universität wien

DISSERTATION

Titel der Dissertation

“Human alpha satellite-derived transcripts interact with RNA polymerase II”

Verfasserin

Mag. Katarzyna Matylla-Kulinska, M.Sc

angestrebter akademischer Grad

Doktorin der Naturwissenschaften (Dr.rer.nat.)

Wien, im April 2014

Studienkennzahl lt. Studienblatt:	A 091490
Dissertationsgebiet lt. Studienblatt:	Dr.- Studium der Naturwissenschaftern UniSt Molekulare Biologie
Betreuerin:	Prof. Dr. Renée Schroeder

Abstract

Whether repetitive regions of the human genome have a function is a very intriguing question. At least 50 % of the human genome is repetitive and following the ENCODE consortium report, up to 60-70 % of the genome is transcribed. As a consequence, the human transcriptome contains a significant fraction of repeat-derived transcripts. Human α satellites consist of 171 bp monomers arranged tandemly in a head-to-tail manner, organized into arrays of higher order repeats spanning hundreds of kilobases to megabases. They predominantly localize near centromeres on every human chromosome, where they assure proper chromosome segregation as a site for the spindle attachment.

Here, we demonstrate that α satellite arrays are transcribed from both DNA strands into long non-coding αsatRNAs (more than 8 kb). Their expression is more pronounced under cellular stress conditions and peak during the S phase of the cell cycle. We observe that the transcription rate of αsatRNAs might be down-regulated to the typical levels only one hour after stress release. Transcription of α satellites is sensitive to α amanitin treatment, indicating that they are RNA polymerase II transcripts. Examination of the 5' termini of αsatRNAs reveals that they possess a cap structure. However, unlike most RNA polymerase II transcripts, they are not polyadenylated and are retained in the nucleus. In order to grasp a putative αsatRNAs function, we searched for a protein interacting partner. In genomic SELEX using RNA polymerase II as bait, we had isolated several aptamers derived from α satellite repeats, suggesting that αsatRNAs interact with RNA polymerase II. Moreover, we found that αsatRNAs bind RNA polymerase II in the active site, serving as substrates for its activities. Using HeLa cells nucleofection with chimeric templates combining α satellite aptamer with an artificial 15-mer, we show that the α satellite RNA is a substrate for RNA-dependent RNA polymerase (RdRP) activity *in vivo*. In addition, we detected 5,6-dichloro-1-beta-D-ribofuranosyl-benzimidazole (DRB)-sensitive 3' extension of the αsatRNAs *in vitro*. Both activities are held by RNA polymerase II. As yet, we do not provide evidence that the products of these reactions are functional. However, we envision that the αsatRNAs-

RNA polymerase II interaction may control the transcriptional rate of functional centromeres. Upon studying the phylogeny and analyzing the secondary structure of α satellites and alphoid sequence from other primates, we realized that αsatRNAs not only fold into the hairpin-hinge-hairpin structure, similarly to snoRNAs, but also contain H/ACA boxes. Therefore, we propose that αsatRNAs are likely descendents of snoRNA mobilized by transposable elements.

Zusammenfassung

Es ist nach wie vor eine offene Frage ob repetitive Regionen des humanen Genoms eine Funktion haben. Diese Frage stellt sich nach zwei kürzlich beobachteten Fakten: 60 bis 70 % der humanen DNA wird in RNA überschrieben und die Hälfte des humanen Genoms ist repetitiv oder stammt von repetitiven Elementen ab. Humane α Satelliten sind 171 bp lange Einheiten, die sich in Tandem Kopf-Schwanz Anordnung zu langen Reihen ordnen und dabei mehrere Hunderttausend bis Millionen Basen umfassen. Man findet sie vor allem in der Nähe von Zentromeren auf jedem humanen Chromosom. Diese α Satelliten garantieren die korrekte Segregation der Chromosomen weil sie der Ort der Spindelanhaftung darstellen.

Hier zeigen wir, dass α Satelliten von beiden DNA Strängen in lange Transkripte (über 8 kb) überschrieben werden. Ihre Expression ist stärker unter Stressbedingungen and auf die S Phase des Zellzyklus beschränkt. Die Transkritption der α Satelliten ist α Amanitin sensitiv, was auf RNA polymerase II Transkritpe hinweist. Ihre 5' Enden weisen eine typische Cap-struktur auf. Anders als viele RNA polymerase II Transkripte sind sie nicht polyadenyliert und sie bleiben im Zellkern lokalisiert. Um einer Funktion dieser Transkripte näher zu kommen, haben wir Proteine als Bindungspartner gesucht. In einem genomischen SELEX Experiment mit RNA polymerase II als Köder haben wir mehrere RNA Aptamere aus α Satelliten erhalten. Weiterhin haben wir gefunden, dass RNAs aus α Satelliten mit dem aktiven Zentrum von RNA polymerase II interagieren und zu einer DRB (5,6-dichloro-1-beta-D-ribofuranosylbenzimidazole) sensitiven Markierung des 3' Endes und/oder einer Synthese des zweiten Stranges führen. Das suggeriert, dass die RNA polymerase II eine RNA-abhängige RNA Polymeraseaktivität (RdRP) an endogenen Substraten ausführen kann. Diese Aktivität haben *in vivo* mit HeLa Zellen und einer Nukleofektion mit chimeren Transkripten, welche α Satelliten RNA und artifiziellen Sequenzen beinhalten gezeigt. Dieses Experiment zeigt die RdRP Aktivität mit genomischer RNA als Templat. Diese Beobachtungen schlagen

vor, dass RNAs aus α Satelliten in Transkriptionsregulation involviert sein können.

Während phylogenetischer Studien um die Sekundärstruktur von α Satelliten und deren Konservierung zu analysieren, haben wir entdeckt, dass αsatRNAs nicht nur die Hairpin–Angel-Hairpin Struktur aufweisen, sondern auch die H/ACA Motive, die typisch für snoRNAs sind, beinhalten. Wir haben daher die Hypothese aufgestellt, dass αsatRNAs höchstwahrscheinlich von snoRNAs abstammen und durch Retroptranposons mobilisiert wurden.

Introduction

1.1 The human genome comprises mainly non-protein coding DNA

The genome of every living organism belonging to any of the three kingdoms of life is a unique, complete set of DNA that constitutes chromosomes and determines the individual's characteristics and traits. The human genome sequence was solved by the Human Genome Project and published in 2003 (Venter, 2003). The sequence was determined for 99 % of the human genomic content with very high quality of the assessment (Schmutz *et al*, 2004). The 1 % of sequences not determined was mainly derived from centromeres; because the low sequence variability of centromeric DNA and lack of unique sequences embedded in long centromeric arrays hinder sequencing and mapping techniques. The human genome project revealed that the human genome is composed of 3.3 billion bases, and 30 % of it is taken up by the canonical genes. However as little as 3 % of the genome encodes for proteins (specifically, coding exons), much less than previously expected (Figure 1). The corollary is that vast majority of the human genome is the non-coding content. Moreover, 70-90 % of the human DNA is reported to be transcribed in the process of development (Djebali *et al*, 2012; The ENCODE Project Consortium, 2007), giving rise to the highly complex, overlapping and intertwining network of the human transcriptome.

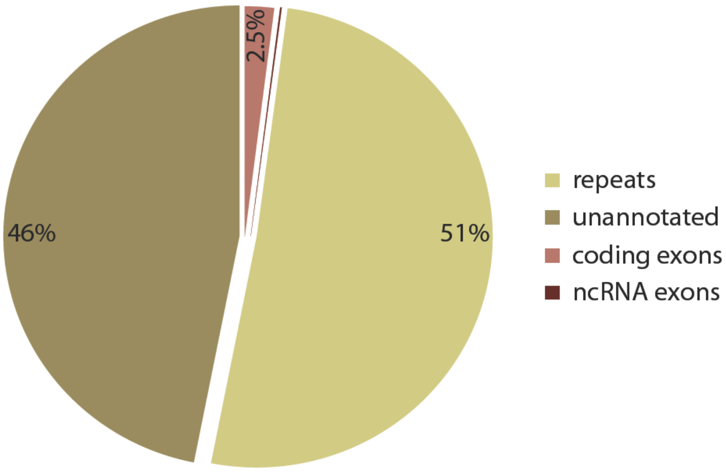


Figure 1 Composition of the human genome. The coding content comprises 3 % of the human genome. At least half of the genome is burdened with repetitive sequences. [Adapted from Matylla-Kulinska et al. 2014, *accepted*]

Recent efforts were aimed to investigate how much of this non-coding content is biologically important. The ENCODE ("ENCyclopedia Of DNA Elements") project claims that as much as 80 % of the human DNA is functional, understood as being transcribed into RNA, associated with regulatory complexes or it contributes to other biochemical activities (The ENCODE Project Consortium, 2007). However, this interpretation of the ENCODE results received broad criticism (Doolittle, 2013; Eddy, 2012). The main negative assessment focused on the definition of the functional elements since the existence of a transcript is not equivalent to its function. It is still a matter of lively debates as to what fraction of the reported transcripts is functional, what portion of the human transcriptome is transcriptional and biochemical noise, and how to delineate meaningful transcripts from the background.

1.1.1 Non-coding DNA is valuable

Surprisingly, in Eukaryotes, as opposed to the Prokaryotes, there is an inconsistent correlation between the genome size and its coding content. Large genomes are composed of a substantial fraction of non-coding DNA, consisting mainly of repetitive sequences. For a long time, there was no interest in looking for functional domains other than within protein-coding regions, because primarily only protein-coding DNA was considered as functional. Therefore non-coding DNA was perceived either as "junk" (Ohno, 1972) or as selfish "genomic parasites" (Orgel & Crick, 1980). However, comparative genomics has changed the perspective.

Upon the burst of large-scale genomic sequencing it became apparent that there is a positive correlation between the developmental complexity and the ratio of non-coding to protein-coding DNA (Figure 2). The portion of the protein-coding DNA declines linearly: in bacteria protein-coding content accounts for ~90 % of the genome, in yeast ~68 %, in insects ~17 %, in humans as little as ~2.5 % (Taft *et al*, 2007). The prevalent interpretation of these data is that the regulatory networks switched from the protein-triggered mode into the RNA-centred control. As a consequence non-protein coding content expands and takes up more and more of the genome. For example: promoter-enhancer regions of developmental genes expand, introns' lengths enlarge, or the length of 3' UTRs grow leaving room for *cis*-regulation (Taft *et al*, 2007; Taylor *et al*, 2006; Cheng *et al*, 2005). Moreover, Mattick suggests that the shift towards RNA-triggered regulation was essential for development of complex, multicellular organisms (Mattick, 2004).

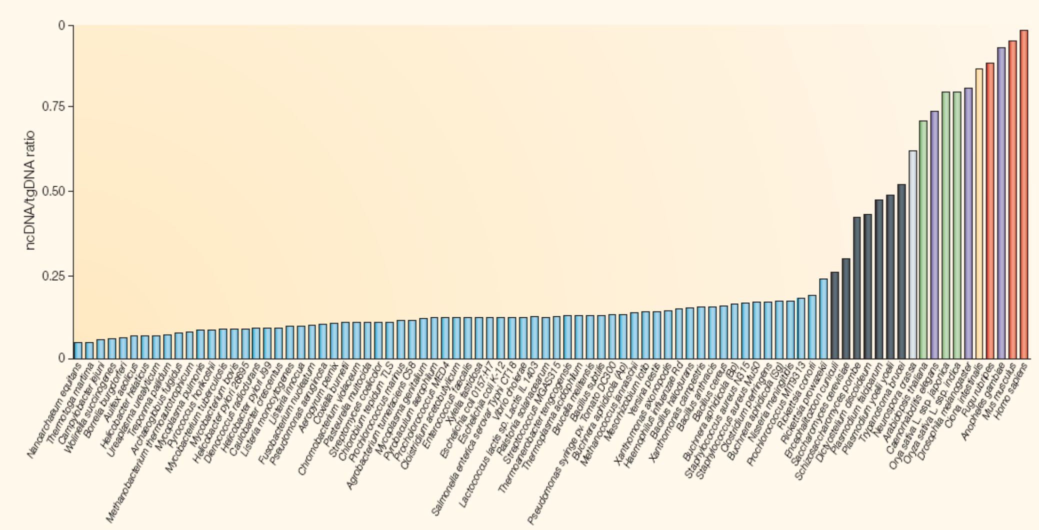


Figure 2 There is a positive correlation between developmental complexity and the ratio of non-coding to protein-coding DNA. Prokaryotic genomes contain less than 25 % non-coding DNA. In Eukaryotes the non-coding DNA occupies a substantial fraction of the genome. [Taken from (Mattick, 2004)]

1.1.2 Non-coding RNAs regulate various cellular processes

As many genomes are being explored, there is a plethora of non-coding RNAs (ncRNAs) being found and widely recognized in various cellular processes. ncRNAs exert very diverse and highly specific functions by: i) hybridizing to their RNA targets, ii) interacting with a small set of proteins or iii) by transcriptional interference. ncRNAs are transcribed either by RNA polymerase II or RNA polymerase III, not only from the introns of protein-coding genes, but also from the exons and introns of non-coding genes (Mattick & Gagen, 2005; Carninci *et al*, 2005), as well as from heterochromatic repetitive regions (Reinhart & Bartel, 2002; Volpe *et al*, 2002).

ncRNAs, like rRNAs and tRNAs engaged in translation of mRNAs, snoRNAs engaged in rRNA modifications, snRNAs implicated in splicing, are already well-described and firmly considered functional. However, the gene regulatory potential of ncRNAs was appreciated once miRNAs and siRNAs were reported. Since their discovery there is lots of attention on the regulatory potential of ncRNAs and this field of research is being extensively explored. Table 1 summarizes many classes of ncRNAs identified thus far.

Class		Mechanism of action	Function
micro, miRNAs		Translational repression, mRNA cleavage	Translational repression
small interfering, siRNAs	endogenous, trans-acting siRNAs	mRNA cleavage	Gene silencing
	PIWI associated, piRNAs	Transposon RNA cleavage	Maintaining of the germline DNA by transposon silencing
	repeat associated, rasiRNAs	Histone, DNA modification	Silencing of retrotransposons, repetitive genes; establishing and maintaining of the heterochromatin structure
	small scan, scnRNAs	Histone methylation, DNA elimination	DNA elimination, genome re-arrangement
antisense, asRNAs		Forming duplex with the coding DNA strand	Transcriptional, translational repression
guide, gRNA		Cleavage of target RNA, insertion of deletion of uridines and relegation of edited RNA	RNA editing in mitochondria of trypanosomes
long non-coding, lncRNAs		Diverse	Transcriptional interference, chromatin remodelling, scaffolding, small RNA precursor, generation of endo-siRNAs, alteration of protein localization

Table 1 Main classes of functional non-coding RNAs. [Adapted from (Zhou *et al*, 2010)]

Importantly, ncRNAs are also key players in setting the proper epigenetic state of the genome. Recently much attention has been given to long ncRNAs (lncRNAs). There are several classes of lncRNAs acting via different mechanisms. In some instances solely the act of transcription of an lncRNA influences significantly the expression of nearby genes either by transcriptional interference or by the chromatin remodeling. For example, the lncRNA transcribed upstream to the human DHFR locus represses the major promoter of the downstream dihydrofolate reductase (Martianov *et al*, 2007). lncRNAs are interacting partners for many proteins and as a consequence they can guide proteins to a specific site or they may influence the activity of their protein partners. HOTAIR lncRNA is one example. It targets the Polycomb Repressive Complex 2 (PRC2) to the distant HOXD locus and thus turns off HOXD transcription and marks this domain for silencing (Rinn *et al*, 2007). Shamovsky and coworkers reported that mammalian heat-shock RNA 1 (HSR1) is crucial for sensing the temperature and the subsequent heat shock response. Upon temperature rising, ubiquitously expressed HSR1 changes its conformation and binds to heat shock factor 1 (HSF1) leading to its trimerization. Only the trimeric HSF1 complex is able to activate transcription of heat shock genes (Shamovsky *et al*, 2006). Genome-wide sequencing studies also imply that lncRNAs serve as precursors for small RNAs. For instance, tRNA-like *mascRNAs* (MALAT1 associated small cytoplasmic RNAs) were reported to be processed from the nuclear long non-coding MALAT1 transcript (Wilusz *et al*, 2008). lncRNAs have been also shown to scaffold subcellular bodies. Presence of e.g. NEAT1, MEN ε/β ensures the integrity of nuclear paraspecles, which are presumably sites for RNA storage (Sunwoo *et al*, 2009; Hutchinson *et al*, 2007). Finally, some lncRNAs are particularly suitable for the epigenetic regulation *in cis*. Their distinct features enable them to efficiently modulate the chromatin structure of the locus of origin, because: i) they reside often at the site of transcription, ii) they are often transcribed at low copy numbers and have fast turnover and iii) they are highly specific, most often being targeted to the unique site.

The first lncRNA discovered in mammalian genome was Xist, the X-inactivate specific transcript (Brown *et al*, 1992). Xist is expressed only from the X chromosome, which will be inactivated, giving rise to a 17-20 kb RNA. It coats the X chromosome *in cis*, targets Polycomb Repressive Complex 2 (PRC2) to the locus through a conserved Repeat A domain and thus triggers the chromosome silencing (Clemson *et al*, 1996). The simplified scheme of action of lncRNAs tethering chromatin-modifying complexes *in cis* is presented in Figure 3.

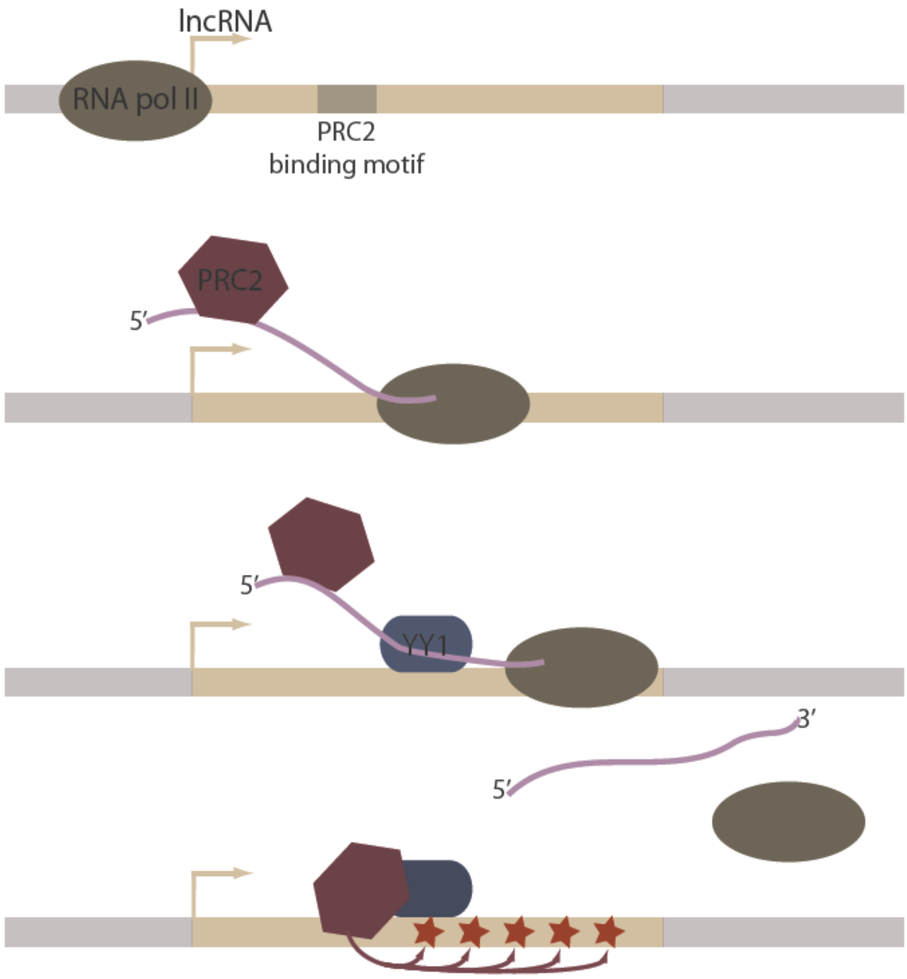


Figure 3 Scheme on an epigenetic regulation triggered by Xist lncRNA *in cis*. RNA polymerase II transcribes lncRNA. The nascent Xist transcript interacts with an epigenetic complex PRC2 and tethers it to the locus. Xist-PRC2 complex is docked onto the chromatin via DNA-binding factor YY1. The chromatin modifications are deposited only *in cis* and the locus is repressed. LncRNA is quickly degraded and RNA polymerase II dissociates. [Adapted from (Lee, 2012)]

Alternatively lncRNA can influence the chromatin state of a locus of origin by serving as a heterochromatin assembly platform. This is the case of lncRNA transcribed from repetitive DNA underlying the centromere of *Schizosaccharomyces pombe*.

1.1.3 lncRNAs serve as platforms for pericentromeric heterochromatin assembly

In *S. pombe* heterochromatin domains cover repetitive sequences and transposons and are located at centromeres, telomeres and mating-type loci. Heterochromatin at these loci is characterized by hypoacetylation of the histones and the presence of an ultraconserved di- or tri-methylation of histone 3 lysine 9 (H3K9). This mark is deposited by methyltransferase Clr4, which in turn is a binding platform for Swi6, Chp1 and Chp2. Swi6 possesses two important domains: chromodomain recognizing the H3K9 mark and a chromoshadow domain interacting with other proteins. (Sadaie *et al*, 2004; Bjerling *et al*, 2002; Nakayama *et al*, 2001; Partridge *et al*, 2000). Heterochromatin assembly requires a so called nucleation site, where repressor proteins are first recruited and it is from this site methylation is further spread on the chromatin fibre in a sequence-independent manner. The heterochromatin status of a given locus is inherited to daughter cell and this "memory" of the chromatin state does not require sequence information of the DNA of this locus. In *S. pombe* the RNAi machinery is crucial for the spreading and epigenetic inheritance of the chromatin state.

Pericentromeric repeats *dg* and *dh* in fission yeast are transcribed by RNA polymerase II into long non-coding RNAs (Figure 4) (Djupedal *et al*, 2005; Kato *et al*, 2005). These lncRNAs are complemented with a second strand by RNA-directed RNA complex (RdRC) and further processed into the centromeric siRNAs by Dicer (Dcr1), which is physically tethered to the RdRC complex (Verdel *et al*, 2009). siRNAs are recognized by the Argonaute protein Ago1, which is a component of RNA-induced transcriptional silencing complex RITS and thus RITS is loaded with centromeric siRNA (Verdel *et al*, 2004). RITS and RdRC complexes are tethered to ncRNAs, as well as to the centromeric DNA, and thus reinforce RNAi and trigger the assembly of heterochromatin at the locus (Motamedi *et al*, 2004). In addition, those complexes are necessary to recruit Clr4 methyltransferase, which is a member of Clr4-Rik1-Cul4 complex (CLRC). Clr4 methylates histone 3 on lysine 9 (Nakayama *et al*, 2001) marking the chromatin repressive state. This conserved mark is then recognized by HP1 proteins, namely Swi6 and Chp2

(Fischer *et al*, 2009; Bannister *et al*, 2001; Thon & Verhein-Hansen, 2000), which it turn can either reinforce the interaction with RITS or the degradation of the ncRNAs via Snf2-histone deacetylase repressor complex (SHREC2) (Sugiyama *et al*, 2007). In addition, heterochromatic transcripts are also degraded by the exosome pathway to further silence the centromeric locus. It is important to note, that all protein complexes required for heterochromatinization, e.g. Swi6, Clr4, RITS, RdRC are tethered to the site of heterochromatin assembly via chromatin-bound centromeric lncRNA.

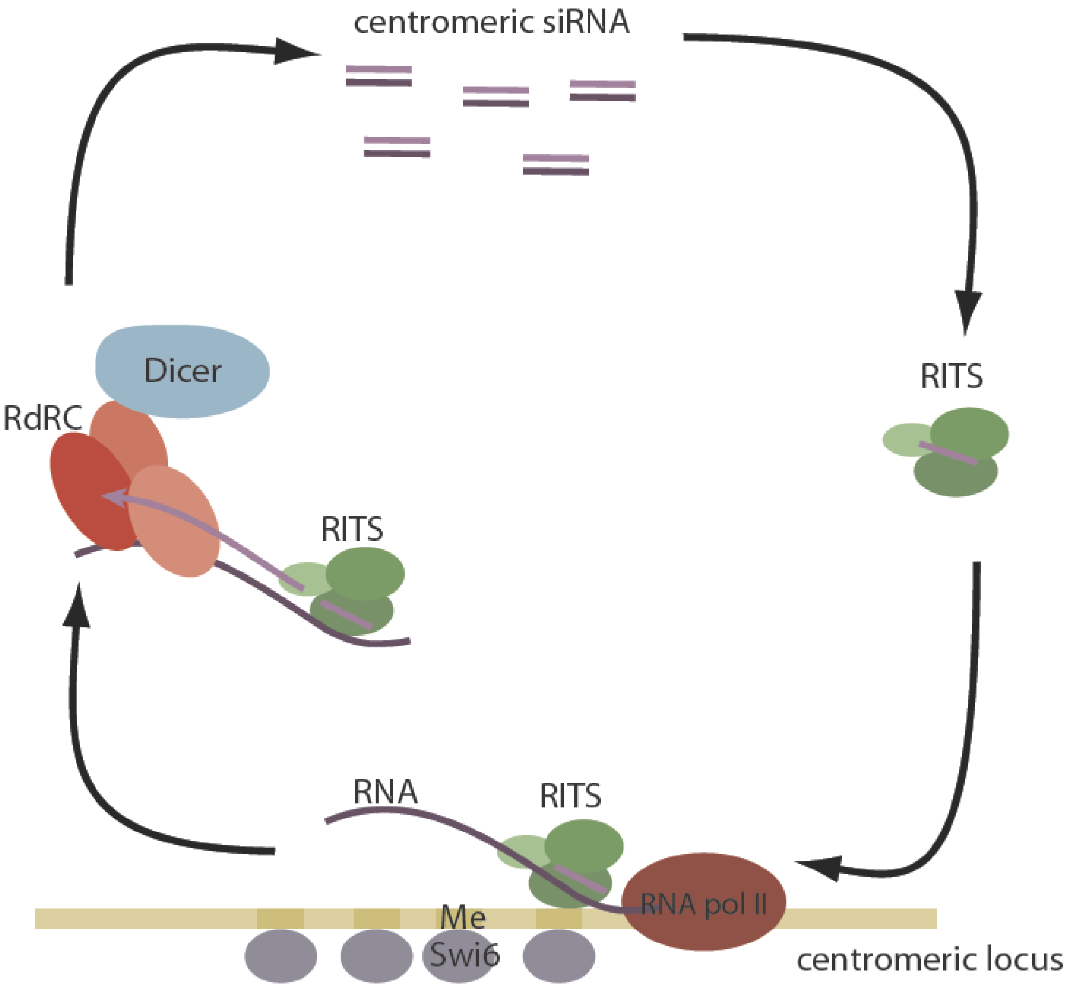


Figure 4 lncRNA transcribed from pericentromeric repeats in *S. pombe* serves as a platform for heterochromatin assembly. Chromatin-bound nascent centromeric transcript recruits RITS and RdRC complexes. Dicer processes double-stranded precursor to centromeric siRNA which are next loaded to RITS. [Adapted from (Moazed, 2009)]

The RNAi pathway contribution in the establishment of pericentromeric heterochromatin has been investigated in other organisms as well, either by exploring the centromere-related phenotype of RNAi mutants (Pal-Bhadra *et al*, 2004) or by detecting pericentromeric siRNAs (Lee *et al*, 2006; May *et al*, 2005; Topp *et al*, 2004). The RNAi-mediated heterochromatin formation at centromeric locus seems to be conserved throughout evolution at least to some extent. However, it should be noted that studying the influence of RNAi on centromeric heterochromatin formation in complex organisms is difficult, mostly due to poor characterization of functional domains within centromeres. In addition, genetic approaches are impeded by high redundancy of numerous homologues of RNAi machinery components.

