens, M. & Bentley, D. L. The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature* **385,** 357–361 (1997).

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

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

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

38. Allemand, E., Myers, M. P., Garcia-Bernardo, J., Harel-Bellan, A., Krainer, A. R. & Muchardt, C. A Broad Set of Chromatin Factors Influences Splicing. *PLoS Genet.* **12,** (2016).

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

40. Xu, Q., Modrek, B. & Lee, C. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.* **30,** 3754–66 (2002).

41. Cáceres, J. F. & Kornblihtt, A. R. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet.* **18,** 186–93 (2002).

42. Kornblihtt, A. R., Schor, I. E., Alló, M., Dujardin, G., Petrillo, E. & Muñoz, M. J. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.* **14,** 153–65 (2013).

43. Hubé, F. & Francastel, C. Mammalian Introns: When the Junk Generates Molecular Diversity. *Int. J. Mol. Sci.* **16,** 4429–4452 (2015).

44. Hattori, D., Millard, S. S., Wojtowicz, W. M. & Zipursky, S. L. Dscam-mediated cell recognition regulates neural circuit formation. *Annu. Rev. Cell Dev. Biol.* **24,** 597–620 (2008).

45. Zipursky, S. L. & Grueber, W. B. The molecular basis of self-avoidance. *Annu. Rev. Neurosci.* **36,** 547–68 (2013).

46. Tress, M. L., Abascal, F. & Valencia, A. Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends Biochem. Sci.* **42,** 98–110 (2017).

47. Weatheritt, R. J., Sterne-Weiler, T. & Blencowe, B. J. The ribosome-engaged landscape of alternative splicing. *Nat. Struct. Mol. Biol.* (2016). doi:10.1038/nsmb.3317

48. Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).

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

50. Shepard, S., McCreary, M. & Fedorov, A. The peculiarities of large intron splicing in animals. *PLoS One* **4,** (2009).

51. Sibley, C. R., Emmett, W., Blazquez, L., Faro, A., Haberman, N., Briese, M., Trabzuni, D., Ryten, M., Weale, M. E., Hardy, J., Modic, M., Curk, T., Wilson, S. W., Plagnol, V. & Ule, J. Recursive splicing in long vertebrate genes. *Nature* (2015). doi:10.1038/nature14466

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

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

54. Osella, M. & Caselle, M. Entropic contributions to the splicing process. *Phys. Biol.* **6,** 46018 (2009).

55. Parra, M. K., Tan, J. S., Mohandas, N. & Conboy, J. G. Intrasplicing coordinates alternative first exons with alternative splicing in the protein 4.1R gene. *EMBO J.* **27,** 122–31 (2008).

56. Parra, M. K., Gallagher, T. L., Amacher, S. L., Mohandas, N. & Conboy, J. G. Deep intron elements mediate nested splicing events at consecutive AG dinucleotides to regulate alternative 3' splice site choice in vertebrate 4.1 genes. *Mol. Cell. Biol.* **32,** 2044–53 (2012).

57. OTT, S., TAMADA, Y., BANNAI, H., NAKAI, K. & MIYANO, S. INTRASPLICING - ANALYSIS OF LONG INTRON SEQUENCES. in *Biocomputing 2003* **339,** 339–350 (WORLD SCIENTIFIC, 2002).

58. Suzuki, H. Nested introns in an intron: evidence of multi-step splicing in a large intron of the human dystrophin pre-mRNA. Supplementary data. 1–2 (2013).

59. Lasda, E. L. & Blumenthal, T. Trans-splicing. *Wiley Interdisciplinary Reviews: RNA* **2,** 417–434 (2011).

60. Berg, M. G., Singh, L. N., Younis, I., Liu, Q., Pinto, A. M., Kaida, D., Zhang, Z., Cho, S., Sherrill-Mix, S., Wan, L. & Dreyfuss, G. U1 snRNP determines mRNA length and regulates isoform expression. *Cell* **150,** 53–64 (2012).

61. Clement, J. Q., Qian, L., Kaplinsky, N. & Wilkinson, M. F. The stability and fate of a spliced intron from vertebrate cells. *RNA* **5,** 206–20 (1999).