A link between RNAi pathway and mammalian centromere function was suggested by a few groups (Kanellopoulou *et al*, 2005; Murchison *et al*, 2005; Fukagawa *et al*, 2004). Since a mouse Dicer knockout results in embryonic lethality (Bernstein *et al*, 2003), studies on RNAi-mediated heterochromatin formation were performed with a conditional Dicer-targeting approach either in mouse embryonic stem cells or in a chicken-human hybrid DT40 cell line containing a single copy of human chromosome 21. Kanellopoulou and co-workers showed that the absence of Dicer in murine ES cells results in changes in DNA methylation and histone modifications at the minor satellite repeats and thus transcription activation of normally repressed centromeric loci and transposons (Kanellopoulou *et al*, 2005). In Dicer-deficient cells, centromere-derived transcripts in both orientations were upregulated and double-stranded complexes could not be processed into smaller RNAs of 25-150 nucleotides in length, which were detected in the presence of Dicer. These data stand partially in contradiction to the work reported by Murchison (Murchison *et al*, 2005), which suggests that Dicer, the central component of RNAi machinery, is not essential to preserve heterochromatin at mouse centromeres. Using chicken-human DT40 hybrid cells Fukagawa and co-authors presented that the absence of Dicer activity led to chromosome missegregation caused by premature disjunction of sister chromatids. Moreover, in the absence of Dicer there was a mislocalization of cohesion and checkpoint proteins accompanied by the normal localization of kinetochore proteins. In addition, human centromeric transcripts were abundant enough to be detectable. But once Dicer was present, very little amounts of siRNA-like RNAs of satellite sequence were observed. The model inferred by Fukagawa implies that the regulation of the kinetochore core centromere versus pericentromeric chromatin is different and much more complex than in fission yeast. The authors suggest that in pericentromeric domain repeat-derived nascent RNA is a platform for histone-modifying complexes, analogous to fission yeast model. HP1 proteins interact with H3K9 methylation further assuring proper recruitment of cohesion and checkpoint proteins. The kinetochore central region would be preserved from the histone-modifying enzymes by the presence of CENP-A substituted nucleosomes and thus would not be regulated in RNAi-dependent manner. The last example of vertebrate small centromeric RNAs comes from studies on the tammar wallaby. Carone and colleagues observed that depletion of small centromeric RNAs, termed crasiRNAs, correlated with the delocalization of H3K9 methylation (Carone *et al*, 2009).

Taken together these reports offer at least partial evidence for the small RNA-based mechanism of mammalian heterochromatin maintenance. However, there are some vague parts that need to be further elucidated. For example, the problematic detection of small RNAs derived from peri/centromeric satellites, or the amplification of siRNA signal, since a proper RNA-dependent RNA polymerase still remains uncharacterized in mammals.

1.2 Approximately half of the human genome consists of repeats

The coexistence of unique and repetitive DNA within eukaryotic genomes was revealed by Britten and Kohne (Britten & Kohne, 1968). After 40 years of research, upon the complete sequencing of many genomes, it is commonly known that repetitive DNA constitutes a substantial fraction of eukaryotic genomes. At least 51 % of the human genomic DNA is occupied by repeats and repeat-derived sequences. Repeats may be classified by function, sequence similarity or the pattern of how they appear in the genome. According to the organizational criterion there are tandem and dispersed repeats (Figure 5).

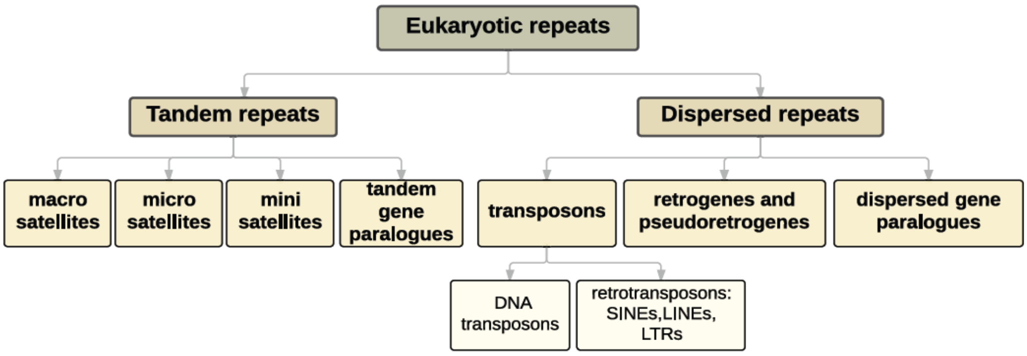


Figure 5 Classification of eukaryotic repetitive DNA.

The most prominent human repeats, constituting at least 45 % of the genome, are transposable elements, that are difficult to recognize due to sequence divergence or partial degradation (Jason de Koning *et al*, 2011). Transposable elements are often referred to as "jumping genes", as their discoverer Barbara McClintock described them. Transposons can move from one genomic location to another either by a cut-and-paste mechanism (DNA transposons) or via RNA intermediates (retrotransposons). Mammalian retrotransposons are: long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs) or long terminal repeats (LTRs) (Figure 6).

Tandem repeats are composed of adjacent copies of a DNA monomer that are organized either in the same orientation or in the opposing direction, in case of inverted tandem repeats. Tandem arrays contain moderately repetitive sequences, like rRNA or telomeric repeats, as well as highly repetitive centromeric satellites.

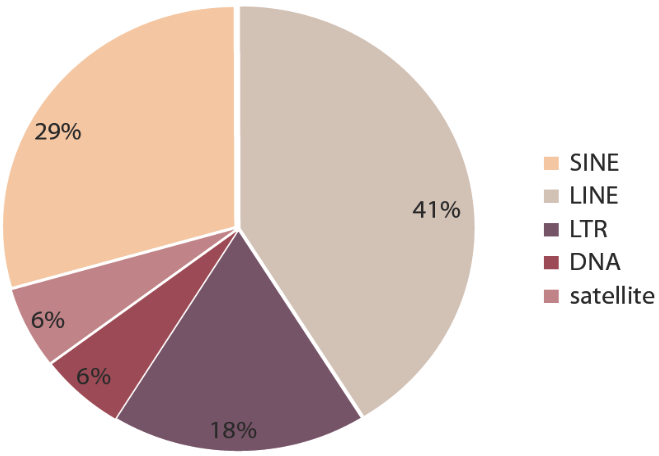


Figure 6 Contribution of different repeat classes into the human genome. [Adapted from Matylla-Kulinska et al. 2014, *accepted*]

1.2.1 Satellite repeats compose heterochromatic domains

The name satellite originates from the second, satellite band formed by these sequences when genomic DNA separates on a cesium chloride density gradient (Waring & Britten, 1966). The partition from the bulk DNA is due to different nucleotide content of satellite DNA. As and Ts are overrepresented when compared to the remaining genomic DNA. According to the total length of arrays, satellites are further categorized into: macro-, mini- and microsatellites (Table 2).

Satellite class	Family	Monomer	Array length
(Macro)satellites	α	171 bp	Kb-Mb
	β, *Sau3 A*	69 bp	
	satellite I	25-48 bp	
	satellite II	attcc	
	satellite III	5 bp	
Minisatellites	telomeric	ttaggg	0,1-20 Kb
	hypervariable	9-64 bp	
Microsatellites		1-4 bp	<150 bp

Table 2 Main characteristics of satellite DNA classes.

Satellites, or macrosatellites, cover about 5 % of the human genome (Lander *et al*, 2001), but it is important to realize that the actual abundance is presumably higher. Typically, satellites are constructed as long arrays of repeated monomers oriented in the head-to-tail manner. They are the dominant element of heterochromatic regions spanning up to several megabases (Charlesworth *et al*, 1994).

Satellites are among the most dynamic elements in the genome. It has been proposed that alterations in the sequence within satellite monomers appear due to the non-reciprocal exchange mechanisms (unequal crossing-over, transposon-mediated reinsertions, gene conversion, and rolling-circle replication followed by re-insertion). Next, these sequence variations are homogenized and fixed within a sexually reproducing population (Dover, 2002; Drouin & de Sá, 1995). This process of concerted evolution (Figure 7) explains well why closely related species differ substantially in the satellite sequence and monomer copy number, while satellite arrays within one genome are more similar (Ugarkovic, 2005).

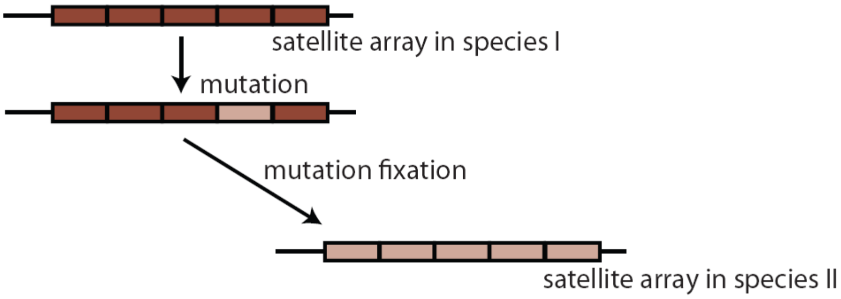


Figure 7 Concerted evolution of satellite repeats. Homogenization of mutations results in a high similarity in the nucleotide sequence among monomers. As a consequence of fixation, satellite arrays within a genome show higher sequence homogeneity than within closely related species.

1.2.2 Functional elements reside in satellite repeats

The vast amount of repetitive sequences within eukaryotic genomes was the basis for a notion that genomes have accumulated and retained repeats as an inert burden (Orgel & Crick, 1980; Doolittle & Sapienza, 1980). However, after Sverdlov observed that LTRs of human endogenous retroviruses may serve as promoters or enhancers for the nearby genes (Sverdlov, 1998), it became apparent that repeats, at least transposable elements, localized in vicinity of host genes can modulate their expression. The positive contribution of repetitive sequences in shaping genomes has been appreciated since. Repeat-derived elements have been described as putative gene enhancers, insulator elements, alternative promoters or alternative splice sites and transcriptional silencers (Schumann *et al*, 2010). Just to illustrate the benefits of repetitive elements for gene expression regulation, the analysis of the human genome sequence disclosed that nearly 25 % of human promoters comprise transposon-derived sequences (van de Lagemaat *et al*, 2003) and about 10 % of transcription binding sites originated from repeats (Polavarapu *et al*, 2008).

Most attention has been given to transposon-derived elements as potential regulators. However, there are a few observations implying that satellite repeats could be of regulatory importance as well. First, by matching the monomer sequences of various human satellite elements, Romanova and colleagues observed that nucleotide mutations have not accumulated uniformly along the sequence (Romanova *et al*, 1996). Satellite sequences contain domains with rare nucleotide alterations as well as more variable parts, presumably implying some functional constrains on the evolution of the sequence. Variable domains could be a consequence of the positive selection for an interaction with a rapidly evolving protein, like the centromere-specific CenH3 histone variant (Cooper & Henikoff, 2004). On the other hand, the constant domain might contribute to binding to a conserved protein, such as centromere protein CENP-B or its homologues. Both CENP-B protein as well as CENP-B box motif within satellite DNA have been identified as evolutionary conserved (Mravinac *et al*, 2005; Kipling & Warburton, 1997).

Sequence analysis of many different satellite species also revealed a characteristic AT nucleotide tract distribution. This AT periodicity influences the bending of the DNA helix leading to the formation of a superhelical tertiary structure, which in turn facilitates tight, heterochromatic conformation of the satellite chromatin (Fitzgerald *et al*, 1994). In addition, satellite DNA from many organisms contain palindromic sequences (Zhu *et al*, 1996; Tal *et al*, 1994). Recently, Bulut-Karslioglu and co-workers reported that Pax homeodomain transcription factors interact with palindromic motifs located in murine major satellites (Bulut-Karslioglu *et al*, 2012). Similarly, human α satellite DNA includes some palindromes forming hairpin structures that attract topoisomerase II to the centromeric locus where it may effect the sister chromatid cohesion (Jonstrup *et al*, 2008).

Although satellite DNA is a major component of heterochromatic domains there is growing evidence that it is not transcriptionally silent (Vourc'h & Biamonti, 2011; Eymery *et al*, 2010; Rizzi *et al*, 2004; Jolly *et al*, 2004). Most satellite-derived ncRNAs are transcribed in a developmental- and tissue-specific manner (Probst *et al*, 2010; Lu & Gilbert, 2007; Li & Kirby, 2003; Rudert *et al*, 1995) implying a potential regulatory role of those transcripts. Furthermore, it was also reported that transcription of satellite DNA is activated by stress conditions (Tittel-Elmer *et al*, 2010; Eymery *et al*, 2010; Valgardsdottir *et al*, 2008; Rizzi *et al*, 2004). Heat-shock specific transcripts derived from human satellite III where shown to re-localize and retain SR family splicing factors in the nuclear stress granules, presumable regulating splicing upon stress condition (Valgardsdottir *et al*, 2005; Metz *et al*, 2004; Chiodi *et al*, 2004). Moreover satellite transcription has been also associated with cancer (Ting *et al*, 2011; Eymery *et al*, 2009a). However in this case, transcriptional activation of normally repressed domains may be a consequence of massive methylation rearrangements of the genome rather than evidence for the satellite functionality. As described above in section 1.1.3, lncRNAs derived from centromeric tandem repeats could be involved in the process of heterochromatin assembly in many organisms suggesting that lowly abundant satellite transcripts play crucial, epigenetic roles in repressing certain chromatin domains.

In the context of centromeric localization of satellite repeats, they also serve as structural component of centromeres. Satellite DNA contains a short CENP-B binding motif present in alternating monomers (Ikeno *et al*, 1994) that mediate the assembly of centromeric chromatin (Masumoto *et al*, 2004). In addition, satellite-derived transcripts are also important for centromere structure. Long single-stranded α satellite transcripts were reported by Wong as crucial components of the kinetochore proteins assembly. Moreover α satellite-derived RNAs were shown to mediate the accumulation of centromere-specific RNAs and proteins at the interphase nucleolus by the time of mitosis when the kinetochore is assembled at centromeres (Wong *et al*, 2007). Similarly, Du and colleagues showed that satellite transcripts in maize interact with centromeric protein CENP-C to increase its affinity to the centromeric DNA (Du *et al*, 2010).

Interestingly, in insects, flatworms and amphibians, satellite DNA is transcribed into the hammerhead ribozyme structure. These transcripts were shown to be catalytically active as they self-cleave into satellite monomers (Rojas *et al*, 2000; Ferbeyre *et al*, 1998; Epstein & Gall, 1987). However, biological relevance of those satellite-encoded ribozymes remains to be explored.

1.3 Centromeres contain repetitive DNA

Centromeres are the primary constriction on every eukaryotic chromosome that together with the telomeres, guard the integrity of the chromosomes and thus the integrity of the genome. By definition, centromeres are chromosomal domains marked by specific CenH3 nucleosomes at the site of formed functional kinetochores. However, in order to ensure precise inheritance of the genome during mitosis, the centromere core domain needs to be surrounded by pericentromeric heterochromatin regions. Centromeres perform a number of functions: i) kinetochore assembly, ii) microtubules attachment, iii) sister chromatid disjunction, iv) pulling of resolved chromatids to the opposite poles during cell division and v) regulation of the beginning of anaphase via checkpoint of the cell cycle progression. On the other hand, the pericentromeric regions are responsible for providing a context for sister chromatid cohesion (Lippman & Martienssen, 2004), defeating the recombination process in the region (Ellermeier *et al*, 2010), and separating the centromere core from the euchromatic context (Chen *et al*, 2008).

There is a range of different centromere types described for diverse species (Table 3). Point centromeres, described in budding yeast, are characterized by the presence of a short, specific DNA that is recognized by centromere-specific protein Cse4 forming a single nucleosome that binds a single microtubule (Furuyama & Biggins, 2007). Most multicellular eukaryotes characterized so far possess so-called regional centromers, which are formed on favored repetitive DNA sequences, but could also occasionally assemble *de novo* elsewhere on the chromosomal arms, the latter are referred to as neocentromeres (Amor & Choo, 2002). All primary human centromeres are built on arrays of 171 bp long α satellite monomers repeated head-to-tail, mostly in a unidirectional orientation. The predominant families of human centromeric DNA were described by Prosser and colleagues (Prosser *et al*, 1986) and the α satellite consensus sequence was constructed based on monomers isolated from non-homologues chromosomes (Prosser *et al*, 1986; Vissel & Choo, 1987). α satellite monomers differ substantially between each other, on average by 20-40 % (Wayel & Willard, 1987), and are hierarchically organized into higher-order repeats (HORs) specific for each chromosome (Willard & Waye, 1987). However, the sequence similarity of the monomers contained within a HOR reaches 99 %, which is explained by the concerted evolution of α satellite DNA (Figure 7) (Ugarkovic, 2008; Durfy & Willard, 1990). The significant centromeric sequence diversity, both among human individual chromosomes, as well as among vertebrates, gave rise to the idea that centromeres are specified by a non-DNA sequence component (Karpen & Allshire, 1997). The third centromere type is described as diffused and is present in worms. There, the entire holocentric chromosome works as a centromere.

Species	Centromere core size	Number of kinetochores	Structure
S. cerevisiae	~125 bp	1	Specific sequence
S. pombe	~4-7 kb	2-3	Unique core flanked by repeats
C. albicans	~3-5 kb	1	Unique sequence
N. crassa	~150-300 kb	ND	AT-rich repeats
D. melanogaster	~500 kb	ND	Simple repeats
C. elegans	Whole chromosome	ND	Diffused
X. laevis	ND	ND	Repeat arrays
G. gallus	~30-500 kb	4-5	Repeats interchanged with unique sequence
O. sativa	~0.75-2 Mb	ND	Repeats arrays interchanged with active genes
H. sapiens	~0.5-10 Mb	15-20	α satellite arrays

Table 3 Diversity of eukaryotic centromeres. ND stands for "not determined". [Adapted from (Burrack & Berman, 2012)]

Upon changes in the centromere organization, centromeric DNA has co-evolved allowing adaptability to the structural variations. Centromeric DNA belongs to the most rapidly evolving genomic domains, varying substantially in the nucleotide sequence and length (Plohl *et al*, 2008). Although there is an apparent lack of the phylogenetic conservation of the centromeric sequence and size, there are some general elements that are preserved: i) the kinetochore-related centromere core is flanked by heterochromatin regions, ii) centromere-associated proteins are similar (Puechberty *et al*, 1999) and iii) most centromeres are contained within (A+T)-rich repetitive sequences, as shown by chromatin immunoprecipitation (ChIP) data (Zhong *et al*, 2002; Vafa & Sullivan, 1997). However, it is important to note that repeats are solely the favorable sequence for the centromere assembly (Ohzeki *et al*, 2002). Ectopically localized centromeres, neocentromeres, were documented to form at loci lacking any α satellites or even other satellite repeats (Choo, 2001; Koch, 2000). Moreover, the sequence of the centromere core in comparison with the surrounding pericentromeric domains seems to be highly similar (Puechberty *et al*, 1999; Schueler *et al*, 2001).

Comparison between centromeric satellite DNA among vertebrates has not revealed any common sequence motif, except for a short 17 bp binding site for the CENP-B protein (Mravinac *et al*, 2005; Kipling & Warburton, 1997; Masumoto *et al*, 1989) or binding site for CP1 in budding yeast (Baker *et al*, 1989). But strikingly, in most organisms the length of a monomer ranges between 140-180 bp resembling the lengths of a nucleosomal unit. This length correlation may be a sign of the evolutionary constrain towards satellite DNA to perform its structural function (Shelby *et al*, 1997; Henikoff *et al*, 2001).

1.3.1 Centromeric DNA alone is not competent to specify the functional centromere

Several lines of observations suggest that centromeres are epigenetically determined. First, centromeric DNA differs substantially among species. Second, functional centromere domains are contained within much longer satellite arrays. Human centromeres consist of several Mb of α satellite arrays, however only a subset of 30-50 % of α satellite repeats form the actual core structure leaving the rest in the pericentromeric heterochromatic context (Lam *et al*, 2006). Third, functional centromeres can rarely assemble at ectopic loci lacking any sequence similarity to satellite DNA (Koch, 2000). Moreover, studies on dicentric chromosomes showed that in spite of the presence of two loci competent to form centromeres, only one centromere remains active and the other one is suspended (Sullivan & Schwartz, 1995). Finally, Earnshaw discovered a few proteins to be exclusively present at centromeres (Earnshaw & Rothfield, 1985) and named them centromere proteins CENP-A, B and C (Earnshaw *et al*, 1986). Later, many proteins associated with the centromere/kinetorochore structures were reported, such as: CENP-E, F, H, I (hMis6), hMis 12 and shown to be highly conserved (Foltz *et al*, 2006; Izuta *et al*, 2006; Okada *et al*, 2006; Chan *et al*, 2005). The multiprotein kinetochore complex forms on the centromere (Figure 8) and since centromeric DNA is highly variable the core kinetochore proteins must tolerate significant sequence diversity with most binding in the sequence-independent manner (Glynn *et al*, 2010).

The functional centromeres, regardless of the underlying DNA sequence (Lo *et al*, 2001), have been shown to associate with nucleosomes at sites where the canonical histone H3 was substituted by the centromeric variant CenH3: CENP-A in vertebrates, Cid in *D. melanogaster*, HCP-3 in *C. elegans*, Cnp1/SpCENP-A in *S. pombe* and Cse4 in *S. cerevisiae*. CENP-A has been co-purified with the nucleosome core being its integral component (Palmer *et al*, 1991). CENP-A protein shares 62 % sequence similarity with histone H3 and was shown to replace H3 (Shelby *et al*, 1997). The structure of CENP-A chromatin has been recently extensively studied and reinforces the hypothesis that the centromeres identity is conferred on the epigenetic uniqueness of the CENP-A nucleosomes. Recently, Hasson and co-workers demonstrated that the major form of CENP-A nucleosome is the octamere with loose termini (Hasson *et al*, 2013) and compared to canonical H3 nucleosome, has a reduced height implying the physical distinction between centromeric and canonical nucleosomes (Miell *et al*, 2013).

The presence of CENP-A marks exclusively active centromeres since the protein is not detectable on mutated or inactivated centromeres (Tyler-Smith *et al*, 1999; Sullivan & Willard, 1998; Sullivan & Schwartz, 1995). Chromosome missegregation and failures in kinetochore formation are phenotypes observed from mutations or deletions of CENP-A (Howman *et al*, 2000; Regnier *et al*, 2005). Therefore, CENP-A protein is considered to be a founder component required for further steps in the centromere/kinetochore specification. In addition CENP-A is regarded an upstream factor for the proper localization of other centromeric proteins, such as Mis12, CENP-C, CENP-H and CENP-I in a co-dependent manner (Amor *et al*, 2004). However, when human CENP-A was mistargeted to an ectopic location, only a portion of the kinetochore proteins were recruited and centromere activity was not observed (Van Hooser *et al*, 2001). Recent tethering approaches and chromosome engineering studies have characterized more factors contributing to the functional centromere assembly. Okada and colleagues demonstrated that *de novo* formation of an artificial kinetochore (human artificial chromosome) depends on the presence of α satellites providing CENP-B binding site (Okada *et al*, 2007). Moreover, targeting of chaperone protein HJURP, primarily described as a Holiday Junction Recognition Protein, to the Lac operon was shown to promote loading of CENP-A to that locus. *De novo* incorporated CENP-A recruited another 16 proteins, known as constitutively centromere-associated network (CCAN) (Foltz *et al*, 2006) and subsequently formed the kinetochore-microtubule attachment (Barnhart *et al*, 2011). On the other hand Gascoigne and co-workers demonstrated that by the ectopic recruitment of CENP-C and CENP-T the kinetochore formation occurs with CENP-A being omitted (Gascoigne *et al*, 2011) implying that CENP-A recruits CENP-C and T.

Recent data from a gene targeting approach data yielded insight into the central role of CENP-A. Fachinetti and colleagues, tracking H3-CENP-A at endogenous centromeres, demonstrated that CENP-A comprises three distinct regions required for its function (Figure 8) (Fachinetti *et al*, 2013). CENP-A targeting domain (CATD) is a histone-specific region that is sufficient to determine the centromere identity. When CATD is fused to histone H3, it not only targets the protein to the centromere but also rescues the proper centromere function upon CENP-A depletion (Black *et al*, 2007). Moreover, CATD domain is also required for the localization of CENP-A in a cell-cycle dependent manner triggered by a chaperone protein HJURP (primarily identified as a Holiday Junction Recognition Protein) (Jansen *et al*, 2007; Dunleavy *et al*, 2009). Furthermore, the CENP-N protein selectively binds CATD domain in a DNA-sequence independent fashion. Thus CENP-N discriminates between CENP-A nucleosomes and H3 nucleosomes and recruits other centromeric proteins to the centromeric chromatin (Carroll *et* *al*, 2009). The second domain, C-terminus of CENP-A (CAC domain), is required for the interaction with CENP-C. Histone H3 fused with CAT and CAC domains were shown to provide centromere survival by establishing the proper interactions with centromere and kinetochore proteins and being correctly loaded into the chromatin via HJURP (Fachinetti *et al*, 2013).

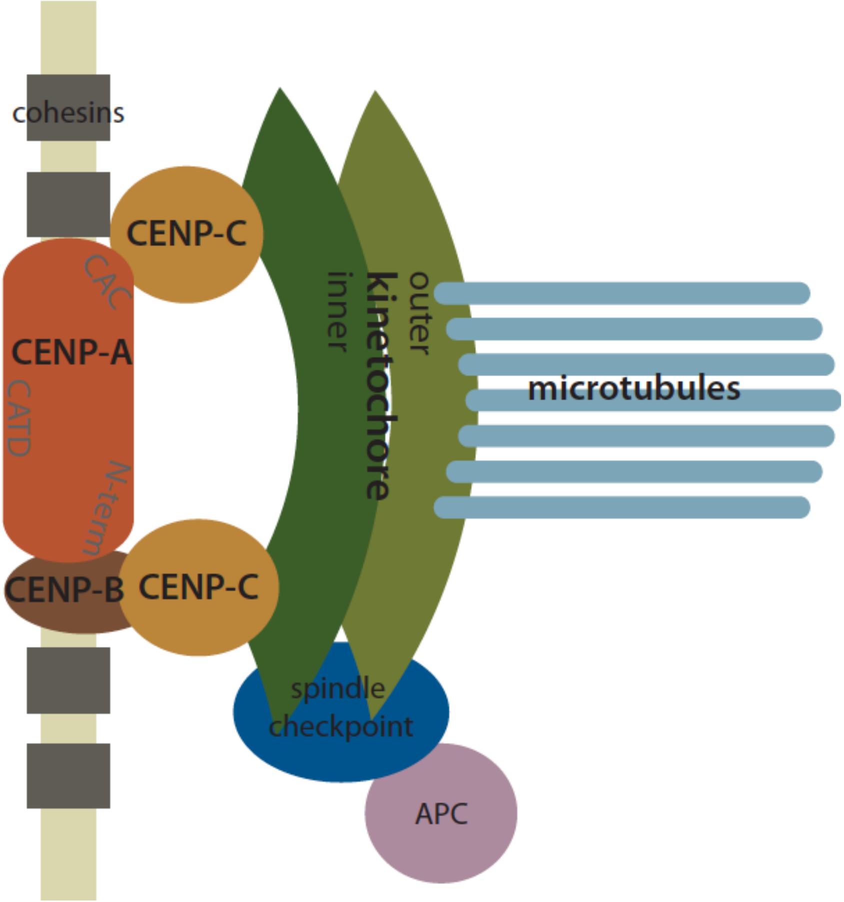


Figure 8 Scheme on the organization of centromere-kinetochore complex. CENP-A is recruited to the centromere via CENP-B bound to the satellite DNA and mediated through HJURP associated with CATD domain. CENP-C, which links with the kinetochore complex, is recruited either by CENP-A via CAC domain or by CENP-B. Additionally, centromeres assure sister chromatid cohesion via cohesions. Moreover, a subset of kinetochore components controls the mitotic progression via anaphase-promoting complex (APC). [Adapted from (Verdaasdonk & Bloom, 2012)]

Importantly, Fachinetti provides evidence for two parallel mechanisms to recruit kinetochore formation. CANP-C can interact with CAC domain and in this case be recruited via CENP-A, as described above. But alternatively, CENP-C binds CENP-B that in turn interacts with the CENP-A N-terminus domain. CENP-B is a DNA-binding protein targeted to the centromere through a CENP-box motif contained within α satellites (Muro *et al*, 1992). The co-existence of two kinetochore recruitment pathways may explain some previous observations, such as α satellites arrays on chromosome Y are devoid of CENP-B box (Haaf *et al*, 1995) and thus do not bind CENP-B protein, but still recruit all other CENP proteins (Earnshaw *et al*, 1989). Moreover, neocentromeres assembled on non-satellite DNA are deficient in CENP-B but nevertheless are able to recruit functional kinetochores (Saffery *et al*, 2001; Choo, 2001). In contrast, Okada and co-workers reported that CENP-B attracted to the α satellite motif is necessary not only for *de novo* centromere assembly but also for the formation of the heterochromatin (Ohzeki *et al*, 2002; Okada *et al*, 2007). However, it is not required for maintenance of functional kinetochore (Hudson *et al*, 1998).

Taken together, the centromere-kinetochore assembly is crucial for the fidelity of chromosome segregation and thus for the integrity of the genome. Proper centromere-kinetochore assembly requires concerted action of many factors allowing precise regulation and many checkpoints steps. Moreover, the centromere/kinetochore assembly pathway needs to be adaptive in order to buffer potential variations in e.g. nucleotide composition, complex position, kinetochore protein number (Tomonaga *et al*, 2003). There is still much to be investigated and understood in centromere assembly and maintenance, but it seems that the epigenetic regulation of this process assures its accuracy and robustness.

1.3.2 Centromeres are not transcriptionally inert

Although centromeric and pericentromeric regions both contribute to the process of the accurate chromosome segregation (Ekwall *et al*, 1997; Blower & Karpen, 2001; Bernard *et al*, 2001), they do not share the same chromatic features. It is evolutionary conserved that the centromere core chromatin is substantially different from the flanking repressive pericentromeric regions (associated with methylation of lysine 9 histone 3 and lysine 27 histone 3) and demarcated from it by the dimethylation of the lysine 9 on the histone 3 (Lam *et al*, 2006; Saffery *et al*, 2003). Lam proposed that the centromere core constitutes a distinct chromatin form, distinguishable from euchromatin as well as from heterochromatin (Lam *et al*, 2006). The major hallmark of the centromere chromatin is CENP-A-containing nucleosomes intermingled with the histone 3 modifications: methylation of the lysine 4 (H3K4me1, H3K4me2) and lysine 36 (H3K36me2, H3K36me3) (Gopalakrishnan *et al*, 2009; Bergmann *et al*, 2011, 2012). Centromeric DNA was also reported to be highly methylated, which is usually linked to transcriptional silencing. Nevertheless, Wong and colleagues proposed that the centromeric transcription could occur in the pockets of non-methylated DNA sequence (Wong *et al*, 2006). Identification of the chromatin remodelling complex, referred to as FACT (facilitates chromatin transcription), may be taken as more evidence for the transcriptional activity at centromeres (Okada *et al*, 2009). Currently there is mounting evidence that transcription within centromeres is not only permissible but also promoted; similar to the monoubiquitination of the histone H2B, which associates with α satellites within CENP-A domains to regulate the process of centromeric transcription in the cell-cycle regulated manner (Sadeghi *et al*, 2014).

Pericentromeric transcription in *S. pombe* has been described in detail in the context of its contribution to the heterochromatin formation (Volpe *et al*, 2002; Volpe & Martienssen, 2011). In vertebrates it is unclear whether pericentromeric transcription is taking part in heterochromatin formation (Kanellopoulou *et al*, 2005; Murchison *et al*, 2005; Fukagawa *et al*, 2004; Carone *et al*, 2009). Instead, transcription activity within pericentromeric regions was detected mainly during development, under cellular stress conditions or in connection to cancer. Murine major satellites were shown to be transcribed in the developing mouse embryo (Rudert *et al*, 1995) or in the aging heart (Gaubatz & Cutler, 1990). As opposed to mice, where major and minor satellites constitute both pericentromeric and centromere cores, respectively, in humans there is no obvious boundary between these regions as both pericentromeric and centromeric domains are burdened with α satellites. As a consequence it is problematic to determine the origin of the detected human satellite transcripts. Human pericentromeric arrays are rarely interspersed with other repeats, like satellite III, SINEs or LINEs (Prades *et al*, 1996; Tagarro *et al*, 1994) and transcription from these elements were reported to be induced under cellular stress (Rizzi *et al*, 2004; Jolly *et al*, 2004; Valgardsdottir *et al*, 2008). In addition, global derepression of heterochromatic regions caused by cancerogenesis results in the activation of α satellites arrays, as detected in breast, epithelial and pancreatic cancer (Zhu *et al*, 2011; Ting *et al*, 2011; Eymery *et al*, 2009b).

The transcription from the centromere core is much more difficult to demonstrate and therefore much less studied and understood. In *S. cerevisiae* depletion of *Cbf1*, a transcription factor interacting with the centromere core, led not only to the repression of centromeric transcription, but also to chromosome missegregation and consequently chromosome loss. Interestingly, this severe phenotype was rescued with centromeric transcription from an artificial promoter (Ohkuni & Kitagawa, 2011), suggesting that transcription is crucial to maintain the centromere function. Reports from *S. pombe* imply that there is a correlation between CENP-A incorporation and the transcriptional repression as an ade 6 maker gene placed within a centromere core was repressed. Furthermore, Castillo and co-workers demonstrated that CENP-A chromatin formed on those repressed marker genes depended on the relative abundance of histone H3 within the context of the sequence (Allshire *et al*, 1994; Castillo *et al*, 2007). Only recently, transcripts from core CENP-A chromatin were detected in fission yeast (Choi *et al*, 2011). In mammals, some initial insight was gained on the transcriptional activity of the centromere core with studies on neocentromeres. Genes embedded in the active centromere domains were reported to be transcribed, for example the L1 retrotransposon from mardel(10) neocentromere, which is transcribed into FL-L1 RNA and incorporated into neocentromere chromatin. FL-L1 knock-down led to a decrease in the CENP-A levels at the mardel (10) neocentromere suggesting that the transcript influenced the structure of the centromere (Chueh *et al*, 2009). Furthermore, active RNA polymerase II associated with transcription factors was shown to localize to human kinetochore during mitosis. Moreover inhibition of the polymerase resulted in the reduction of α satellite-derived RNAs and CENP-A levels. Finally, centromeric RNAs per se were detected in mouse (Bouzinba-Segard *et al*, 2006; Ferri *et al*, 2009), tammar wallaby (Carone *et al*, 2009) and human (Wong *et al*, 2007; Horard *et al*, 2009), as well as in plants (Topp *et al*, 2004; Lee *et al*, 2006).

1.3.3 Repeat-derived transcripts are integral elements of the centromeric chromatin and the kinetochore

The seminal work of Riedel demonstrated that RNA was an integral component of the kinetochore (Rieder, 1979). Later, since Maison and colleagues put forward the idea that RNA is also an essential constituent of the higher-order heterochromatin structure at mouse pericentromeres (Maison *et al*, 2002), heterochromatic loci have been extensively explored considering their potential RNA elements. Currently, there is growing body of data demonstrating that transcripts derived from centromeres have a vital function in the centromere and kinetochore structure.

CentC satellite- and centromeric retrotransposons-derived transcripts ranging in size from 40-200 nucleotides were reported to co-immunoprecipitate together with the CenH3 protein of the maize kinetochore (Topp *et al*, 2004). Also in rice CentO transcripts were found to be associated with CenH3 nucleosomes. Moreover siRNAs cognate for CentO RNAs were detected implying the RNAi-mediated heterochromatin formation (Lee *et al*, 2006).

In mammals, transcripts derived from centromere cores were described to be associated not only with CENP-A, but also with other centromere/kinetochore-specific proteins. In the tammar wallaby, transcripts produced from sat23 and marsupial-specific KERV1 retrotransposon, located at the centromere core, interact with CENP-B and short (< 42 nucleotides) centromere repeat-associated small-interacting RNAs, termed crasi-RNAs. crasiRNAs, which are processed from the longer transcripts are required for the localization of centromere proteins (Carone *et al*, 2009). Murine centromeric minor satellites are transcribed either into large RNAs (2-4 kb) or into 120 nucleotides transcripts in the differentiated or stressed cells. Minor satellites-derived RNAs were demonstrated not only to be contained within the CENP-A chromatin fraction but also to interact with Aurora B/Survivin/INCENP complex during G2/M phase regulating Aurora B activity. Furthermore, Ferri and co-workers inferred that CENP-A within centromere core chromatin provides a scaffold for centromeric RNA-dependent assembly of passenger protein complexes at the onset of mitosis

(Bouzinba-Segard *et al*, 2006; Ferri *et al*, 2009). In humans, α satellite-derived transcripts were detected in the interphase nucleolus together with CENP-C and INCENP proteins to later target these proteins to the mitotic kinetochore (Wong *et al*, 2007). Furthermore, a decrease in α satellite-derived transcripts during mitosis, by blocking RNA polymerase II, led to destabilization of CENP-C binding at centromeres and consequently to the chromosome instability (Chan *et al*, 2012).

1.3.4 The level of centromeric transcription needs to be correct

Interestingly, there are many lines of evidence for a close correlation between the level transcriptional activity and centromere function (Hall *et al*, 2012). In budding yeast transcription from the centromeric locus is driven by the transcription factor Cbf1. Decreased transcription activity triggered by Cbf1 deletion manifested in a severe chromosome instability phenotype, which was overcome by induction of the centromeric transcription from a substituted *MET25* promoter. On the other hand, overexpression of centromeric transcripts also resulted in chromosome missegregation, implying that the maintenance of the exact level of centromeric transcription is compatible with the fidelity of chromosome segregation (Ohkuni & Kitagawa, 2011). Similarly, Chan and colleagues demonstrated that inhibition of RNA polymerase II leads to increased mitotic index caused by the drop in CENP-C deposition (Chan *et al*, 2012).

Indicative seems to be a relation between ectopic *de novo* centromere formation and the transcriptional activity of this locus. Ishii and coworkers examined DNA sequences within fission yeast genome prone to the neocentromere formation and observed that *de novo* centromeres assemble preferentially within lowly-transcribed genes (Ishii *et al*, 2008).

Experiments on human artificial chromosomes, named HACs, also support the requirement for tightly regulated transcriptional activity for the fidelity of the centromere function. Engineering of the epigenetic nature of the centromeric chromatin gave much insight into the transcriptional demands of the centromere loci. Directing either a transcriptional activator or repressor to the HAC alphoid sequence resulted in the HAC missegregation and subsequent loss (Nakano *et al*, 2008). Furthermore, depletion of the specific epigenetic mark solely from the chromatin in the centromere core, di-methylation of the lysine 4 on histone 3 by the centromere tethering the lysine-specific demethylase 1 (LSD1) resulted in the suppression of transcription, which in turn was effective of decrease in CENP-A loading and CENP-C localization at the kinetochore (Bergmann *et al*, 2011).

The detrimental consequences of perturbing the balance in the centromeric transcriptional activity suggest that the proper state of the chromatin within centromeric core undoubtedly contributes to the maintenance of the critical transcriptional level. However, very little is understood about the regulation of RNA polymerase II engaged in the centromeric transcription. So far there are only a few reports on this matter. Thorsen and colleagues demonstrated that Mediator complex may adjust recruitment of RNA polymerase II at the proper level to centromeric repeat in fission yeast (Thorsen *et al*, 2012).

1.4 Transcription of most of the genome is mediated by RNA polymerase II

RNA polymerase II is responsible for the transcription of mRNAs and some of the non-coding RNA species. Thanks to a number of crystal structures the structure and mode of the function of this enzyme is fairly well inferred (Cramer *et al*, 2001; Murakami *et al*, 2013). RNA polymerase II forms a large complex of around 500 kDa comprising 12 subunits Rbp1-12 (Figure 9). The central cleft where DNA enters and the RNA synthesis happens is marked by metal ions (Mg^{2+}) and is formed by Rbp1 and Rbp2 core subunits. The largest DNA-directed subunit Rbp1 contains a carboxyl-terminal domain, referred to as CTD, consisting of 52 heptapeptide tandem repeats. The CTD plays an essential role in the process of transcription and is crucial for survival, since deletion of this domain in rodents leads to neonatal death (Litingtung *et al*, 1999). The CTD domain is a target for a specific phosphorylation code that varies throughout the transcription phases and regulates the initiation of transcription and processive elongation of the nascent RNA. Moreover the CTD is a platform for interactions with many protein partners coupling transcription with the splicing and RNA maturation processes (Egloff & Murphy, 2008).

Transcription initiates once regulatory factors bind in proximity to the transcription start site (TSS). These factors recruit proteins from the transcription complex to a promoter sequence, but can also attract chromatin modifiers to facilitate the process of transcription. The preinitiation complex forms around the promoter core ensuring a proper site for RNA polymerase II. When all transcription factors associate with RNA polymerase II at the TSS, the DNA strands melt to form a 11-15 bp bubble within the, so-called, open complex. The template strand enters the active site cleft and RNA synthesis begins. Often, the transcription machinery produces many short oligoribonucleotides in the process referred to as abortive transcription (Holstege *et al*, 1997). RNA polymerase II is processive once it loses the contact with promoter region and dissociates from most of the transcription factors, which happens when the nascent RNA reaches about 30 nucleotides in length. The elongating polymerase interacts with multiple factors along the process for termination of transcription and maturation of the RNA.

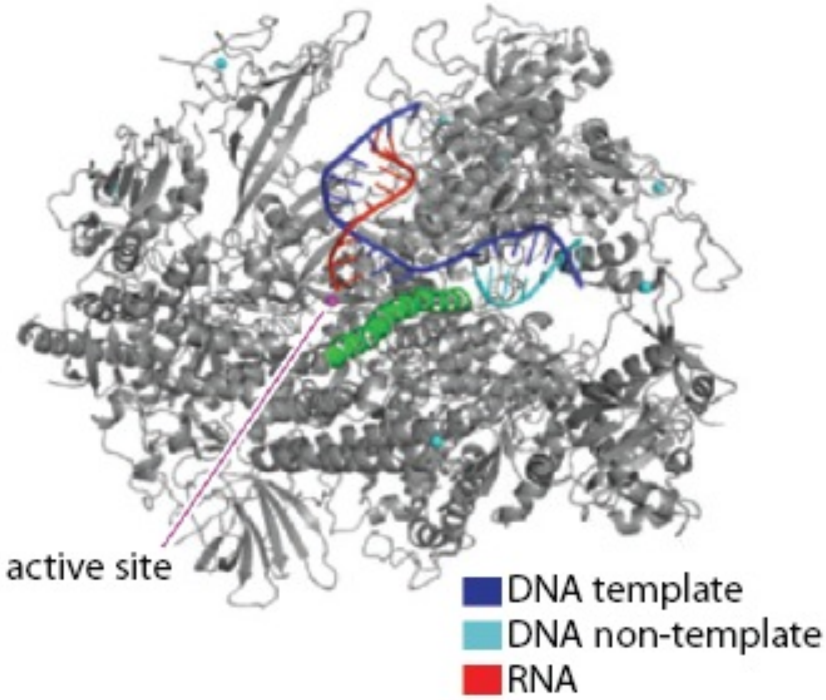


Figure 9 Structure of the RNA polymerase II elongation complex with the adjustable active site. [Taken from (Cramer *et al*, 2008)]

The DNA-dependent RNA polymerase activity is a main action performed by RNA polymerase II. However, there is some structural evidence suggesting that the RNA polymerase II active site is adjustable and may accommodate not only single DNA template strand. Lehmann and colleagues demonstrated that RNA molecule can enter the site where the DNA-RNA complex resides at the time of the canonical transcription (Lehmann *et al*, 2007). This observation suggests that RNA polymerase II may act as RNA-dependent enzyme supporting the hypothesis that RNA polymerase II is a descendent of an ancient replicase.

1.4.1 RNA polymerase II performs RNA-dependent RNA polymerization

In mammals the canonical RNA-dependent RNA polymerase has not yet been characterized. Nevertheless, some functional homologues have been described (Maida *et al*, 2010; Wagner *et al*, 2013). Interestingly, RNA polymerase II has been reported to harbor a RNA-dependent RNA polymerase (RdRP) activity.

Dezelee and co-workers showed that the yeast RNA polymerase II uses the synthetic, single-stranded, homopolymeric $(rC)_n$ RNA as a template for the GTP incorporation (Dezélée & Sentenac, 1974). Yeast polymerase was reported to elongate an RNA template-product “scaffolds” containing HDV antigenome and synthetic FC aptamer, yet with a lower efficiency than the DNA-dependent transcription (Lehmann *et al*, 2007). Moreover, replication of hepatitis delta virus (HDV), which has an RNA genome, was shown to be mediated by the host RNA polymerase II. (-) RNA strand derived from the HDV genome was taken as a template for the *in vitro* specific RNA synthesis in an α amanitin sensitive manner in HeLa nuclear extract (Filipovska & Konarska, 2000). α amanitin is a cyclic peptide, which specifically interferes with RNA polymerase II transcription by blocking the RNA translocation (Bushnell *et al*, 2002). Further, α amanitin-sensitivity of the replication of the HDV genome within the host cells also suggest that RNA polymerase II can act as RNA-dependent RNA polymerase (Lai, 2005). In addition, RNA polymerase II extracted from plants was shown to take the linear viroid (-) RNA strand as a template for the full length (+) RNA strand synthesis (Rackwitz *et al*, 1981).

Intriguingly, RNA polymerase II is also capable of extending RNA from its 3' end. This activity was first observed by Johnson and Chamberlin (Johnson & Chamberlin, 1994). Furthermore, a study on the yeast RNA polymerase II demonstrated that the 3' end elongation of the RNA bound in the active site of the polymerase occurs in the templated manner (Lehmann *et al*, 2007). This extension activity may represent a more general mechanism, since bacterial RNA polymerase was also shown to extend RNAs trapped within the active site of the enzyme (Windbichler *et al*, 2008).

Recently, Wagner and co-workers reported that human RNA polymerase II interacts with the murine B2 non-coding RNA and uses it as the RdRP template. Moreover, the polymerase elongates the B2 RNA by about 18 nucleotides in the templated manner, which results in the destabilization of the complex. The authors proposed, by observing some analogy with the 6S RNA (Wassarman & Saecker, 2006), that the RNA-dependent RNA polymerase activity of RNA polymerase II is a mechanism to rid the enzyme of RNAs trapped in the active site (Wagner *et al*, 2013).

1.5 Aim of the project

The aim of this thesis was the characterization of α satellite-derived transcripts and exploration of their function in human HeLa cells.

A major outcome and surprise from the Human Genome Project was the very low contribution of protein coding DNA in the genomic content (Lander *et al*, 2001). Additionally, it is reported that at least half of the human genome is covered with repetitive elements (Jason de Koning *et al*, 2011). On the other hand, ENCODE reports that 50-75 % of the human genome is transcribed (Djebali *et al*, 2012). From these findings it is conceivable that repeat elements might be transcribed and display novel functions.

Due to technical obstacles, highly repetitive regions, in contrast to protein- and RNA-coding genes, remain still largely unexplored. Many experiments and analyses mask or neglect repeats, thus the knowledge about repeat-derived transcripts and their functional relevance is fairly poor.

As the central interest of my PhD dissertation, I aimed to characterize transcripts derived from human α satellites (αsatRNAs) and address their functionality. α satellite DNA is known to play an important role in binding CENP-B proteins (Sullivan & Glass, 1991) and thereby in the centromere structure, but little is known about a potential function of α satellites at the RNA level.

My thesis consists of two parts: i) a descriptive characterization of αsatRNAs and ii) the biochemical analyses of αsatRNAs activities.

The major motivation for studying αsatRNAs was based on a previous finding in our laboratory. Transcripts derived from α satellites were found to directly bind RNA polymerase II. Genomic SELEX against RNA polymerase II was performed to isolate putative transcription regulators that directly interact with RNA polymerase II. Genomic SELEX, combined with deep sequencing, explores the potential transcriptome irrespective of its expression levels and is especially helpful when searching for RNAs that are encoded in silenced or repressed parts of the genome. In the genomic SELEX assay we isolated 314 RNA polymerase II aptamers derived from α satellite arrays. Additionally, considering that i) transcripts from centromeric regions in other species serve as templates for RNA-dependent RNA synthesis, ii) RNA polymerase II exerts RNA-dependent RNA polymerase (RdRP) activity, and iii) αsatRNAs directly bind RNA polymerase II, I hypothesized that αsatRNAs can serve as substrates for RNA polymerase II.

2 Results

2.1 Characterization of transcripts derived from human α satellite arrays

2.1.1 α satellites in the human genome

Human α satellite DNA exists either as a single 171 bp unit located in pericentromeric regions or as centromeric higher order repeats (HOR) consisting of head-to-tail organized monomers (Schueler *et al*, 2005). For the purpose of this project, α satellite sequences were reannotated using the dfamscan.pl script (http://dfam.janelia.org/help/tools) (Wheeler *et al*, 2013). There are 44058 annotated genomic loci matching α satellite sequences that are clustered into 1301 arrays covering around 0.1 % of the human genome (7.44 Mb). The average array contains 33 α satellite monomers. Each monomer spans 171 bp and differs from any other one by 20-40 % (Wayel & Willard, 1987). α satellites localize mostly to centromeres on every chromosome, but can be also found elsewhere on chromosomal arms (Figure 10). It is worth noting, that α satellite regions belong to poorly annotated parts of the genome mainly due to their low complexity and repetitiveness.