62. Ruskin, B. & Green, M. An RNA processing activity that debranches RNA lariats. *Science (80-.).* **229,** 135–140 (1985).

63. Kulesza, C. a & Shenk, T. Murine cytomegalovirus encodes a stable intron that facilitates persistent replication in the mouse. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 18302–18307 (2006).

64. Awan, A. R., Manfredo, A. & Pleiss, J. A. Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 12762–7 (2013).

65. Suzuki, H. & Tsukahara, T. A View of Pre-mRNA Splicing from RNase R Resistant RNAs. *Int. J. Mol. Sci.* **15,** 9331–42 (2014).

66. Taggart, A. J., DeSimone, A. M., Shih, J. S., Filloux, M. E. & Fairbrother, W. G. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nat. Struct. Mol. Biol.* **19,** 719–21 (2012).

67. Mercer, T. R., Clark, M. B., Andersen, S. B., Brunck, M. E., Haerty, W., Crawford, J., Taft, R. J., Nielsen, L. K., Dinger, M. E. & Mattick, J. S. Genome-wide discovery of human splicing branchpoints. *Genome Res.* gr.182899.114 (2015). doi:10.1101/gr.182899.114

68. Pulyakhina, I., Gazzoli, I., Hoen, P.-B. t., Verwey, N., den Dunnen, J., Aartsma-Rus, A. & Laros, J. SplicePie: a novel analytical approach for the detection of alternative, non-sequential and recursive splicing. *Nucleic Acids Res.* gkv242- (2015). doi:10.1093/nar/gkv242

69. Saulière, J., Murigneux, V., Wang, Z., Marquenet, E., Barbosa, I., Le Tonquèze, O., Audic, Y., Paillard, L., Roest Crollius, H. & Le Hir, H. CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. TL - 19. *Nat. Struct. Mol. Biol.* **19 VN-r,** 1124–1131 (2012).

70. Fedorova, L. & Fedorov, A. Puzzles of the Human Genome : Why Do We Need Our Introns ? 589–595 (2005).

71. Hollerer, I., Grund, K., Hentze, M. W. & Kulozik, A. E. mRNA 3'end processing: A tale of the tail reaches the clinic. *EMBO Mol. Med.* **6,** 16–26 (2014).

72. Kyburz, A., Friedlein, A., Langen, H. & Keller, W. Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3' end processing and splicing. *Mol. Cell* **23,** 195–205 (2006).

73. Lu, S. & Cullen, B. R. Analysis of the stimulatory effect of splicing on mRNA production and

utilization in mammalian cells. *RNA* **9,** 618–30 (2003).

74. Cooke, C., Hans, H. & Alwine, J. C. Utilization of splicing elements and polyadenylation signal elements in the coupling of polyadenylation and last-intron removal. *Mol. Cell. Biol.* **19,** 4971–9 (1999).

CONTRIBUTION TO MANUSCRIPTS

I am the first author of the manuscript “Genome-wide identification of intrasplicing events in the human transcriptome and hints to their regulatory potential”. I performed the bioinformatical and experimental work and wrote the manuscript.

I am co-author of the manuscript “Human α satellite transcripts are substrates for RNA Pol II and contain remnants of snoRNAs”. I contributed to the performing of the experiments.

I am co-author of the manuscript “RNA polymerase II-binding aptamers in human ACRO1 satellites disrupt transcription in *cis*”. I contributed to the performing of the experiments and writing of the manuscript.

I am co-author of the manuscript “The spliceosome-associated protein Nrl1 suppresses homologous recombination-dependent R-loop formation in fission yeast”. I contributed to the performing of the experiments.

ACKNOWLEDGEMENTS

I want to thank Renée for giving me the nowadays rare opportunity to simply research and showing us that other values than impact factors may matter.

I want to thank the members of the Schroeder Lab for good company and self defying bearing of really good music.

I want to thank my PhD committee members Jens Bohne, for his long-term support and our occasional but enjoyable conventions, and Andea Barta, for keeping splicing alive at the MFPL.

I want to thank Nicky, Zhara, Gerlinde and Paulina for making all of our lives easier.

I want to thank Corinna and Reiner for their support and ongoing attempts of understanding what I actually do.

And Vanessa, for joining me on yet another ride and true companionship.