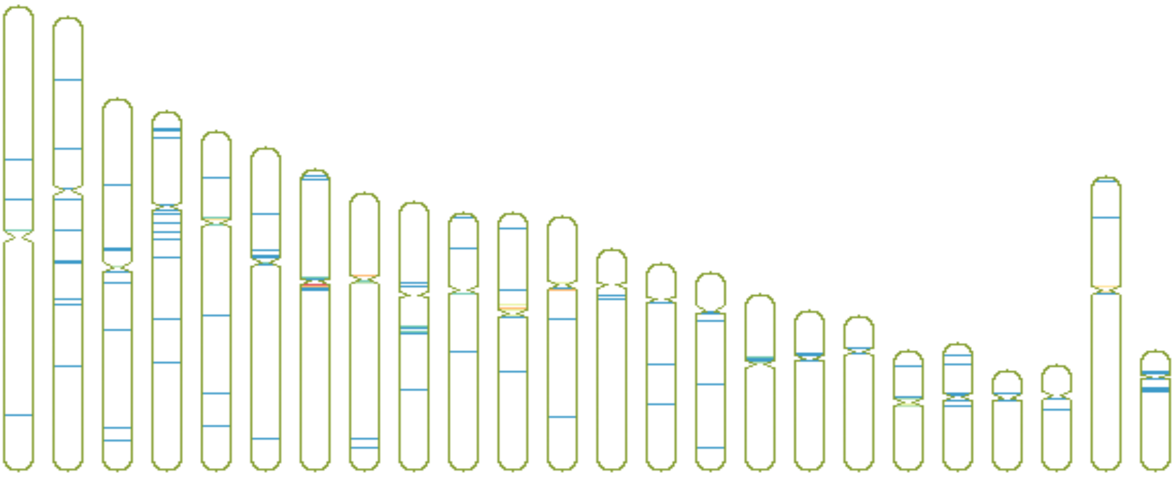


Figure 10 Map of all α satellite hits on human chromosomes. There are 44,058 genomic loci that match with α satellite sequence; similar to the 43,482 found by DFAM hits. [Taken from http://dfam.janelia.org/]

2.1.2 Detection and mapping of transcripts derived from human α satellites

Repetitive arrays are mainly enclosed within constitutive heterochromatin domains and therefore to date have been mistakenly considered to be transcriptionally inert. To test whether human α satellite arrays are transcribed we searched for α satellite transcripts (αsatRNAs) in HeLa cells via strand specific RT-PCR. Considering that: i) α satellite DNA is AT-rich, which results in a relatively low melting temperature of primers, making them more prone for mispriming; ii) α satellite arrays in the genome are poorly annotated, iii) sequence homology between any α satellite monomer is approximately 60 % (Wayel & Willard, 1987), I designed primers to the consensus α satellite sequence retrieved from (Prosser *et al*, 1986). This strategy allows αsatRNAs' detection *en masse*. Additionally, we also used some α satellite-specific primers for RT-PCR reactions. To confirm the α satellite origin of RT-PCR amplicons, products were subsequently cloned, sequenced and mapped to the reference human assembly hg19. Sequences of unique αsatRNAs amplified in HeLa cells are presented in Table 8 (Appendix). It should be noted that most of those sequences were obtained more than once in the analyses.

Assigning a single genomic location for αsatRNAs isolated in RT-PCRs is problematic because of the repetitive nature of α satellites. Furthermore, assembly of highly repetitive regions in the hg19 human genome is not extensive. None of αsatRNAs identified in our experiments mapped back to the reference genome with 100 % identity. However, considering differences between hg19 and HeLa genomic sequences, as well as accumulation of single nucleotide polymorphisms (SNPs) in repetitive arrays, this is to be expected.

All RT-PCR products that were confirmed to be α satellite-derived by mapping to the hg19 human genome, were next aligned to the consensus sequence of α satellite unit (Prosser *et al*, 1986). Due to the repetitive nature of α satellite arrays, RT-PCR amplification yielded products longer than one unit. Therefore a concatamer of the consensus α satellite units was used for the alignment. As presented in the graphical overview (Figure 11), the vast majority of amplicons (89) were mapped in the reverse complement orientation to the unit (RC-αsatRNA) while only few (18) shared the same orientation as the α satellite concatamer (D-αsatRNA). This can either reflect the enrichment of RC-αsatRNAs within HeLa cells, or an RT-PCR bias stemming from either better efficiency of the forward primers, or lower structuredness of the RC-αsatRNA facilitating its reverse transcription.

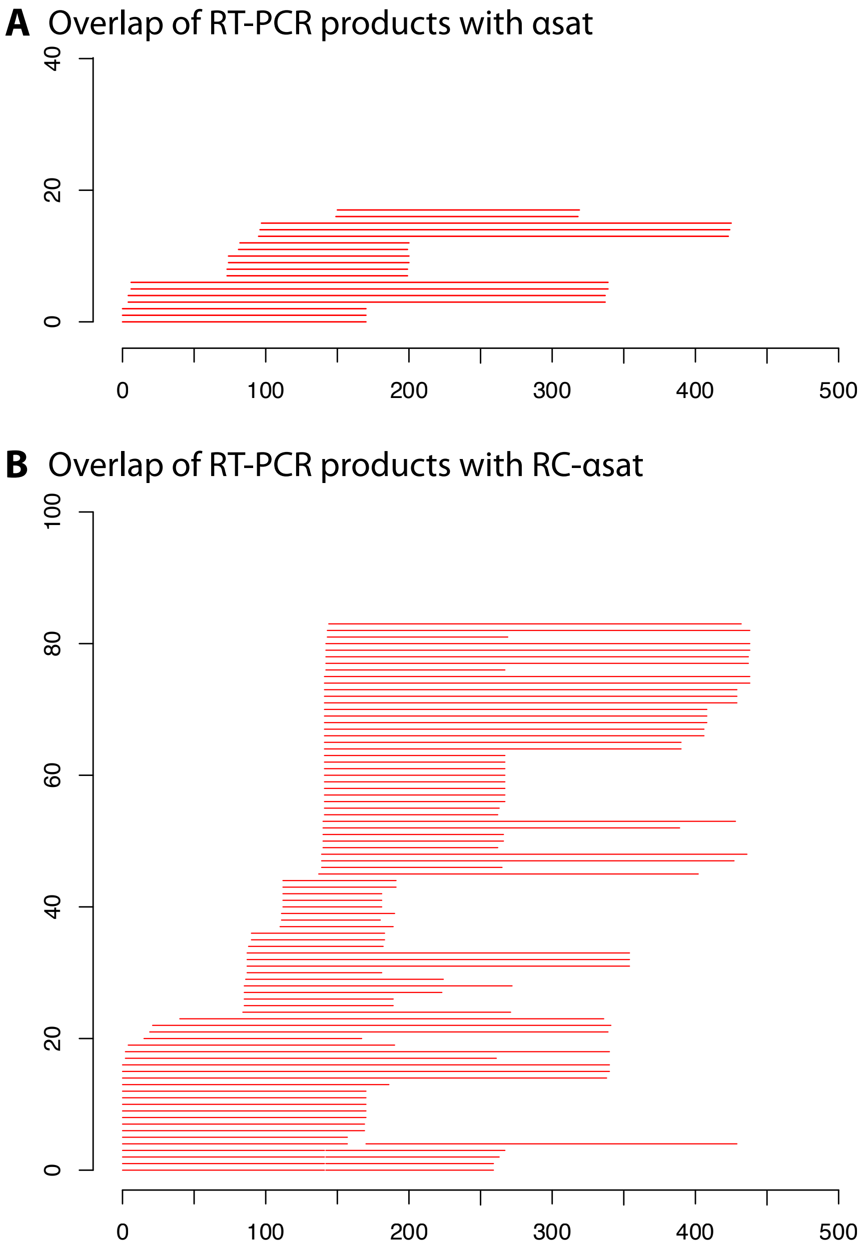


Figure 11 Representation of α satellite amplicons aligned to the concatamer of α satellite consensus units. A. D-αsatRNA-derived RT-PCR amplicons share the same orientation as the concatamer of α satellite units. **B.** Majority of transcripts retrieved in RT-PCR align to the concatamer of α satellite units in the reverse complement orientation (RC-αsatRNA). D-αsatRNA contains the direct sequence of α satellite unit, whereas its reverse complement is names RC-αsatRNA.

Although α satellite arrays span up to several kilobases, the longest fragment retrieved from RT-PCRs spaned only 3 α satellite units. This may be a consequence of inefficient cloning of longer multimers or by their instability in a plasmid. The length of the native αsatRNA was assessed by Northern blot analyses and is presented in the section 2.1.4. Further, it cannot be concluded whether any part of α satellite unit is preferentially transcribed, since the coverage on the consensus sequence is strictly determined by the RT-PCR primers' sequence.

2.1.3 α satellites are transcribed from both DNA strands

Using strand-specific RT-PCR with radioactively labeled [α-^{32}P] GTP, α satellites from both DNA strands were amplified. PCR in the presence of labeled nucleotides enables sensitive detection with significant reduction of amplification cycles (from 35 to 18) making any amplification bias less pronounced. D-αsatRNAs were reverse transcribed with the reverse primer (Rev) in the RT step, whereas RC-αsatRNAs were primed with the forward primer (Fwd). To ensure strand specificity by excluding that RNA snaps back upon itself to serve as template, an RT reaction without a primer (-) was carried out. Separately, to control for any genomic DNA contamination, the RT enzyme was omitted in the reverse transcription step (-RT).

Figure 12 shows a representative RT-PCR result obtained with the degenerate α satellite primer pair. α satellite transcripts from both DNA strands (Rev + and Fwd + lanes) were present in all conditions of HeLa culture: i) control (37 °C, ∞) (Figure 12B), ii) heat shock (45 °C, 30') and iii) heat shock followed by a recovery (45 °C, 30'; 37 °C, 60'). Products ranging from 171 bp to 1026 bp in size were detected and demonstrated the typical ladder-like pattern of bands, which is a result of primers amplifying both α satellite monomer (171 bp) and tandem repeats (342, 513 bp etc), as schematically illustrated in Figure 12A.

It should be mentioned that the direct quantification of RC-αsatRNA versus D-αsatRNA, is not possible due to the repetitive nature of α satellite arrays, the difference in forward and reverse primers efficiency, and the lack of loading control.

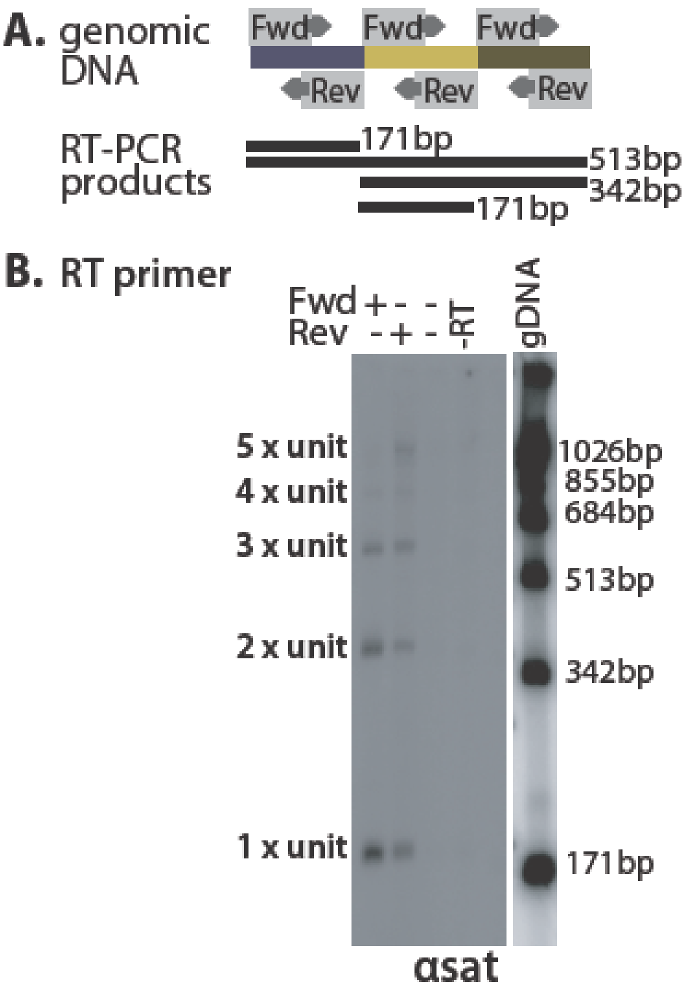


Figure 12 α satellites are transcribed from both DNA strands. A. Schematic representation of amplification from α satellite array. Primers designed to amplify α satellite unit (171 bp) hybridize also in adjacent units giving rise to a population of products varying by 171 bp. **B.** α satellite-specific RT-PCR amplicons show a typical ladder-like pattern. Radioactive RT-PCR amplicons were analyzed on 5 % native PAGE and visualized by autoradiography. Fwd primer in RT step detects RC-αsatRNA, while Rev primer – D-αsatRNA. As a control for strand specificity: RT reaction without any primer was performed (-). RT enzyme was omitted in -RT control.

2.1.4 Long transcripts arise from α satellite arrays in HeLa cells.

To determine the length of transcripts containing αsatRNA, Northern blot analyses with a consensus probe modified with LNA nucleotides were performed (Figure 13). Total RNA isolated from i) control (37 °C, ∞), ii) heat shocked (45 °C, 30') and iii) heat-shocked followed by a recovery (45 °C, 30'; 37 °C, 60') HeLa cells was resolved on a formaldehyde agarose gel, blotted onto a nylon membrane and hybridized with degenerate consensus probes in both orientations.

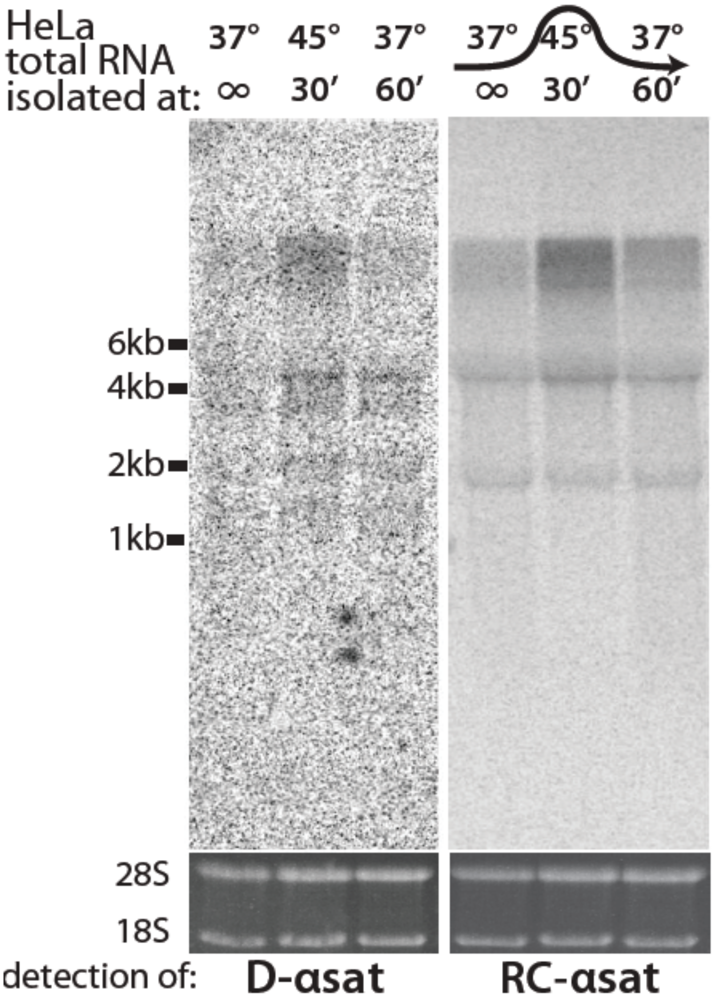


Figure 13 α satellites are transcribed into high molecular weight products in HeLa cells. An RNA blot of 30 μg total RNA from HeLa cells grown at: i) 37 °C, ∞, ii) 45 °C, 30' and iii) 45 °C, 30' followed by 37 °C, 60' hybridized with degenerate LNA-DNA probes detecting D- and RC-αsatRNAs. The profile of the ethidium bromide stained ribosomal 18S and 28S served as a loading control and is presented below the blots.

Large molecular weight bands were detected in all conditions suggesting that α satellites are transcribed from both DNA strands into long non-coding RNAs (more than 8 Kb). In addition, comparison of the band intensity between samples from different conditions, suggests that α satellites accumulate upon cellular stress, in this case heat shock. Moreover, levels of

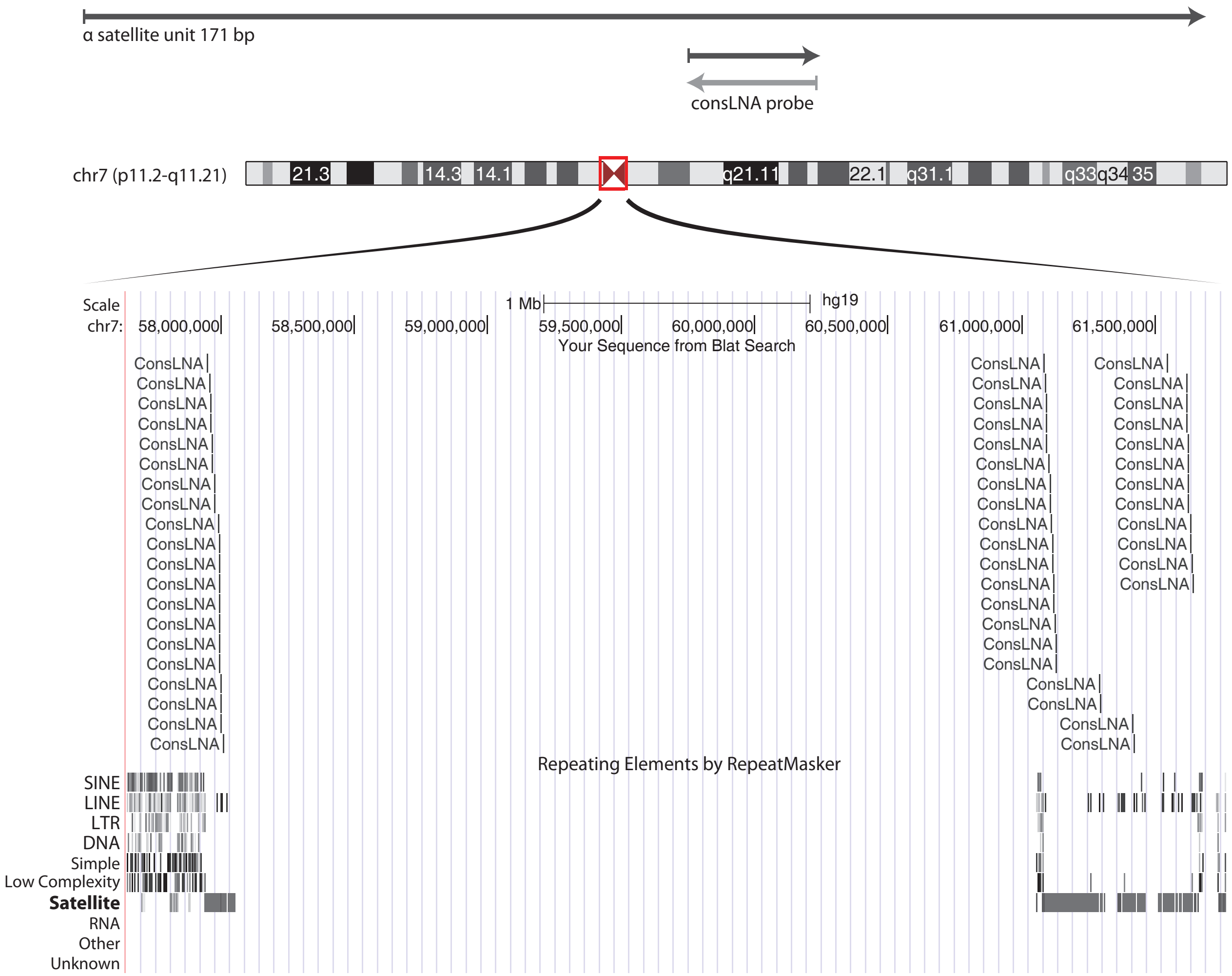


Figure 14 α satellite loci on chromosome 7 with the perfect match to the consensus sequence of Northern probes. Degenerate consensus probes map to the reference hg19 genome approximately 1700 times, maximizing the probability to detect a hybridization signal in spite of the weak expression level of α satellite transcripts. [Screenshot taken from UCSC Genome Browser].

αsatRNAs prior to and after heat shock are comparable implying a controlled expression of αsatRNAs transcription under normal conditions.

Consensus probes map approximately 1700 times with 100 % identity to all human chromosomes, except chromosome 4 and 13, enabling detection of α satellites *en masse* and thereby raising the detection probability. Additionally, presence of LNA nucleotides embedded in the probes further improves the detection sensitivity. Figure 14 demonstrates the matches of the consensus probes to the hg19 reference genome on chromosome 7, drawn in scale to α satellite unit.

2.1.5 α satellite transcripts peaks during S phase of the cell cycle

In order to determine whether α satellite repeats are constitutively expressed or follow some transient expression pattern, HeLa cells were chemically synchronized with double thymidine block and thymidine-nocodazole treatments (Wendt *et al*, 2008). Subsequently, total RNA was extracted at each phase of the cell cycle and strand specific RT-PCR analysis was carried out.

The efficiency of HeLa synchronization was verified by FACS analysis (data not presented) and immunofluorescence (Figure 15B). Cell nuclei were visualized with DAPI DNA staining. In addition, cells were stained with antibodies against PCNA (the proliferating cell nuclear antigen) or Aurora B proteins. PCNA staining identifies cells in the S phase. PCNA protein is translated in G1 phase, however during S phase it exhibits a typical granular distribution that is displaced to the nucleolus in the late S phase. The level of Aurora B protein peaks at the transition from metaphase to the end of mitosis and thus it serves as a specific G2/M phase marker.

Figure 15A demonstrates a representative RT-PCR result obtained with consensus α satellite primers. D-αsatRNAs, as well as RC-αsatRNAs were reverse transcribed and amplified from RNA at each cell cycle phase. The data suggests that the level of α satellite expression peaks during the S phase, and reduces significantly with the progression to the end of mitosis, being hardly detectable in the G1 phase. In the log phase sample, containing HeLa cells at different cell cycle phases, α satellite transcripts in both orientations were amplified as well.

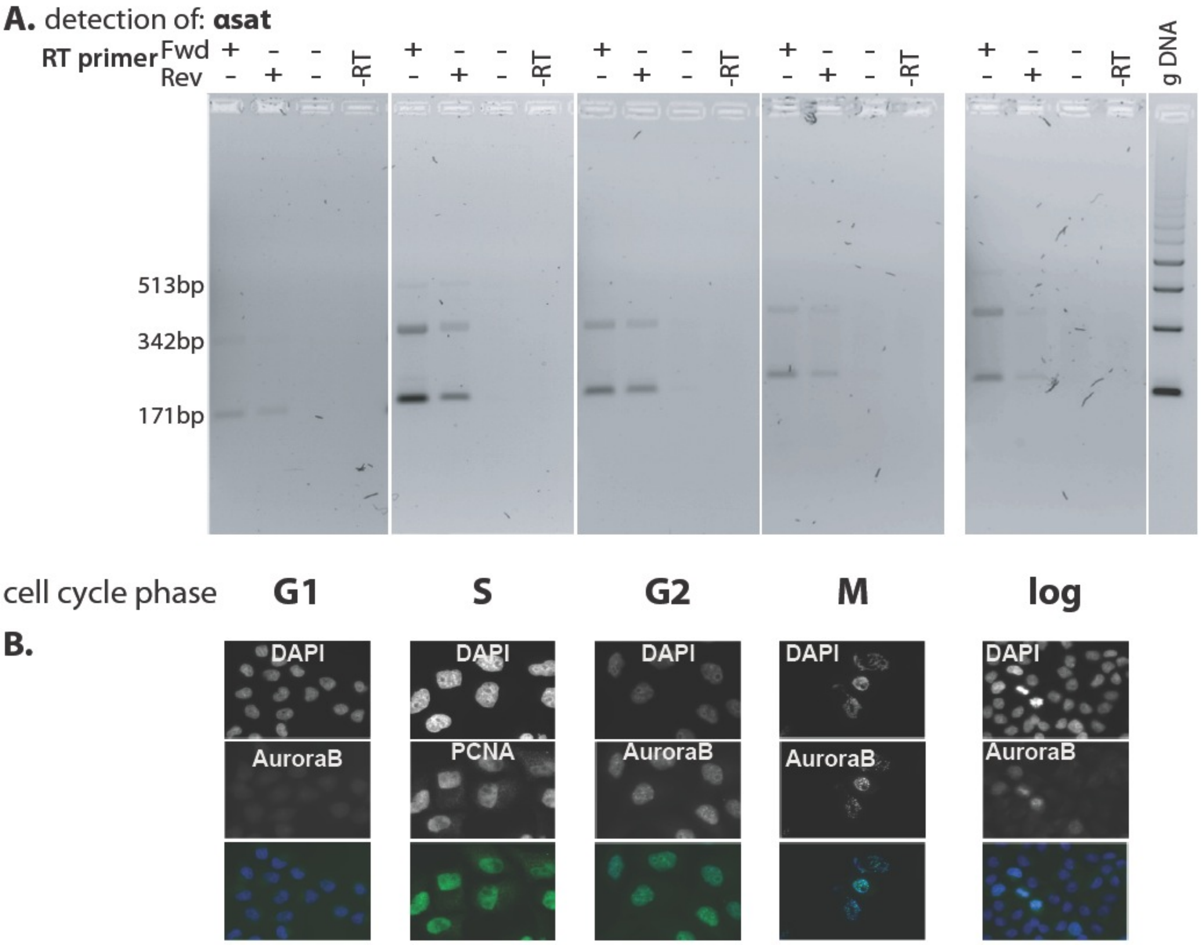

A. detection of: αsat
RT primer Fwd + - - -RT + - - -RT + - - -RT + - - -RT + - - -RT
Rev - + - - + - - + - - + - - + -
g DNA
513bp
342bp
171bp
cell cycle phase G1 S G2 M log
B.
DAPI
AuroraB
PCNA

Figure 15 αsatRNAs peak during S phase of the cell cycle. A. A representative RT-PCR result obtained with degenerate α satellite primers on RNA isolated at different cell cycle points. **B.** Synchronization of HeLa cells was verified by immunofluorescence microscopy. Nuclei are visualized with DAPI.

2.1.6 α satellites localize to the nucleus

To learn about subcellular localization of transcripts derived from α satellite repeats, total RNA from HeLa cells was separated into nuclear and cytosolic fractions. Fractionated RNA was then analyzed via Northern. As presented in Figure 16A, strong high molecular weight bands (more than 8 Kb) were detected in the nuclear fraction, and less intense signals of the corresponding size were observed in the total RNA samples. An *in vitro* transcribed α satellite monomer in D and RC orientation (invD-αsat/ invRC-αsat) served as a hybridization positive control. This result indicates that α satellites localize in the nucleus as high molecular weight transcripts.

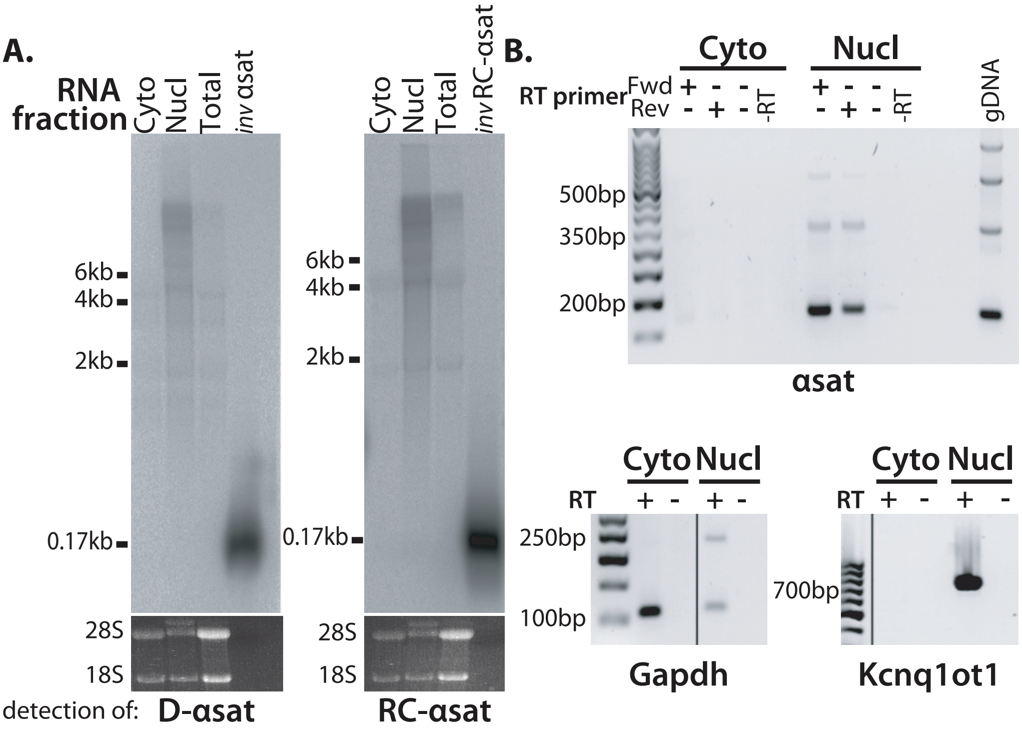


Figure 16 α satellites transcripts localize to the nucleus. A. Northern blot of 15 µg cytosolic/ nuclear/ total HeLa RNA hybridized with degenerate LNA-DNA probes detecting D- and RC-αsatRNAs. 20 ng of *in vitro* transcribed α satellite unit (inv D-αsat/RC-αsat) served as a positive control on each blot. The profile of the ethidium bromide stained ribosomal 18S and 28S served as a loading control and is presented below the blots. **B.** Strand-specific RT-PCR products on cytosolic/nuclear RNA fractions obtained with degenerate α satellite primers. Fractionation efficiency was assessed by *Gapdh* and *Kcnq1ot1* localization.

The result was additionally confirmed via strand specific RT-PCR analysis with degenerate α satellite primers. As shown in Figure 16B, α satellite transcripts derived from both DNA strands were successfully amplified from nuclear RNA fraction, but were absent in the cytosolic one. Fractionation accuracy was assessed via amplification of *Gapdh* and *Kcnq1ot1*. The *Gapdh* mRNA-specific band of 110 bp was detected in the cytosolic pool, but also in the nuclear fraction as a fainter band together with an unspliced variant of 240 bp in size. *Kcnq1ot1* RNA was observed solely in the nuclear fraction.

2.1.7 Are α satellites RNA polymerase II transcripts?

In order to find out which polymerase is engaged in transcribing α satellite arrays, we analyzed the nature of the 5’ and 3’ termini of the transcripts and the sensitivity of αsatRNAs to α amanitin, a specific inhibitor of RNA polymerase II.

To determine whether RNA polymerase II is transcribing α satellite repeats, I analyzed levels of αsatRNAs in total RNA isolated from HeLa cells treated with α amanitin at 20 μg/ml concentration, a concentration that specifically blocks RNA polymerase II (Bortolin-Cavaillé *et al*, 2009). As presented in Figure 17A, levels of α satellite transcripts diminished throughout α amanitin treatment. 6 hours post α amanitin treatment the level of αsatRNAs remained constant. After 24 hours D-αsatRNA levels were hardly detectable, and RC-αsatRNA levels reduced dramatically after 48 hours of α amanitin incubation. Strand specificity (no primer) and DNA contamination (-RT) controls were performed along, but are not presented. In parallel, well-characterized RNA polymerase II-dependent and independent transcripts were monitored as controls. The level of *Gapdh* was analyzed and shown to reduce after 48 hours post α amanitin treatment, whereas 5S RNA, transcript of RNA polymerase III, remained unchanged, serving as a negative control.

A poly(A) tail is a hallmark of transcripts transcribed by RNA polymerase II, therefore we examined whether αsatRNAs are polyadenylated. A pull down with biotinylated oligo(T) probe was performed to enrich transcripts comprising poly(A) tails. αsatRNAs expression was assayed either in the pulled down (PD) or flow through (FT) fractions. As shown in Figure 17B, α satellites from both DNA strands were more enriched in the FT pool, suggesting that they are devoid of poly(A) tails. Moreover, the bioinformatic analysis of α satellite arrays confirmed the lack of a canonical polyadenylation signal. The weak signal detected in the PD fraction might be explained by the fact that α satellites are A-rich and therefore may have some affinity to oligo(dT) probe.

Typical RNA polymerase II transcripts contain m7GpppN cap added enzymatically to their 5' end. To assess the nature of 5' terminus of αsatRNAs, we utilized a 5' adaptor ligation reaction was used (FirstChoice RLM-RACE kit). The ability to ligate a 5' adaptor to the 5' terminus of a transcript strictly depends on the nature of the 5' end of the transcript. A 5' monophosphate is required for ligation to the adapter. An RNA with a 5' cap structure will not ligate to the adapter unless first treated with Tobacco Acid Pyrophosphatase (TAP), which removes the 5' cap structure leaving a 5' monophosphate. In addition, calf intestinal phosphatase (CIP) removes a monophosphate from RNAs that do not have 5' cap structure. RT-

PCR analysis on CIP/TAP treated ligation reactions, followed by sequencing of PCR products collected from the gel, revealed α satellite-derived products only in samples treated with TAP (Figure 17C). This result strongly implies that αsatRNAs are capped.

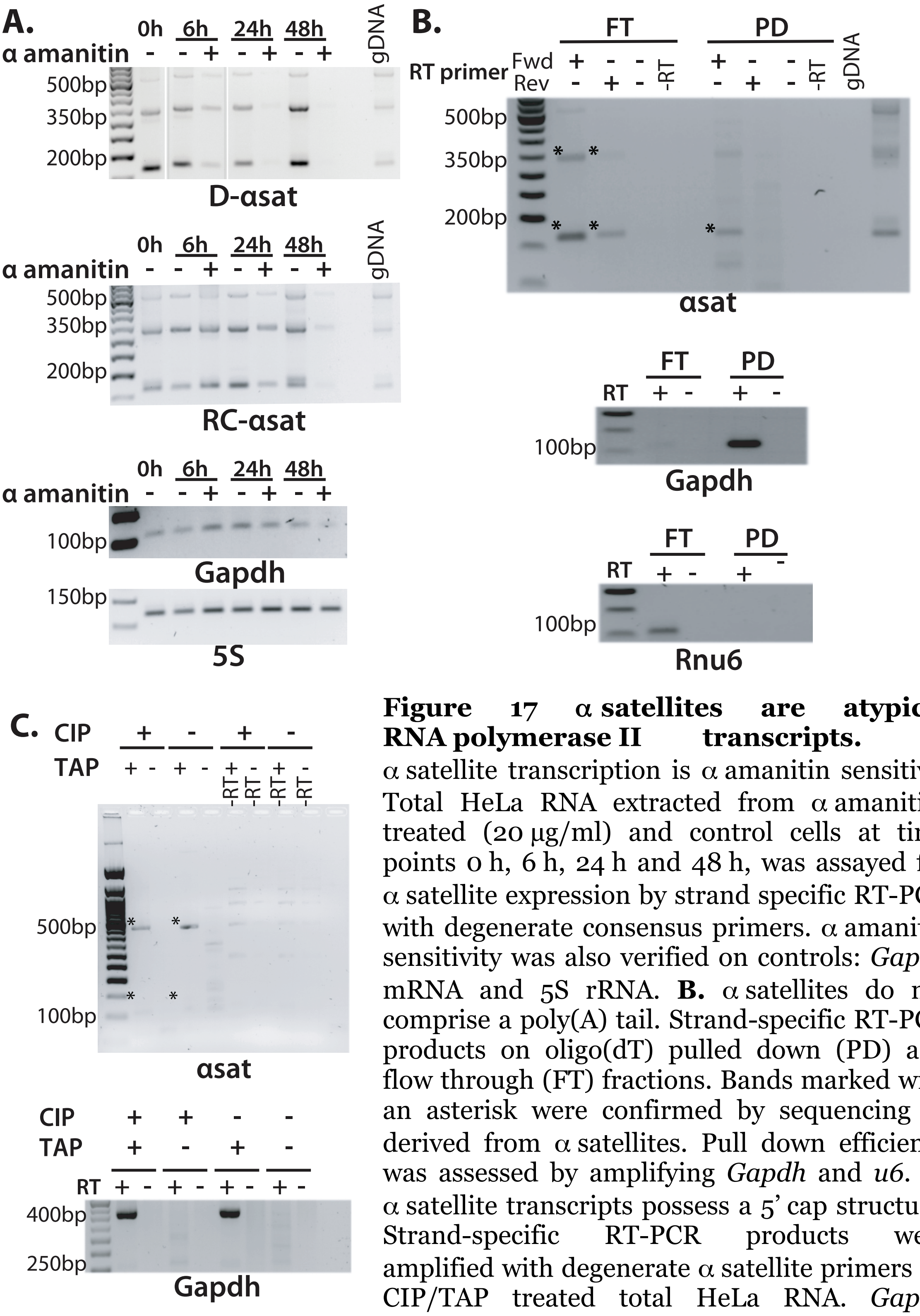


Figure 17 α satellites are atypical RNA polymerase II transcripts. A. α satellite transcription is α amanitin sensitive. Total HeLa RNA extracted from α amanitin-treated (20 µg/ml) and control cells at time points 0 h, 6 h, 24 h and 48 h, was assayed for α satellite expression by strand specific RT-PCR with degenerate consensus primers. α amanitin sensitivity was also verified on controls: *Gapdh* mRNA and 5S rRNA. **B.** α satellites do not comprise a poly(A) tail. Strand-specific RT-PCR products on oligo(dT) pulled down (PD) and flow through (FT) fractions. Bands marked with an asterisk were confirmed by sequencing as derived from α satellites. Pull down efficiency was assessed by amplifying *Gapdh* and *u6*. **C.** α satellite transcripts possess a 5' cap structure. Strand-specific RT-PCR products were amplified with degenerate α satellite primers on CIP/TAP treated total HeLa RNA. *Gapdh* transcript served as positive control.

Taken together the results demonstrate that αsatRNAs are atypical RNA polymerase II transcripts. Like most RNA polymerase II products, they are sensitive to α amanitin and capped; however devoid of poly(A) tail. The lack of poly(A) tail commonly correlates with decreased RNA stability. However, analyses of α amanitin treatment (Figure 17A) revealed an estimated half-life of 12/24 hours for D-αsatRNAs/RC-αsatRNAs, respectively. In addition, lack of poly(A) tail causes the nuclear retention of given transcripts, which was indeed confirmed and shown in Figure 16 (in section 2.1.6).

2.2 Discussion

After the ENCODE project reported that 60-70 % of the DNA was transcribed into RNA (Djebali *et al*, 2012), most of the newly identified non-coding RNAs await to be described and their biological significance assessed. The question of what portion of these RNAs is functional and how to discriminate between functional transcripts and a background noise is highly debated in the field (Struhl, 2007; Willingham & Gingeras, 2006). The RNA polymerase II machinery produces both pervasive and functional transcripts. Thus, RNAs arising from both processes share common biochemical features and are indistinguishable. As a consequence there is no straightforward approach to discriminate functional RNAs from noise.

The human genome is burdened with repetitive sequences, that either consist of simple repeats (i.e. satellite family) or interspersed transposable elements (Lander *et al*, 2001). Our knowledge about repeat-derived transcripts (repRNAs) is still lagging behind. Technical obstacles hamper analyses of repetitive regions and therefore most of the performed research simply neglects them. Nevertheless, several lines of evidence suggest that the repeat-derived transcriptome plays an important role (Bennetzen, 1996; Feschotte, 2008; Gao & Voytas, 2005; Shapiro & von Sternberg, 2005) and therefore it should be studied and not considered junk regions anymore. In order to explore the functionality of repRNA, detailed investigations on individual RNA species should to be carried out.

Results described in section 2.1 present an unbiased analysis of transcripts derived from α satellite arrays of HeLa cells. So far, there has been only a few reports on transcripts derived from human α satellite arrays (Ting *et al*, 2011; Wong *et al*, 2007; Chan *et al*, 2012). However, to our knowledge, data demonstrated here provides the first analysis characterizing α satellite-derived transcripts and determining their properties in detail.

I show that α satellite arrays are transcribed from both DNA strands into long non-coding RNAs, larger than 8 Kb (Figure 13). Northern blot hybridization did not yield any signals corresponding to the size of α satellite monomer suggesting that α satellites are not transcribed as single units. Lacking the precise assembly of centromere sequences, I assumed that the identified transcripts are derived from centromeric higher order repeats (HORs). Although the average length of centromeric array is 5.7 Kb, the Northern blot signal demonstrates that the αsatRNAs are larger than 8 Kb. This may be explained either by the poor annotation of centromeres, or by read-through from other transcriptional units.

Furthermore, if αsatRNAs in direct and reverse complement orientation were concurrently present in the cell, they could possibly form long double-stranded RNAs and further activate the RNA interference (RNAi) pathway. In this case, long double-stranded αsatRNAs could be recognized by Drosha and give rise to small interfering siRNAs. Presence of centromeric small siRNAs was 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; Murchison *et al*, 2005; Pal-Bhadra *et al*, 2004). These data demonstrate that the RNAi machinery cooperate with centromeric siRNAs to alter the local chromatin structure of the locus that codes for those RNAs. However, in the standard Northern blot assay I was unable to detect centromeric siRNAs implying that either i) long αsatRNAs are not processed into siRNAs in HeLa cells and thus human centromeric heterochromatin formation does not rely on siRNA machinery, as reported in (Wang *et al*, 2006); or ii) siRNAs are undetectable in the assays.

Comparison of the intensity of signals obtained in the Northern blot analyses (Figure 13) revealed that the steady-state levels of αsatRNAs are higher upon cellular stress, in this case heat shock. This result is in line with other studies demonstrating that the accumulation of satellite sequences is a consequence of DNA demethylation, cellular stress or genomic instability (observed in cancer) (Bouzinba-Segard *et al*, 2006; Jolly *et al*, 2004; Valgardsdottir *et al*, 2008; Ting *et al*, 2011). Our analyses were carried out in HeLa cells, the cell line derived from cervical cancer cells. However, bioinformatic interrogation of ENCODE metadata (Tafer H, personal communication) suggests that α satellite expression is probably not a result of a tumor transformation as α satellite transcripts are also found in the pool of transcripts identified in GM12878 cells, the mesoderm cell lineage (Rozowsky *et al*, 2011). Interestingly, levels of αsatRNAs detected in HeLa cells prior to heat shock and after 1 hour of the recovery time are comparable. This implies that under normal growth conditions there is a tight control of the transcriptional rate from α satellite loci, which was already demonstrated to be critical for the centromere function. Studies on human artificial chromosomes (HAC) revealed that an increase in the rate of transcription, by targeting the transcription activators to HAC centromeres and thereby opening the chromatin structure, led to the loss of kinetochore function. Interestingly, a similar phenotype was observed when transcriptional silencers were targeted to HAC centromeres to change the chromatin state into a highly repressed form. As a consequence the rapid depletion of centromeric-specific proteins CENP-A, CENP-B and CENP-C was observed (Nakano *et al*, 2008). These results show that for proper centromere function, a tightly regulated equilibrium between open and closed chromatin state is required. Another observation that active neocentromeres are formed only at lowly transcribed loci (Saffery *et al*, 2003; Ishii *et al*, 2008) further highlights the compatibility between low transcriptional rate with centromere function.

In order to estimate the expression levels of αsatRNA, deep sequenced nuclear RNA libraries from THP1 cells (Taft *et al*, 2010) and 5-8F cells (Liao *et al*, 2010) were analyzed and the abundance of αsatRNAs were compared in relation to other RNA families i.e. snoRNA, snRNA, 10 nuclear lnRNAs: KCNQ1OT1, NEAT1, MALAT1, HOTAIR, MIAT, SRA1, AIRN, HOTTIP, NRON and XIST (taken from (Ip & Nakagawa, 2012)) (H. Tafer- personal communication). To this end the following parameters were computed: i) the total number of overlapping reads, ii) the total number of reads versus the number of annotation elements in each family, iii) the total number of reads normalized to the number of nucleotides of the RNA family. 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. Considering the amount of genomic DNA covered by α satellite arrays (0.1 %), this level of αsatRNAs expression is surprisingly low.

Importantly, a lowly expressed transcript may still exert a biological function, i.e. locally in changing the chromatin state of the loci of their origin. In support of this scenario it has already been demonstrated that human centromeric RNAs are required for the proper localization of CENP-C (Centromere Protein C) at kinetochores (Wong *et al*, 2007). CENP-C is a key inner kinetochore protein that nonspecifically bind double-stranded DNA (Kwon *et al*, 2007). Du and co-workers (Du *et al*, 2010) showed that single-stranded, centromeric RNAs bind maize CENP-C to alter its structure in a way that facilitate its DNA binding affinity. The model for centromeric RNA-mediated DNA binding inferred by Du et al (Du *et al*, 2010) suggests that centromeric RNA remains at the site of transcription and acts as an epigenetic mark that converts CENP-C:DNA binding to a stable, functional state.

On the other hand, a lowly expressed RNA may be just a by-product of the process of transcription that *per se* is important. Foltz and colleagues showed that FACT (Facilitates Chromatin Transcription complex) coimmunoprecipitates together with CENP-A nucleosomes. CENP-A is the histone H3 variant that is a hallmark of an active centromere (Allshire & Karpen, 2009). Data reported by Foltz implies that RNA polymerase II transcription, resulting in the opening of the chromatin structure, promotes deposition of the CENP-A into centromeric nucleosomes (Foltz *et al*, 2006), which is similar to the observation from fission yeast (Folco *et al*, 2008).

Analysis of αsatRNAs expression throughout the cell cycle (Figure 15) revealed that transcripts level reach the peak during the S-phase of the cell cycle. Also centromeric regions in fission yeast (Chen *et al*, 2008), as well as in mouse (Lu & Gilbert, 2007) were shown to be transcribed in the S-phase dependent manner. Generally, during DNA replication in the S-phase, silencing marks are shortly reduced being just placed on a newly replicated strand, which results in the opening of the chromatin structure. As a consequence, the S-phase may be a window of opportunity for the transcription start to occur. Lyn Chan et al. provided evidence that there is an active RNA polymerase II transcribing centromeric α satellite arrays at the kinetochores of metaphase and anaphase during mitosis in human cells. This observation may suggest that αsatRNAs play a role in the kinetochore protein binding, similar to previously published data in (Du *et al*, 2010).

Data presented in the section 2.1.7 provide evidence that αsatRNAs are synthesized by RNA polymerase II. As most RNA polymerase II products, αsatRNAs possess a typical cap at the 5' end and their synthesis is inhibited by the α amanitin. But as shown in Figure 17, αsatRNAs are detected in the pool of transcripts devoid of poly(A) tails, which is atypical for RNA polymerase II transcripts. The lack of a poly(A) tail may cause the nuclear retention of those transcripts, which was indeed detected and presented in Figure 16. Additionally, transcripts devoid of poly(A) tails are commonly less stable. However, this does not seem to be the case for αsatRNAs. As delineated from α amanitin experiments (Figure 17), the half-life of αsatRNAs is over 12 hours. Upon confirmation of

RNA polymerase II-mediated α satellite arrays transcription, we performed a bioinformatic interrogation to look for transcription factor binding sites within α satellite arrays. Transcription factor binding motifs were retrieved from JASPAR database (Sandelin *et al*, 2004). The analysis revealed random composition of weak transcription factor binding sites what could result in uncoordinated transcription start sites (H.Tafer, personal communication). 5' end adaptor ligation analysis confirmed the heterogeneous nature of the 5' ends of αsatRNAs. Our observation is well supported with results reported by Bulut-Karslioglu et al (Bulut-Karslioglu *et al*, 2012). They showed that mouse repetitive heterochromatin regions contain many transcription factors binding sites that are not synergistically coordinated into regulatory modules, such as promoters and enhancers. They inferred a model for transcription factor-based heterochromatin formation. Uncoordinated transcription does not result in generation of tightly controlled and efficient synthesis of transcripts; instead it triggers the silencing via transcription factor-coupled (Delattre *et al*, 2004) or RNA-mediated (Nagano *et al*, 2008) recruitment of histone methyltransferases.

2.3 Human α satellite transcripts are substrates for RNA polymerase II

2.3.1 αsatRNAs interact with the RNA polymerase II

α satellite DNA plays a well-described role in binding CENP proteins crucial for the kinetochore formation (Stimpson & Sullivan, 2010). However, knowledge about α satellites role at the RNA level is fairly poor. In order to learn about a potential function of α satellite transcripts, an approach to identify a protein interacting partner was chosen.

Initially, we exploited the data from Genomic SELEX combined with deep sequencing performed to isolate regulatory RNAs with high affinity to RNA polymerase II, previously obtained in our group (Boots JL, von Pelchrzim F, Weiss A, Zimmermann B et al, *in preparation*). RNA polymerase II Binding Elements termed PBEs, from an enriched pool after seven rounds of amplification and selection were deep sequenced and mapped back to the hg19 version of the human genome. In the final pool of PBEs, 314 unique α satellite-derived aptamers (Table 9, Appendix) were identified and mapped to both strands of α satellites (177 for direct, 137 for reverse complement orientation).

Upon aligning all α satellites PBEs to the consensus sequence of two tandem α satellite units, it was apparent that most of the aptamers span the junction of the units. Additionally, high variability within the sequence of α satellite aptamers allowed identifying a sequence motif using the MEME search tool, as presented in Figure 18. For both D- and RC-αsatRNAs the identified motifs are characterized by an overrepresentation of G/A on their 5' and T/C on their 3' part. The motif is a putative binding site for the RNA polymerase II.

2.3.2αsatRNAs are used as templates for transcription in HeLa nuclear extract

RNAs with a potential to directly interact with RNA polymerase II might regulate the process of transcription. In order to test the regulatory potential of PBEs, *in vitro* transcription assays, performed in HeLa nuclear extracts, were established.

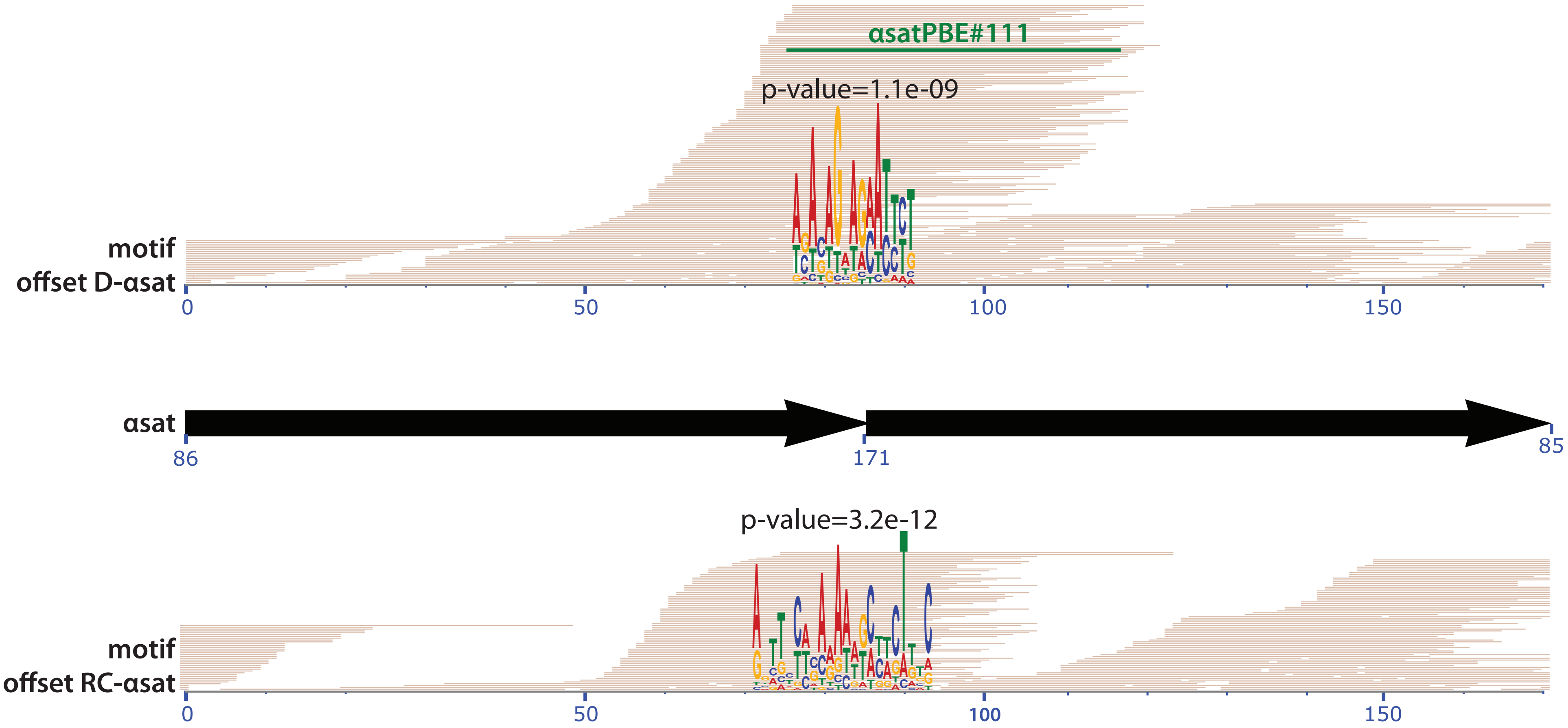


Figure 18 Overlaps between the RNA polymerase II-binding aptamers. The partial α satellite units are shown in brown. Because the motifs are located at the boundaries between two α satellite units, the position of the overlaps is shown with respect to their location in the offset α satellite, i.e. satellite 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.

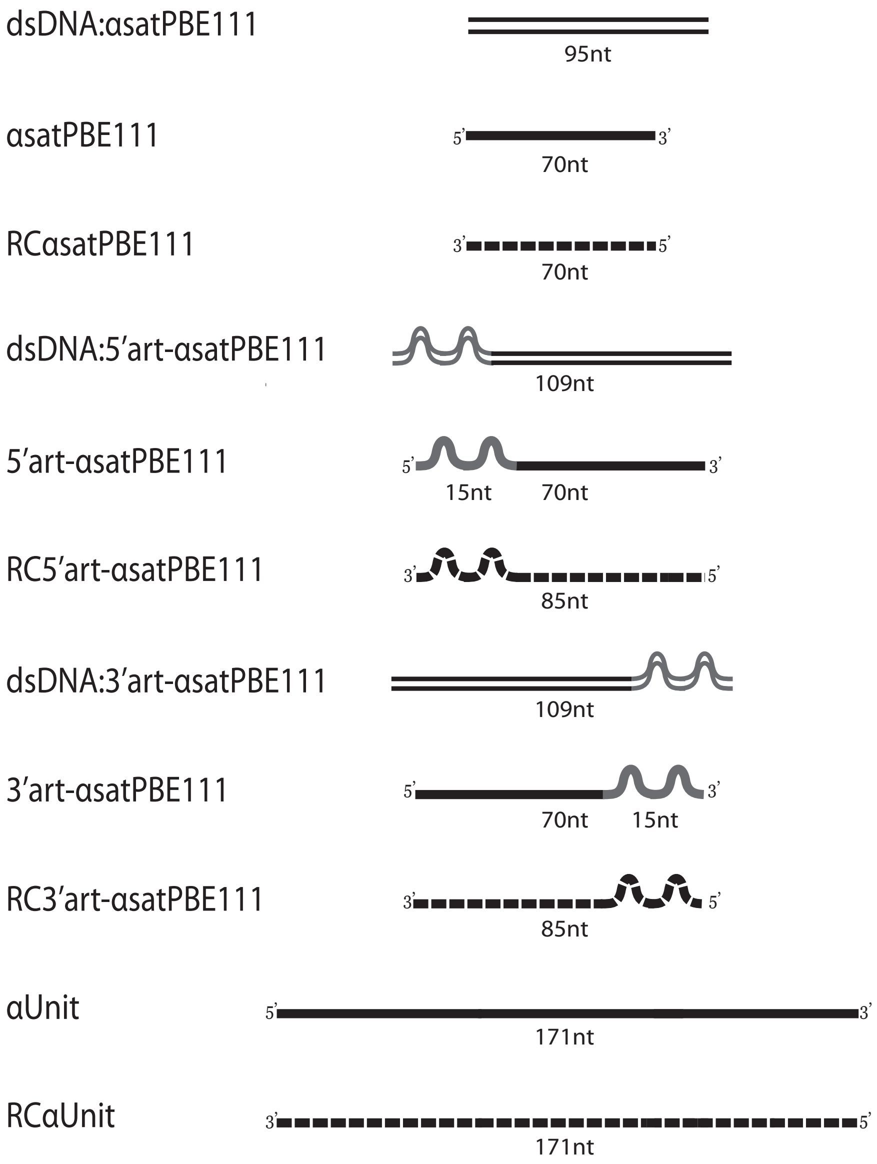


Figure 19 Schematic representation of αsatRNA templates for the transcription assay. Black lines represent α satellite sequences, grey lines denote artificial sequence absent from the human genome, dashed lines denote reverse complement sequence, and double lines are double stranded DNA fragments containing T7 promoter sequence used to transcribe the given RNA.

Several different transcription templates, including the α satellite sequences, were constructed to assess the impact of α satellite-derived PBEs on RNA polymerase II activity: i) 'αsatPBE #111' a 70 nucleotide long aptamer from the PBE pool (marked in Figure 18) mapping to chromosome 19 and the reverse complement to it 'RCαsatPBE #111'; ii) several chimeric templates consisting of 'αsatPBE #111' fused with an artificial 15 nucleotides, not present in the human genome, positioned at the 5' or 3' termini of the 'αsatPBE #111' RNA, termed: '5'art-αsatPBE #111', '3'art-αsatPBE #111' and their reverse complement counterparts; iii) finally, a 171 nucleotide long α satellite unit in direct and reverse complement orientation, termed 'αUnit' and 'RCαUnit'. All templates tested in the *in vitro* transcription experiments are depicted in Figure 19.

2.3.3α satellite RNAs are labeled when incubated in HeLa nuclear extract

Incubation of the 70 nucleotides long αsatPBE #111 RNA in HeLa nuclear extract in the presence of a radioactively labeled [α-^{32}P] GTP resulted in the appearance of labeled products (Figure 20A). Titration of increasing amounts of HeLa nuclear extract led to a stronger labeling of αsatPBE #111. The length of the labeled products correlate well with the size of the input RNA, however products appear as a smear indicating that there is a population of RNAs differing in length by a few residues. There was no radioactively labeled product when a double-stranded DNA fragment was used as a transcription template. Similar results were obtained upon incubation of the 171 nucleotides long αUnit RNA, its reverse complement RCαUnit (Figure 20B), as well as the 40 nucleotides long αsatPBE #276 in HeLa nuclear extract (data not shown). The pattern of labeled bands in each experiment corresponds to the size of the input RNA, but appearing as smear combined with a discrete band. Importantly, every nuclear extract preparation batch contains some residual nucleic acids that give raise to background transcripts present in every reaction (bands marked with asterisks in Figure 20A).

To determine if the observed labeling is specific to α satellite RNA, we tested several other RNA templates, which yielded no signals in the assay. Among them were not only members of repetitive class: SINE element Alu (25-50 nM) and acromeric satellite ACRO (50 nM) that belongs to PBEs, but also 60-mer RNA with a sequence not matching the hg19 human genome at 50 nM concentration (data not shown).

Taken together these data demonstrated that α satellite RNA is specifically labeled in HeLa nuclear extract suggesting the RNA is either a substrate for RNA-dependent RNA synthesis (RdRP) or is extended at its 3' terminus.

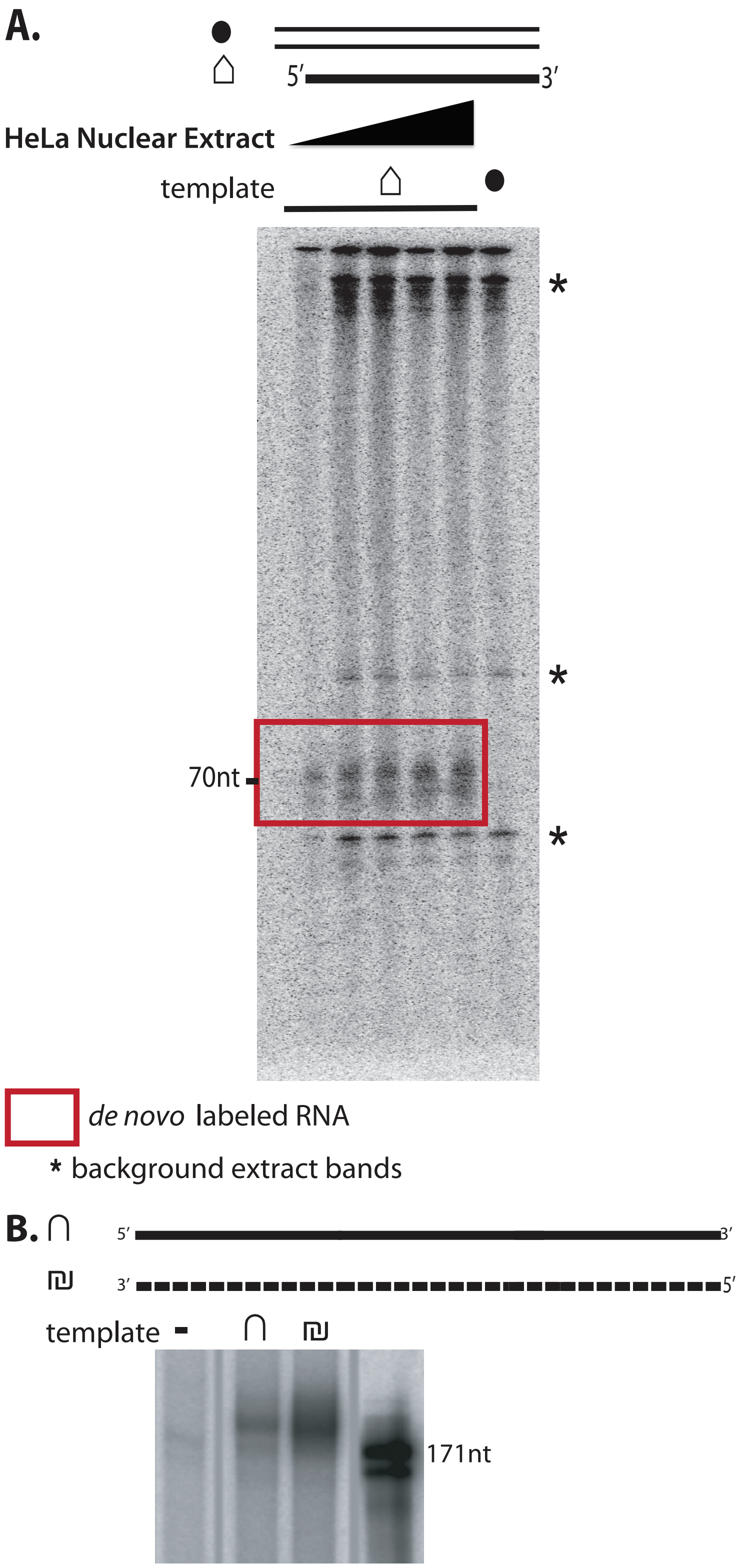


Figure 20 α satellites are labeled in HeLa nuclear extract in the presence of radioactive nucleotides. A. Products of incubation of 25 nM αsatPBE #111 template with increasing amount of HeLa nuclear extract in the presence of $[\alpha\text{-}^{32}P]$ GTP resolved on 8 M urea 8 % polyacrylamide gel. Red box depicts labeled αsatPBE #111. Asterisks in the no template lane (-) represent background products transcribed from nucleic acids remaining in the nuclear

extract preparation. **B.** Products of incubation of 50 nM αUnit RNA and its reverse complement RCαUnit in HeLa nuclear extract in the presence of [α-^{32}P] UTP resolved on 8 M urea 8 % polyacrylamide gel. Reaction without added template (-) serves as a background transcription control. End-labeled αUnit (last lane) is a marker for the 171 nucleotides long RNA.

2.3.4Location of the α satellite-specific sequence determines the size of the labeled products

In order to test whether the observed labeling activity of the α satellite sequence is a result of extending the RNA and/or is used as a template (RdRP), we designed chimeric templates consisting of α satellite-derived 70-mer (αsatPBE #111) and artificial 15 nucleotides (absent from the human genome) fused either to the 5' or 3' termini of the RNA. Thus, 5'art-αsatPBE #111 and 3'art-αsatPBE #111 vary solely by the position of the non-human sequence. As shown in Figure 21A, incubation of 100 nM 5'art-αsatPBE #111 and 3'art-αsatPBE #111 within HeLa nuclear extract in the presence of [α-^{32}P] UTP resulted in transcripts differing in size depending on where the artificial sequence was fused in the template. The same observation is valid for αsatPBE #276 (data not shown).

When the non-human sequence was located at the 3' terminus of the template (3'art-αsatPBE #111), there was a discrete band corresponding to the size of 70 nucleotides, and a smear of products of higher molecular weight. Shuffling the artificial sequence to the 5' end of the construct (5'art-αsatPBE #111) resulted in a radioactive product matching the size of an entire template accompanied by a smear of higher bands. Incubation of the double stranded DNA (dsDNA:5'art-αsatPBE #111) fragment did not yield any labeled bands.

All together the presented data suggest that the position of the α satellite aptamer within the template determines the size of labeled transcripts. We propose that the α satellite sequence recruits RNA polymerase II, presumably engaged in RNA-dependent RNA polymerization. This hypothesis is supported by the fact that RNA polymerase II polymerase-binding motif identified *in silico* (Figure 21B) is contained in templates used in the assays.

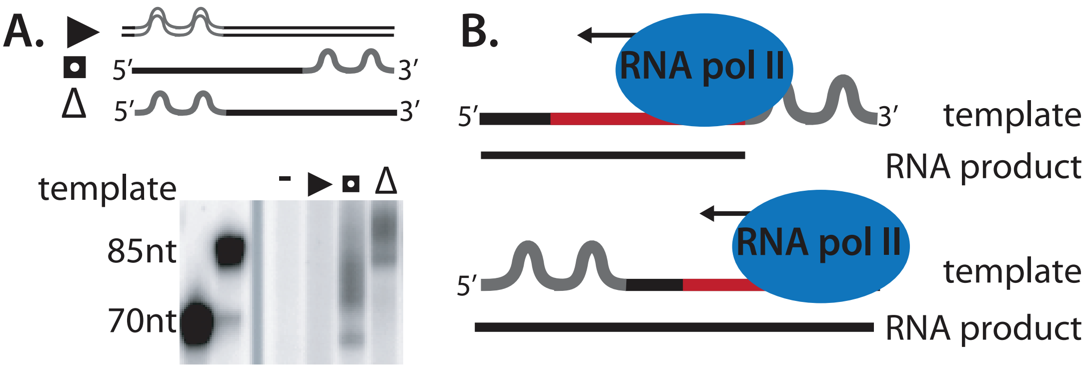


Figure 21 Location of α satellite aptamer determines the size of a labeled product. A. Schematically depicted templates illustrate: double stranded DNA fragment containing T7 promoter (▶), 70-mer of αsatPBE #111 with 15 nucleotides fused (absent from human genome) to the 3' or 5' terminus of the template (◘, Δ, respectively). Products of incubation of 100 nM templates in HeLa nuclear extract in the presence of [α-^{32}P] UTP analyzed on 8 M urea 12 % polyacrylamide gel and visualized by autoradiography. Reaction with no template added (-) controls for background transcription. End-labeled 70 and 85-mer RNAs served as a size marker. **B.** Depending on the location of the α satellite-specific aptamer within the template, the polymerase either omits the artificial 15-mer and starts transcribing at the 3' end of αsatPBE #111 (top) or includes the 15-mer giving rise to a product corresponding in length to the native template (bottom).

2.3.5 Labeling of αsatRNAs is sensitive to DRB inhibitor

To decipher which enzyme is responsible for the observed modification of αsatRNAs in HeLa nuclear extract, transcription reactions were performed in the presence of 60 μM 5,6-dichloro-1-beta-D-ribofuranosylbenzimidazole (DRB). DRB blocks RNA polymerase II elongation by inhibiting CDK9 associated with Positive Transcription Elongation Factor (P-TEFb).

As show in Figure 22, 5'art-αsatPBE #111 incubated in HeLa nuclear extract gives rise to 85 nucleotides long products. However, no products were detected when DRB, a specific inhibitor for an elongating RNA polymerase II, was added to the reaction (comparison between DRB +/- samples). RNA polymerase II may exert two activities leading to the observed modification of αsatRNAs, namely: *de novo* synthesis of the complement strand via RNA-dependent RNA polymerase (RdRP) activity (Filipovska & Konarska, 2000) and/or 3' terminus extension of αsatRNAs. It is worth mentioning that the two mechanisms are not mutually exclusive. But still, a result negating one mechanism, instantly points to the second one. Experiments described in the following sections were conducted to determine which mechanism triggered the observed αsatRNAs modifications.

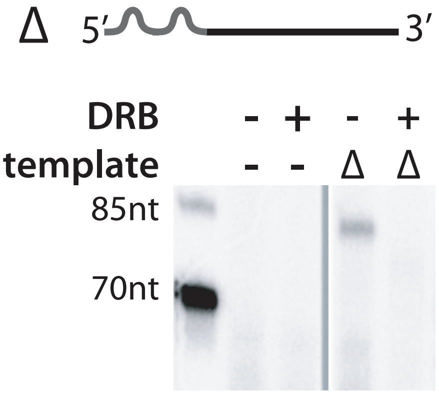


Figure 22 Modification of αsatRNAs is inhibited once DRB is added to the reaction. 50 nM 5'art-αsatPBE #111 template was incubated in HeLa nuclear extract with [α-^{32}P] UTP in the presence or absence of 60 μM DRB inhibitor (+/-). Resulting transcripts were analyzed on 8 M urea 8 % polyacrylamide gel. Reactions without added template (-) served as a control for background. End-labeled 70 and 85-mer RNAs served as a size marker.

2.3.6αsatRNA is a template for *de novo* synthesis of a complement RNA strand

To address the question whether αsatRNAs are substrates for RNA polymerase II-mediated RdRP activity, products of transcription within HeLa nuclear extract were assayed by primer extension.

First, input 5'art-αsatPBE #111 and RC5'art-αsatPBE #111 RNAs were incubated separately in HeLa nuclear extract in the transcription buffer devoid of labeled nucleotides. The transcription reaction was stopped, precipitated and analyzed by primer extension. The isolated RNA pool potentially contains: i) input RNA, ii) shortened input RNA resulting from backtracking RNA polymerase II, iii) input RNA extended at the 3' terminus by a few residues, or finally iv) *de novo* synthesized, complementary strand. Primers hybridizing to input (control) or putative complementary strand (RdRP product) were used to delineate which strand was present in the output RNA pool, thus establishing whether RdRP is occurring. The caveat of this approach is that it does not allow detection of extended residues, it only determines whether there was RdRP activity. A reverse primer anneals at the 3' end of the 5'art-αsatPBE #111 so upstream to the putative extended residues. Extension of the RNA by a few bases might be either templated or random, thus prediction of the extended sequence to design a primer for it was not feasible.

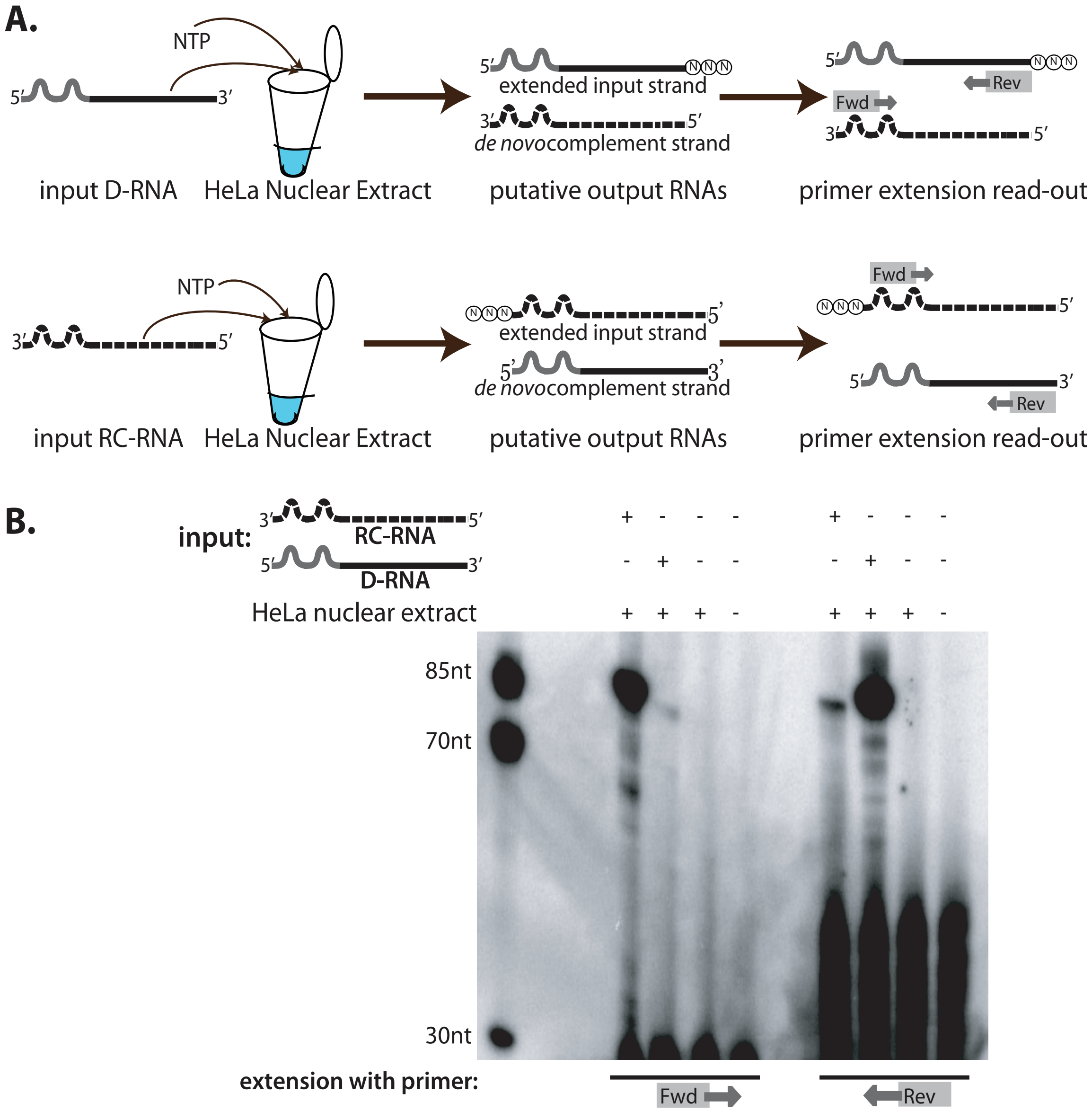


Figure 23 5’art-αsatPBE #111 and RC5’art-αsatPBE #111 are templates for RNA-dependent RNA synthesis. A. Scheme for the experimental setup. **B.** Products of primer extension on RNA pool precipitated from HeLa nuclear extract incubation of 50 nM RC5’art-αsatPBE #111 and 5’art-αsatPBE #111. Samples probed with Fwd are loaded in lanes 2-5, with Rev in lanes 6-9. To control unspecific priming by any background transcript in the output RNA pool, Fwd and Rev primer were hybridized to reaction products when no template was added to HeLa nuclear extract (lane 4, 8). End-labeled, 30-mer Fwd and Rev primers were run in lane 5 and 9.

As schematically shown in Figure 23A, two input RNAs were used to assay for RdRP activity (the synthesis of the complementary strand): the **5’art-**

αsatPBE #111 (direct input RNA, D-input) and its reverse complement **RC5'art-αsatPBE #111** (RC-input). Two different primers were used in a primer extension assay, following incubation of the two input RNAs in HeLa nuclear extract, to determine if the complementary strand was synthesized in the nuclear extract. For the **D-RNA input** the reverse primer recognized the input strand and served as a control. In addition, the forward primer recognized the putative RdRP product and resulted in a band detected at 85 nucleotides, implying that the D-RNA input is a template for RdRP. The corollary is observed for the **RC-RNA input**, the forward primer served as a control for the input, while the reverse primer detected the RdRP-dependent strand. In Figure 23B, samples were loaded according to the primer used in the primer extension reaction. For example, when **RC5'art-αsatPBE #111** (RC-input) was incubated in HeLa nuclear extract, the input strand was detected with forward primer (lane 2) as a strong band of 85 nucleotides accompanied by shorter products resulting from the abortive reverse transcription. Importantly, the reverse primer also was extended to approximately 85-mer, demonstrated as a fainter band in lane 6. This result suggests the presence of *de novo* synthesized RNA strand that is complement to the RC-input RC5'art-αsatPBE #111. Similarly, when **5'art-αsatPBE #111** (D-RNA) was used as a template for *in vitro* transcription reaction, the input strand was recognized by reverse primer and extended to the full size product (lane 7). Shorter bands correspond to products of incomplete reverse transcription. Additionally, a fainter signal was detected with the forward primer (lane 3), again pointing to a product of the RdRP activity.

2.3.7 αsatRNAs are extended at the 3' end upon incubation in HeLa nuclear extract

To be able to directly detect the 3' terminus extension of the input RNA, a reverse RPA approach was developed. The experimental design is shown in Figure 24A. αsatPBE #111 RNA was incubated in HeLa nuclear extract in the presence of [α-^{32}P] UTP. Once the transcription reaction was stopped and RNA precipitated, it was incubated with an unlabeled complementary RPA probe and subjected to the A/T1 RNases treatment. Samples were subsequently analyzed on 8 M urea 8 % PAGE. The RPA probe hybridizes with the input (radioactive) αsatPBE #111 presumably leaving the 3’ overhang formed by extended residues, which will be digested by A/T1 RNases.

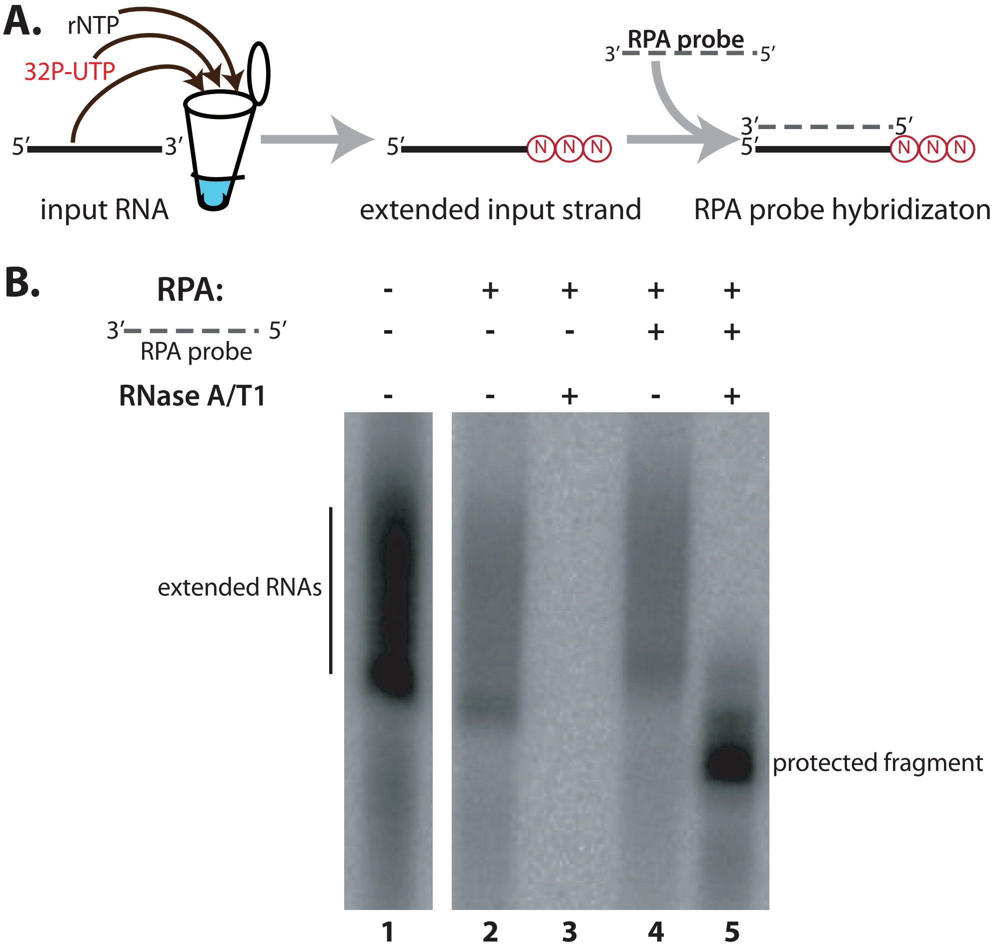


Figure 24 Reverse RPA detects residues added to the 3’ end of αsatPBE #111 RNA. A. Schematic representation of the experimental setup. 100 nM was incubated in HeLa nuclear extract in the presence of [α-^{32}P] UTP. RNA products were precipitated and used as input for RPA: hybridized to complementary, unlabeled RPA probe and digested by A/T1 RNases. **B.** Products of αsatPBE #111 RNA incubation in HeLa nuclear extract are presented in the left panel. Resolved reverse RPA products are shown in the right panel. Lane 1, 2: control samples with no probe; lane 3, 4: samples hybridized with the RPA probe.

The denaturing gel shown in the left panel in Figure 24B shows radioactively labeled products extracted from the incubation of αsatPBE #111 RNA in HeLa nuclear extract (lane 1). Next, those transcripts were hybridized with unlabeled RPA probe (lane 4, 5), or incubated with no probe to serve as a control (lane 2, 3). In the RNA pool extracted from the HeLa nuclear extract reaction there were transcripts protected by the probe, which is demonstrated by the retention of the signal on the gel (lane 5). Moreover the protected fragments are detected as shorter bands and a significantly reduced smear, when compared to RNase untreated control (lane 4). This result clearly indicates that residues added to the 3' terminus of αsatPBE #111, where digested by RNases A/T1. It is important to stress that RNase A cleaves the 3' end of unpaired cytosine and uridin residues, whereas RNase T1 cleaves 3' to guanosine residues. Therefore, products of A/T1 RNases treatment are not blunt-ended and hence still contain labeled residues, what allows its detection. The shift in size of the protected fragment may be explained by the presence of population of transcripts resulting from the backtracking RNA polymerase II.

2.3.8RNA polymerase II extends αsatRNA at the 3'end

An additional result showing 3' extension of the αsatRNA is provided by the experiment conducted by Stacey Wagner in the collaborating laboratory of J. F. Kugel and J. A. Goodrich (Figure 25).

αsatPBE #276 RNA was incubated with the purified RNA polymerase II in the transcription buffer. The experiment was designed in two setups: i) input αsatPBE #276 was incubated with RNA polymerase II together with ribonucleotide mixture containing radioactively labeled [α-^{32}P] CTP, and ii) end-labeled αsatPBE #276 RNA was incubated with RNA polymerase II together with unlabeled nucleotides, in case if there was a specificity for the included residues (Figure 25A). Products of the transcription reaction were analyzed on the denaturing PAGE. As shown in Figure 25B, both variants of the experiment revealed that αsatPBE #276 RNA is extended by RNA polymerase II. Products of the reaction performed in the presence of radioactive ribonucleotides migrated slower than end-labeled αsatPBE #276 template. Size difference between the extended (lane 2) and end-labeled (lane 1) αsatPBE #276 RNA corresponds to approximately 20 nucleotides. In the sample where γ end-labeled αsatPBE #276 RNA was incubated with RNA polymerase II and unlabeled ribonucleotides, there was a visible up-shift of the band. This suggests that RNA polymerase II interacted with a fraction of end-labeled αsatPBE #276 and extended it by several residues.

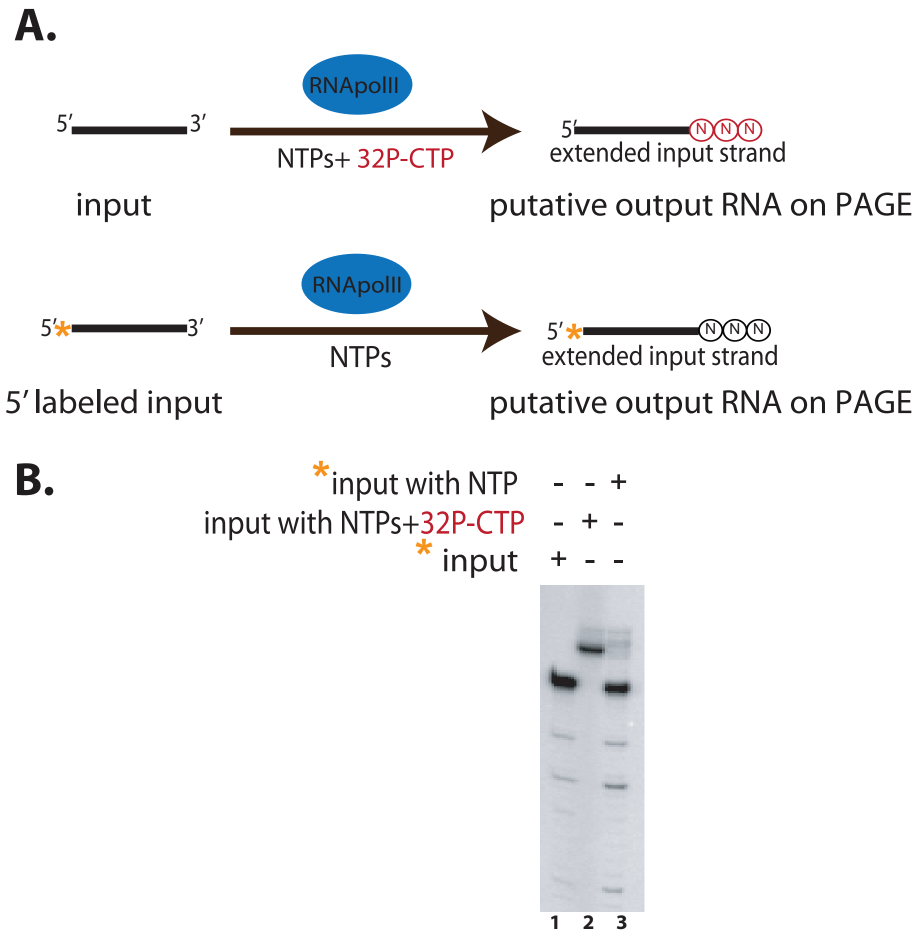


Figure 25 αsatPBE #276 RNA is extended by the purified RNA polymerase II. A. Schematic representation of the experimental setup. **B.** Products of incubation of 50 nM cold/end-labeled αsatPBE #276 with the purified RNA polymerase II in the presence of NTP mixture/NTPs supplemented with [α-^{32}P] CTP analyzed on 8 M urea 12 % polyacrylamide gel. Lane 1: end-labeled αsatPBE #276 as size marker; lane 2: unlabeled αsatPBE #276 incubated with NTPs and [α-^{32}P] CTP; lane 3: end-labeled αsatPBE #276 with NTPs.

For the RNA molecule to be extended by a few residues, a 3’OH group is required. Therefore, to abolish the addition of extra nucleotides, templates lacking 3’ OH were transcribed and incubated in HeLa nuclear extract. In order to obtain RNA with the blocked 3’ terminus, Hepatitis Delta virus (HDV) ribozyme sequence was fused downstream to α satellite templates. Catalytically active HDV ribozyme was self-cleaved during *in vitro* T7 transcription leaving 3’ cyclic phosphate at the 3’ terminus of the 5’art-αsatPBE #111 and RC5’art-αsatPBE #111. As shown in Figure 26, templates with 3’ cyclic phosphate and controls with 3’ OH were incubated in HeLa nuclear extract in the presence labeled [α-^{32}P] UTP. Then, RNA products were precipitated and analyzed on denaturing gel.

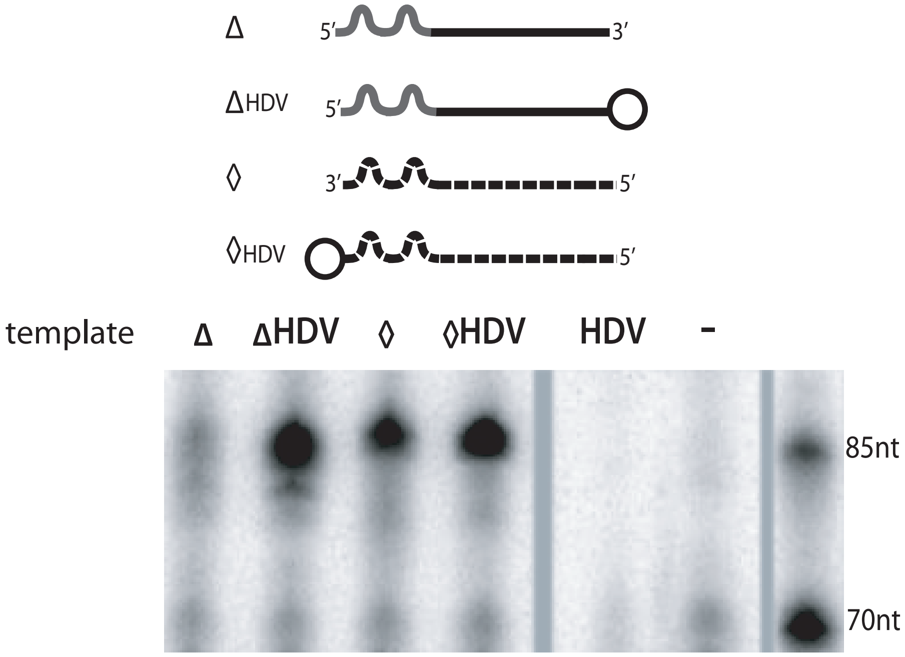


Figure 26 Incubation of αsatRNAs with blocked 3' terminus in HeLa nuclear extract still results in radioactively labeled products. Products of incubation of 50 nM 5'art-αsatPBE #111 and RC5'art--αsatPBE #111 (either with 3' cyclic phosphate – depicted as HDV, or 3' OH) within HeLa nuclear extract in the presence of [α-^{32}P] UTP resolved on 8 M urea 12 % polyacrylamide gel.

Surprisingly, there is no apparent difference between products obtained from the templates with blocked 3' terminus compared to the controls. There are a few explanations for this inconclusive result: i) the 3' cyclic phosphate is unstable, ii) RNA polymerase II removes the 3' cyclic phosphate when backtracking or iii) detected transcripts represent only products of RNA-dependent RNA synthesis.

2.3.9αsatRNAs serve as substrates for RNA-dependent RNA activity *in vivo*

In order to test whether αsatRNAs may serve as substrates for RNA-dependent RNA activity *in vivo*, 5'art-αsatPBE #111 was nucleofected into HeLa cells. There is no sequence similarity between the artificial 15-mer located at the 5' terminus of the template and the human genome. The artificial sequence was designed to enable the discrimination between *de novo* synthesized and endogenous αsatRNA. *In vitro* transcribed 5'art-αsatPBE #111 RNA was purified and nucleofected into HeLa cells. 24 hours post nucleofection, total RNA was isolated and assayed for sense and antisense strands of αsatRNA. Strand-specific RT-PCR was aimed to detect the *de novo* synthesized strand, complementary to the nucleofected 5'art-αsatPBE #111. The forward primer used in the assay hybridizes to the artificial 15-mer, assuring the detection of only the RNA derived from nucleofection. As shown in Figure 27, we detected RNA antisense to the nucleofected 5'art-αsatPBE #111, a clear RdRP product. The second strand was absent in the input used for nucleofection, what is shown on the left panel as a control.

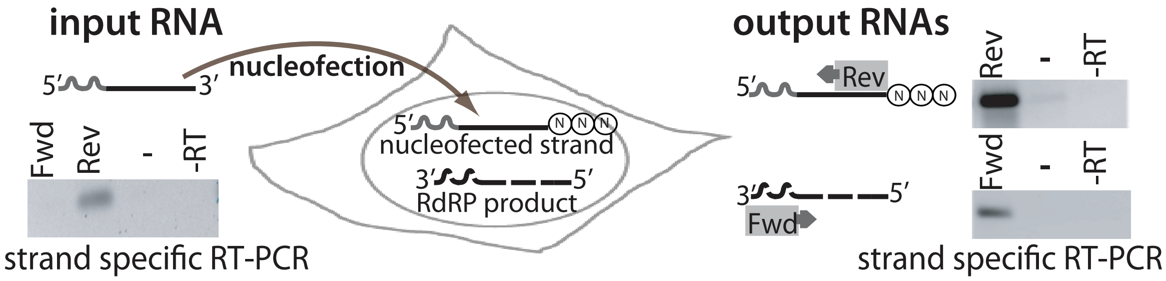


Figure 27 asatRNA is a template for RdRP activity *in vivo*. Strands-specific RT-PCR on total RNA isolated from HeLa cells 24 h post nucleofection with RNA template consisting of αsatPBE #111 fused with non-human 15-mer absent from the human genome. Results obtained on total RNA from nucleofected cells are shown on the right site, on the input RNA - on the left site.

The presented result clearly shows that αsatRNAs are taken as templates for *in vivo* RNA-dependent RNA polymerase activity, presumably held by RNA polymerase II.

2.4 Discussion

The function of α satellite RNAs was unknown prior to this study. We hypothesized that the αsatRNAs may play a role in transcriptional regulation due to the isolation of α satellite-derived aptamers in Genomic SELEX with RNA polymerase II as bait. Considering that i) α satellite-derived RNAs bind RNA polymerase II, ii) RNA polymerase II harbors RNA-dependent RNA polymerase (RdRP) activity (Filipovska & Konarska, 2000; Lehmann *et al*, 2007), and that iii) we were able to detect α satellite transcripts from both strands, we focused on αsatRNAs as putative substrates for RNA polymerase II activities. All results presented in section 2.3 provide evidence that αsatRNAs interact with the active side of RNA polymerase II and are extended by several residues at the 3' end and/or taken as templates for *de novo* synthesis of the complementary strand. I was able to detect RdRP activity *in vitro* in HeLa nuclear extract, as well as in HeLa cells. The most definitive experiment demonstrating that αsatRNAs are templates for RNA-dependent RNA synthesis shows that the size of RdRP product is determined by the position of αsatPBE aptamer on the RNA template (Figure 21). Additionally, I also demonstrated that αsatRNAs are extended at the 3' terminus by several bases, what was clearly assayed in the reverse RPA assay (Figure 24) or confirmed by the incubation of αsatPBE #276 with the purified RNA polymerase II (Figure 25).

Notably, none of our assays demonstrate that the products of the observed RNA polymerase II activities are functional. Their cellular function remains unknown, or they may simply be by-products of the RNA polymerase II dissociating the redundant RNA locked in its active site. It was reported that under stress conditions, the bacterial RNA polymerase is blocked by 6S RNA that mimics the structure of the transcription bubble (Trotochaud & Wassarman, 2004). Once the stress condition is released, the polymerase escapes the 6S RNA by transcribing p19 RNAs on the 6S template via RdRP activity (Wassarman & Saecker, 2006). The process of transcription on the 6S RNA template destabilizes the complex freeing the active site from the small RNA. Whether the p19 RNAs have a function is still unknown. Additionally, bacterial RNA polymerase was also shown to extend several small RNAs at their 3' termini that were bound to its active site. This was observed by Windbichler et al (Windbichler *et al*, 2008) in our laboratory. Moreover, Wagner and colleagues demonstrated a similar mechanism for the human RNA polymerase II. They found that RNA polymerase II uses mouse B2 RNA as substrate and template for its RdRP activity and also extends B2 by 20 nucleotides (Wagner *et al*, 2013). Although the results presented here imply already reported mechanisms, they demonstrate for the first time that human endogenous RNA is used as a template for RNA-dependent RNA synthesis by human RNA polymerase II. This expands the biochemical relevance of the observed phenomenon showing the versatility of RNA polymerase II activities on a bread spectrum of RNA targets. It is therefore to be expected that RNA polymerase II is not just involved in synthesizing RNAs, but that it has a wide spectrum of biochemical activities to resolve inactive complexes.

In conclusion, we hypothesize that αsatPBEs may either recruit RNA polymerase II to centromeres to promote centromeric heterochromatin remodelling or they can block the RNA polymerase II active site contributing to the tight regulation of the low transcription levels from the centromeric loci.

2.5 Human α satellite transcripts contain remnants of snoRNAs

2.5.1 Consensus structure of α satellites includes remnants of functional RNA

In order to look for a putative progenitor of αsatRNAs, we performed *in silico* structural and sequence analyses (H.Tafer, personal communication). Dfam, a database of human repetitive DNA elements (Wheeler *et al*, 2013), classifies α satellites by the sequence similarity into 3 families: ALR, ALRa and ALRb. For each of those families, the consensus sequence was derived with clustalw and subsequently folded with RNAalifold. ALRa and ALRb showed the highest degree of structuredness, containing a conserved stem within the 5' region. To minimize the impact of large sequence variability within α satellites on the computation process, the consensus structure was recomputed by aligning the consensus sequences of all three α satellite families. The resulting fold is composed of a hairpin-hinge-hairpin containing H- and ACA-Boxes. This observation suggests that α satellites may have originated from snoRNAs.

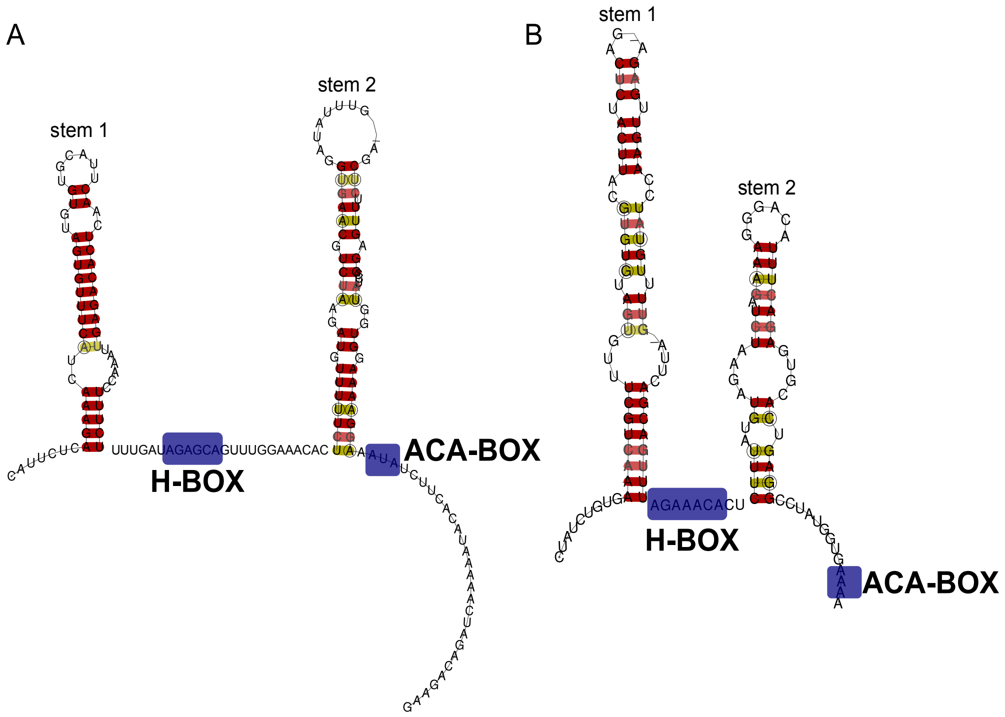


Figure 28 Consensus structure of human α satellite consensus sequences extracted from Dfam database. In blue typical snoRNA- like elements: H and ACA boxes are marked. Red-coloured base pairs show no compensatory mutation. Ochre base pairs have one compensatory mutation at the given position.

To further study the phylogeny of αsatRNAs, primates' genomes were scanned for homologs of human α satellites. All α satellite sequences recovered from Dfam were blasted against primates' phylogenetic tree, but in a reverse order. First, the chimpanzee genome was searched for α satellite homologs. The resulting alphoid sequences identified in the chimpanzee genome were added to the query used to screen the next closest genome. The search was repeated in chimpanzee, gorilla, orang-utan, gibbon, macacca, phillipine tarsier, grey mouse lemur and greater galago. The approach led to the identification of alphoid sequences up to the marmoset genome, what was earlier reported in (Shepelev *et al*, 2009). 5 of the 30 marmoset alphoid sequences were mapped into introns of coding genes. These sequences were further inspected (aligned and folded) for the presence of snoRNAs-like structures. In the consensus fold the hairpin-hinge-hairpin-tail structure of H/ACA snoRNAs was found (Figure 28). These observations hint at snoRNAs being putative ancestries of α satellites.

2.6 Discussion

The last and presumably most innovative part of our α satellites analysis arose from the interest in the evolutionary origin of repetitive RNAs (repRNAs). Brosius hypothesized that many non-coding RNAs are remnants of the RNA world (Brosius, 2003). There are a few non-coding RNAs that are reported to have evolved from other RNAs, e.g Alu originated from 7SL RNA (Ullu & Tschudi, 1984) or BC1 derived from tRNA (DeChiara & Brosius, 1987).

While analysing the structure of α satellites, we observed that they contain structural hallmarks of H/ACA snoRNAs. Intrigued by this observation, we studied their phylogeny and show here that αsatRNAs fulfil all bioinformatic criteria for a snoRNA classification. They are composed of two long stem loops linked by a single-stranded region with the H box (ANANNA) and a tail with the ACA box. They share some sequence complementarity with the ribosomal rRNA that enables its pseudouridylation by base-pairing mechanism (Kiss, 2002). We indeed identified putative pseudouridilation sites on 28S rRNA and U5 snRNA by the RNAsnoop target prediction tool (Tafer *et al*, 2010). However, none of those sites were shown to be modified yet.

Our hypothesis is well supported by several already reported facts. It was reported that mammalian genomes contain new snoRNA copies via the copy-and-paste mechanism (Weber, 2006). Further, Schmitz et al provided evidence that snoRNAs can be transposed into new genomic locations by being mobilized into retrotransposable element, called snoRTE (Schmitz *et al*, 2008). Moreover, the fold of the most evolutionary distant αsatRNAs homologues, found in marmosets, contain a short 40 nucleotide long 3' overhang, presumably being a fossil of the retrotransposable element. Finally, depletion of dyskerin, a core component of H/ACA snoRNPs, disrupts the formation of mitotic spindle in HeLa cells resulting in the raised mitotic index (Alawi & Lin, 2013).

To conclude, we propose that α satellites are most probably descendents of snoRNAs

3 Concluding remarks

The results described in this dissertation represent a novel approach aimed at finding novel functions within RNAs derived from human repetitive regions. A plethora of the ENCODE-reported non-coding transcripts still requires functional identification. Recently, the non-coding transcriptome has been receiving increasing attention, however non-coding RNAs derived from repeats escape most researchers notice. The research on highly repetitive regions of the human genome is lagging behind mainly due to the mistaken conviction that these parts of the genome are of low complexity. There are also technical obstacles in the computational annotation that hamper the analysis. A fact that calls for consideration is the necessity to apply non-canonical methods to explore repetitive parts of the human genome. This work represents an unbiased attempt to screen the entire human genome for a particular property, in this case direct interaction with RNA polymerase II, via genomic SELEX combined with deep sequencing.

From the descriptive characterization of αsatRNAs, we learn that human centromeres are transcribed into long, non-coding RNAs. Our experiments do not provide evidence whether αsatRNAs derived from both DNA strands remain single stranded or form duplexes. We would expect that the long double-stranded RNAs triggered the RNAi-mediated response resulting in the presence of small centromeric RNAs. We failed in detecting those and thus we speculate that αsatRNAs do not form double-stranded complexes. However, in the light of facts on the centromeric heterochromatin formation triggered by the RNAi mechanism, reported even for *S. pombe* (Volpe *et al*, 2002), it would be of importance to settle whether similar mechanisms are required for human centromeric heterochromatin formation. In order to address this, two experiments may be conducted: complexes formed by double-stranded RNAs may be immunoprecipitated by J2 monoclonal antibody against long dsRNA and probed for αsatRNAs. Alternatively, Northern blot for enhanced detection of small RNAs (Pall & Hamilton, 2008) with probes covering the whole α satellite unit could be conducted to yield more information. Additionally, it would be possible to search for hallmarks of A to I editing, which occurs on dsRNAs. However, due to the high variability of αsatRNA sequences, it is not easily feasible to determine the origin of a reverse-transcribed αsatRNA and thus prove RNA modification events.

Results contained in the functional analysis of αsatRNAs provide evidence that centromeric RNAs interact directly with RNA polymerase II being substrates for its activities: extension of the 3' end of the RNA bound in the active site and/or RNA-dependent RNA synthesis.

Thus far, RNA polymerase II activity was observed on different αsat-derived RNA templates: αsatPBE #111, αsatPBE #276 as well as the whole monomers: αUnit and RCαUnit. Although the putative RNA polymerase II-binding motifs were identified bioinformatically, it would be useful to biochemically map the interaction and define the minimal binding motif. This could be assessed by electrophoretic mobility shift assays with a series of differently sized and mutated α satellite templates. Moreover, on the basis of the observed 3' end extension of tested αsat-derived templates, we assumed that α satellite transcripts bind to the active site of RNA polymerase II. However, competition binding assay with f.e FC(*) aptamer (Kettenberger *et al*, 2006) or UV cross-linking experiment would accurately map the interaction sites.

In addition, the biological meaning for the interaction between α satellites and RNA polymerase II requires further elucidation. Nakano et al demonstrated that the targeting of transcription activators and repressors to centromeric loci disturbed proper centromere function (Nakano *et al*, 2008). It is conceivable that this tight control of transcription is triggered by the ribo-regulation of RNA polymerase II by local RNAs. As shown by RNA FISH (Wong *et al*, 2007), αsatRNAs localize to centromeric loci at mitotic chromosomes to exert the local function. It is also important to emphasize that assessing a function for αsatRNAs is particularly difficult since centromeric DNA plays such a prominent role in the cell. Knocking out α satellite arrays is not feasible since centromeres are indispensable for chromosomes to segregate properly and thus the cells with chromosomes devoid of centromeric loci cannot divide.

In order to shed light on αsatRNAs function, it would be very informative to identify their protein partners. To this end the ChOP experiment could be performed, as described in (Mariner *et al*, 2008). In the chromatin oligoaffinity precipitation (ChOP), a biotinylated anti-αsatRNA oligonucleotide would serve to purify αsatRNA-associated interactors from formaldehyde-treated cells. The affinity-purified complexes could be further analyzed by mass spec to investigated protein partners or by tilling arrays to identify DNA loci where αsatRNA bound. Alternatively, SILAC-based RNA pull down (Butter *et al*, 2009) may be undertaken.

The αsatRNAs impact on transcription could be assayed *in vitro* in the minimal RNA polymerase II transcription system, similarly as reported by Espinoza et al (Espinoza *et al*, 2004). The rate of transcription from an RNA polymerase II template can be measured in the presence of titrated αsatRNAs and compared to the one affected by already reported RNA polymerase II inhibitors, like Alu (Mariner *et al*, 2008). However, the caveat is what α satellite sequence to choose. The functional αsatRNA domain requires being determined first. Another, yet risky, experiment to test αsatRNAs potential to regulate RNA polymerase II transcription could be performed in the reporter assay in transfected HeLa cells described by Han and co-workers (Han *et al*, 2004).

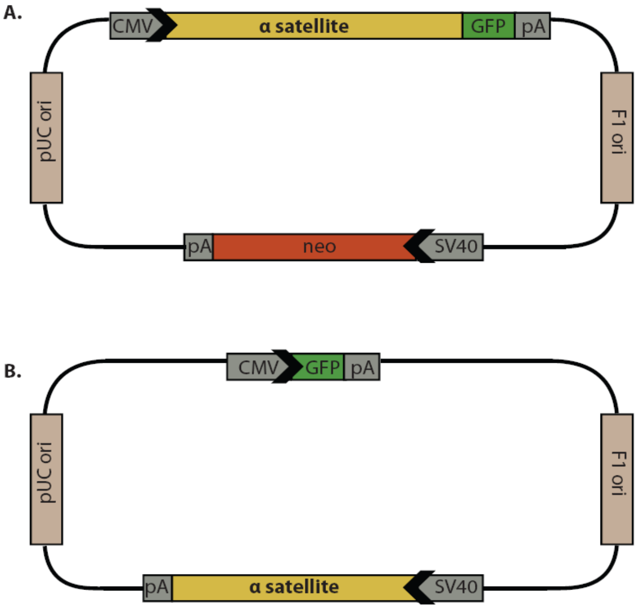


Figure 29 Plasmid maps to be used in the reported assay to test αsatRNAs impact of RNA polymerase II transcription (modified (Han *et al*, 2004)**.**

To learn about the co-transcriptional function of αsatRNAs, a large portion of a native α satellite array could be cloned upstream to GFP in the same transcriptional unit and the GFP fluorescence can be measured to check the change in the transcriptional output (Figure 29A). Using the same reporter assay platform, *in* *trans* impact of αsatRNAs on transcription could be studied as well. In this case, α satellite transcripts would be driven separately from the GFP transcript and monitoring the GFP fluorescence could yield some information on whether αsatRNAs regulate RNA polymerase II engaged on other transcriptional units (Figure 29B). However, some technical obstacles need to be considered for the described experimental setup. It would be burdensome what portion and what α satellite's sequence to amplify and clone. Moreover, designing a proper negative control would be critical. A fragment of a congruent size of a gene could be used on a trial basis.

The last part of my dissertation presents a hypothesis on the phylogeny of αsatRNAs. We propose that αsatRNAs are presumably descendants of H/ACA snoRNAs that migrated within transposable elements. In order to gain some insight into the secondary structure of αsatRNAs the *in vivo* chemical probing could be undertaken. It would be important to test whether the typical snoRNA hairpin-hinge-hairpin structure, as predicted *in silico*, is formed *in vivo*. Moreover, it would be interesting to assess whether dyskerin, a member of H/ACA snoRNPs and a homolog to the yeast centromere binding protein Cbfp5, is among αsatRNAs' interacting partners.

4 Materials and methods

4.1 HeLa cells 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 °C in 5 % CO_2 atmosphere. For the analysis of α satellite RNAs expression, HeLa cells were cultured under native as well as stress conditions. 80 % confluent cultured cells were subjected to heat shock at 45 °C for 30 minutes with subsequent recovery at 37 °C for 1 hour, when stated. Control cells were maintained at 37 °C.

4.1.1 α amanitin treatment

5×10^5 HeLa cells were seeded 42 hours 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 α satellite RNAs expression.

4.1.2 Cell cycle synchronization

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

4.1.3 HeLa cells nucleofection

1×10^6 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 hours, nucleofected and control cells were harvested for total RNA isolation.

RNA for nucleofection	Sequence
5'art-α111	cgcgcgtgaggccatcacgacagaagaattctcagtaacttccttgtgttgtgtgtattcaactcacagagttgaacgttccttt

Table 4 Sequence of the template for nucleofection.

4.2 RNA isolation

4.2.1 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 at the Spectrophotometer ND-1000 (Nanodrop), and by comparing the intensity of the 28S to 18S ribosomal bands on agarose gel electrophoresis. Then, two consecutive DNase I (NEB) treatments were performed at 37 °C for 30 minutes to remove potential genomic DNA contamination. The RNA was purified by standard phenol/chloroform extraction, ethanol precipitation and collected by centrifugation.

4.2.2Nuclear/Cytoplasmic RNA fractionation

Separation of nuclei from cytoplasm was done using the modified Sambrook and Russell protocol. 80 % confluent cells were washed with ice-cold 1 x PBS prior to being scraped off. Cell suspension was spun down at 4 °C for 5 minutes at 2000 x*g*. Then, the cell pellet was resuspended in ice-cold 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 underlain with an equal volume of Lysis Buffer containing 24 % w/v sucrose. Nuclei were fractionated by density gradient with ultrafugation at 4 °C for 20 minutes at 10000 x*g*. The cytoplasmic fraction was recovered and subjected to proteinase K digestion (200 µg/ml). The nuclear pellet was resuspend in Lysis Buffer, nuclei were disrupted and the liberated genomic DNA was sheared mechanically by repeatedly squirting the solution through a thin needle (19 gauge). Then, the nuclear fraction was digested with the proteinase K (200 µg/ml). RNA from both fractions was purified by standard phenol/chloroform extraction, precipitated and collected by centrifugation.

4.3 Strand-specific Reverse Transcription

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-labeled primer was denatured at 70 °C for 10 minutes and quickly chilled on ice. Reverse transcrtiption reaction (OmniScript, Qiagen) was performed at 45 °C for 1 hour in a total volume of 20 µl according to the manufacturer's protocol. "No primer" and "no reverse transcriptase" controls were included. Afterwards, the reverse transcriptase was heat-inactivated and removed by standard phenol/chloroform extraction. 5 μl of the reverse transcription reaction was amplified by PCR and analyzed on agarose gel or, in case of primer extension, on 8 % polyacrylamide gel (1 x TBE, 8 M urea).

Primer	Sequence
αCons_F	cattctcagaaacttctttgtgatgtatac
αConsNest_F	cttctgtctagtttttatatgaagatattcc
αCons_B	cttctgtctagtttttatatgaagatattcc
αConsNest_B	gttgaatgtatacatcacaaagaagtttc
α111_F	cacgacagaagaattctcagtaac
α111_B	gctctgtctaaaggaacgttcaac
αchr19_F	tcatgtaaggtagacagaag
αchr19-outer_F	ggaaacgggatttcttcatataaggcac
αchr19_bF	cgtttcaaaactagacagaatcattcccg
αchr19_eF	gcagatttcagacactcattttgtggaa
αchr19_fF	ctgcaagtggatatttggatctagtaga
αchr19_B	aagggaaggttcaactctgtcagttg
αchr19-outer_B	agcgtgtttcaaatctgctctgtctaaa
αchr19_bB	agagtgtttccaaatggctctatgaaaag
αchr19_eB	ccactatatgaagaaatcccgtttcca
αchr7_F	gaatcactctttttgtagaatacgcatttag
αchr7_B	gcacacatctcaaagaagtttctgag
Gapdh_F	cgggaagcttgtcatcaatgg
Gapdh_B	cgccagtggactccacgac
GapdhNest_B	catattgagggacacaaggttac
Kcnq1ot1_F	acagtggggtactgggatct
Kcnq1ot1_B	cgctattgggatggaagtt
Rnu6_F	gtgctcgcttcggcag
Rnu6_B	aaaatatggaacgcttcacg
5S_F	gtctacggccataccaccctgaa

5S_B	aaagcctacagcacccggtattcc
5Snest_B	tgcttagcttccgagatcagacg

Table 5 Sequences of PCR primers.

4.4 5’ Rapid Amplification of cDNA Ends

Analysis of 5’ ends of transcripts derived from α satellites was performed using Rapid Amplification of cDNA Ends method (FirstChoice RLM-RACE Kit, Invitrogen), according to the manufacturer’s instructions.

4.5 Cloning of PCR products

4.5.1 Preparation of DH5α competent cells

Single colony from DH5α plate was inoculated into 5 ml LB medium and cultured at 37 °C overnight at 180 rpm. The overnight culture was then diluted in 100 ml LB medium and grown until $OD_{600} = 0.6$. Next, cells were incubated on ice for 15 minutes and collected by centrifugation at 4 °C for 10 minutes at 4000 rpm. The pellet was gently resuspended with 10 ml sterile-filtered, ice-cold 0.1 M $CaCl_2$ and incubated on ice for at least 1 hour. Cell suspension was spun down at 4 °C for 10 minutes at 4000 rpm. Cell pellet was again gently resuspended in 5 ml sterile, ice-cold 0.1 M $CaCl_2$ with 15 % glycerol and aliquots were kept on ice until snap-frozen in liquid N2 and transferred to -80 °C.

4.5.2Ligation of PCR products into a plasmid

PCR products were purified over an agarose gel or a column (Wizard® SV Gel and PCR Clean-Up System, Promega) and ligated with the pGEM-T Easy vector (pGEM®-T Easy Vector Systems, Promega) according to the manufacturer’s protocol. Positive and negative controls were included.

4.5.3Chemical transformation of DH5α cells

2 µl of the ligation reaction or 5 ng control plasmid DNA was added to DH5α competent cells. After 30 minutes incubation on ice, cells were heat-shocked in a water bath at 42 °C for 45 seconds and immediately put on ice for 5 minutes. Next, cells were incubated with agitation at 37 °C in 1 ml SOC or LB medium without

antibiotics to allow the recovery. After 1 hour of incubation, cells were plated onto selective LB plates containing X gal and incubated at 37 °C overnight.

4.6 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 minutes and added to the denatured 200 µg of total RNA resuspended in 0.5 x SSC. Magnetic beads were captured and washed 6 times with 0.1 x SSC. Enriched RNA was eluted from the beads with water and 50 ng were analyzed with strand-specific RT-PCR.

4.7 RNA analysis

4.7.1 5' end-labeling

Templates: i) chemically synthesized DNA oligonucleotides, ii) LNA modified oligonucleotides, or iii) *in vitro* transcribed RNAs, were end-labeled with [γ-^{32}P] ATP by T4 Polynucleotide Kinase (NEB) at 37 °C for 30 minutes. Unincorporated nucleotides were removed using spin-column chromatography (Illustra MicroSpin G-25 Columns, GE Healthcare Life Sciences). In case of RNA probes, they were dephosphorylated with Calf Intestinal Phosphatase (NEB) at 37 °C for 1 hour, purified by standard phenol/chloroform extraction and ethanol precipitated prior to the labeling reaction.

4.7.2 Northern blot analysis

RNA samples were dissolved in 2 x denaturing loading dye (95 % formamide, 0.025 % bromophenol blue, 0.025 % xylene cyanol FF, 0.025 % ethidium bromide, 0.5 mM EDTA), denatured at 74 °C for 15 minutes and loaded onto a prerun 0.8 % formaldehyde agarose gel. Electrophoresis was carried out at 175 V, 4 °C for around 5 hours. RNA was transferred onto nitrocellulose membrane Hybond N+ (Amersham) by capillary transfer overnight and covalently cross-linked to the membrane by 254 nmUV, 120000 µJ/cm^2 (UV Stratalinker 2400). After 30 minutes of prehybridization in hybridization buffer (Ambion® ULTRAhyb®-Oligo, Ambion® ULTRAhyb®, Ambion), denatured 5' end-labeled probe was added and hybridized at 42 °C overnight. The blot was washed according to manufacturer's protocol and visualized by autoradiography.

Probe	Sequence
ConsLNA_D	gaa+tct+gca+agt+gga+tat+ttg
ConsLNA_RC	ca+aat+atc+cac+ttg+cag+att+c

Table 6 Sequence of the probes used in Northern blot analysis."+" stands for the LNA™ modified base (Exiqon).

4.7.3 Reverse RNase Protection Assay

Total content of the transcription reaction in HeLa nuclear extract (see section 1.10.B) with [α-^{32}P] UTP was hybridized to a probe complementary to the input αsatRNA at 42 °C overnight, digested with A/T1 RNases mixture at 37 °C for 30 minutes, and precipitated according to the RPA III Ribonuclease Protection Assay Kit (Ambion) protocol. Samples were analyzed on 8 % polyacrylamide gel (1 x TBE, 8 M urea) and visualized on a phosphoimager screen.

4.8 *In vitro* transcription & RNA purification

In vitro transcription was carried out either from linear PCR fragments or annealed chemically synthesized, complementary oligonucleotides containing promoter sequence specific for T7 or SP6 phage polymerases. Transcription reaction was performed at 37 °C for 4 hours or overnight under following conditions: 0.4 µM template, 5 mM each NTP, 5-25 mM $MgCl_2$, 50 µM DTT, 1 x transcription buffer (40 mM Tris-Cl pH 7.5, 25 mM $MgCl_2$, 3 mM sperimidine) 40 U RNase Inhibitor (NEB), 100 U T7 (NEB) or 40 U SP6 (NEB) RNA polymerase in 100 µl reaction. In order to optimize each reaction conditions, the concentration of $MgCl_2$ and/or template was varied. DNA template was removed by two subsequent DNase I (NEB) digestions at 37 °C for 20 minutes each. Transcribed RNA was purified over 8 % polyacrylamide gel (1 x TBE, 8 M urea). Bands were visualized by UV shadowing. The band of the expected size was excised, crashed and soaked in elution buffer (10 mM Tris-Cl pH 7.5, 2 mM EDTA, 0.3 M NaOAc). RNA was eluted overnight at RT, 1400 rpm, ethanol precipitated, collected by centrifugation and dissolved in TE buffer (10 mM Tris-Cl pH 7.5, 1 mM EDTA) or water.

4.9 RNA oxidation with sodium periodate

In vitro transcribed RNA was dissolved in borax buffer pH 8.6 (4.375 mM borax, 50 mM boric acid) and 0.2 M NaIO4 was added. The reaction was carried out in the dark at RT for 10 minutes. Incubation was repeated for 10 minutes after addition of 2 µl glycerol. Next, the mixture was lyophilized at 45 °C for 40 minutes. RNA pellet was dissolved in borax buffer pH 9.5 (33.75 mM borax, 50 mM boric acid, pH adjusted with NaOH) and incubated at 45 °C for 90 minutes. The reaction was terminated by standard phenol/chloroform extraction and ethanol precipitation. The protocol was modified after (Akbergenov *et al*, 2006).

4.10 *In vitro* experiments in HeLa nuclear extract

4.10.1 Preparation of HeLa nuclear extract

HeLa cells grown in suspension were collected at 4 °C for 15 minutes at 2000 rpm, 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 cells were homogenized in Douncer with 12 strokes with pestle B, nuclei were pelleted at 4 °C for 15 minutes at 2800 rpm; 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 with 6 strokes to disperse clumps. The homogenized suspension was then stirred for 30 minutes at 4 °C. When the suspension became less viscous, it was carefully transferred to centrifuge tubes and spun at 4 °C for 30 minutes at 18000 rpm to remove cell debris. Recovered supernatant was dialyzed to remove salts in 1 L of 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 4 °C for 20 minutes at 17000 x*g* and gently resuspended in dialysis buffer. Dialysis was done at 4 °C overnight in a dialysis cassette with 3000 MW cut-off and repeated with a freshly exchanged buffer for next 5 hours. Insoluble debris was pelleted at 4 °C for 20 minutes at 14000 x*g*, whereas the recovered supernatant was snap-frozen in liquid N2 in aliquots to be stored at -80 °C.

4.10.2 Transcription in HeLa nuclear extract

5-50 nM unlabeled, *in vitro* transcribed RNA or control 50 nM PCR product, were pre-incubated at 30 °C for 10 minutes in reaction mixture of total volume of 50 µl containing: 1 x NTPs mix (0.4 mM ATP, 0.4 mM GTP, 0.4 mM CTP, 0.016 mM UTP), 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 1.7 µM [α-^{32}P] UTP was added to the reaction and incubated at 30 °C for 1 hour. 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 purified by standard phenol/chloroform extraction. Next, transcription products were ethanol precipitated, collected by centrifugation and separated on 8-12 % polyacrylamide gel (1 x TBE, 8 M urea) and visualized on a phosphoimager screen.

Template	Sequence
dsDNA: α111	ccaactaatacgactcactataggcacgacagaagaattctcagtaacttccttgtgttgtgtgtattcaactcacagagttgaacgttccttt
α111	cacgacagaagaattctcagtaacttccttgtgttgtgtgtattcaactcacagagttgaacgttccttt
RCα111	aaaggaacgttcaactctgtgagttgaatacacacaacacaaggaagttactgagaattcttctgtcgtg
dsDNA:5'art-α111	ccaactaatacgactcactataggacggaggggcacggtcacgacagaagaattctcagtaacttccttgtgttgtgtgtattcaactcacagagttgaacgttccttt
5'art-α111	acggaggggcacggtcacgacagaagaattctcagtaacttccttgtgttgtgtgtattcaactcacagagttgaacgttccttt
RC5'art-α111	aaaggaacgttcaactctgtgagttgaatacacacaacacaaggaagttactgagaattcttctgtcgtgaccgtgcccctccgt
dsDNA:3'art-α111	ccaactaatacgactcactataggcacgacagaagaattctcagtaacttccttgtgttgtgtgtattcaactcacagagttgaacgttcctttacggaggggcacggt
3'art-α111	cacgacagaagaattctcagtaacttccttgtgttgtgtgtattcaactcacagagttgaacgttcctttacggaggggcacggt
RC3'art-α111	accgtgcccctccgtaaaggaacgttcaactctgtgagttgaatacacacaacacaaggaagttactgagaattcttctgtcgtg
αUnit	cattctcagaaacttctttgtgatgtatacattcaactcacagagttgaaccttccttttcatagagcagttttgaaacactctttttgtagaatctgcaagtggatatttggaccgctttgaggccttggttggaaacgggaatatcttcatataaaaactagacagaag
RCαUnit	cttctgtctagtttttatatgaagatattcccgtttccaaccaaggcctcaaagcggtccaaatatccacttgcagattctacaaaaagagtgtttcaaaactgctctatgaaaaggaaggttcaactctgtgagttgaatgtatacatcacaaa

	gaagtttctgagaatg
α276	taacagagatgaaccttccttttgacagagcagttttgaa
5’art-α276	acggaggggcacggttaacagagatgaaccttccttttgacagagcagttttgaa
3’art-α276	taacagagatgaaccttccttttgacagagcagttttgaaacggaggggcacggt

Table 7 Sequences of templates used for *in vitro* transcription in HeLa nuclear extract.

5 Literature

Akbergenov R, Si-Ammour A, Blevins T, Amin I, Kutter C, Vanderschuren H, Zhang P, Gruissem W, Meins F, Hohn T & Pooggin MM (2006) Molecular characterization of geminivirus-derived small RNAs in different plant species. *Nucleic Acids Res.* **34:** 462–471

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

Allshire RC, Javerzat JP, Redhead NJ & Cranston G (1994) Position effect variegation at fission yeast centromeres. *Cell* **76:** 157–169

Allshire RC & Karpen GH (2009) Epigenetic regulation of centromeric chromatin: old dogs, new tricks? *Nat. Rev. Genet.* **9:** 923–937

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

Amor DJ, Kalitsis P, Sumer H & Choo KHA (2004) Building the centromere: from foundation proteins to 3D organization. *Trends Cell Biol.* **14:** 359–68

Baker R, Fitzgerald-Hayes M & O'Brien T (1989) Purification of the yeast centromere binding protein CP1 and a mutational analysis of its binding site. *J. Biol. Chem.* **264:** 10843–50

Bannister AJ, Zegerman P, Partridge JF, Miska EA, Thomas JO, Allshire RC & Kouzarides T (2001) Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410:** 120–4

Barnhart MC, Kuich PHJ, Stellfox ME, Ward JA, Bassett EA, Black BE & Foltz DR (2011) HJURP is a CENP-A chromatin assembly factor sufficient to form a functional de novo kinetochore. *J. Cell Biol.* **194:** 229–43

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

Bergmann JH, Martins NMC, Larionov V, Masumoto H & Earnshaw WC (2012) HACking the centromere chromatin code: insights from human artificial chromosomes. *Chromosome Res.* **20:** 505–19

Bergmann JH, Rodríguez MG, Martins NMC, Kimura H, Kelly D a, Masumoto H, Larionov V, Jansen LET & Earnshaw WC (2011) Epigenetic engineering shows H3K4me2 is required for HJURP targeting and CENP-A assembly on a synthetic human kinetochore. *EMBO J.* **30:** 328–40

Bernard P, Maure J, Partridge J, Genier S, Javerzat J & Allshire R (2001) Requirement of heterochromatin for cohesion at centromeres. *Science* **294:** 2539–42

Bernstein E, Kim SY, Carmell MA, Murchison EP, Alcorn H, Li MZ, Mills AA, Elledge SJ, Anderson KV & Hannon GJ (2003) Dicer is essential for mouse development. *Nat. Genet.* **35:** 215–7

Bjerling P, Silverstein RA, Thon G, Caudy A, Grewal S & Ekwall K (2002) Functional divergence between histone deacetylases in fission yeast by distinct cellular localization and *in vivo* specificity. *Mol. Cell. Biol.* **22:** 2170–2181

Black BE, Jansen LE, Maddox PS, Foltz DR, Desai AB, Shah JV & Cleveland DW (2007) Centromere identity maintained by nucleosomes assembled with histone H3 containing the CENP-A targeting domain. *Mol. Cell* **25:** 309–22

Blower MD & Karpen GH (2001) The role of Drosophila CID in kinetochore formation, cell cycle progression and heterochromatin interactios. *Nat. Cell Biol.* **3:** 730–739

Bortolin-Cavaillé ML, 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:** 8709–8714

Britten RJ & Kohne DE (1968) Repeated sequences in DNA. *Science* **161:** 17

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

Brown CJ, Hendrich BD, Rupert JL, Lafrenière RG, Xing Y, Lawrence J & Willard HF (1992) The human XIST gene: Analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71:** 527–542

Bulut-Karslioglu A, Perrera V, Scaranaro M, de la Rosa-Velazquez IA, van de Nobelen S, Shukeir N, Popow J, Gerle B, Opravil S, Pagani M, Meidhof S, Brabletz T, Manke T, Lachner M & Jenuwein T (2012) A transcription factor-based mechanism for mouse heterochromatin formation. *Nat. Struct. Mol. Biol.* **19:** 1023–30

Burrack LS & Berman J (2012) Flexibility of centromere and kinetochore structures. *Trends Genet.* **28:** 204–12

Bushnell DA, Cramer P & Kornberg RD (2002) Structural basis of transcription: alpha-amanitin-RNA polymerase II cocrystal at 2.8 A resolution. *Proc. Natl. Acad. Sci. U. S. A.* **99:** 1218–22

Butter F, Scheibe M, Mörl M & Mann M (2009) Unbiased RNA-protein interaction screen by quantitative proteomics. *Proc. Natl. Acad. Sci. U. S. A.* **106:** 10626–10631

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest ARR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, et al (2005) The transcriptional landscape of the mammalian genome. *Science* **309:** 1559–63

Carone DM, Longo MS, Ferreri GC, Hall L, Harris M, Shook N, Bulazel KV, Carone BR, Obergfell C, O'Neill MJ & O'Neill RJ (2009) A new class of retroviral and satellite encoded small RNAs emanates from mammalian centromeres. *Chromosoma* **118:** 113–25

Carroll CW, Silva MC, Godek KM, Jansen LE & Straight AF (2009) Centromere assembly require the direct recognition of CENP-A nucleosomes by CENP-N. *Nat. Cell Biol.* **11:** 896–902

Castillo AG, Mellone BG, Partridge JF, Richardson W, Hamilton GL, Allshire RC & Pidoux AL (2007) Plasticity of fission yeast CENP-A chromatin driven by relative levels of histone H3 and H4. *PLoS Genet.* **3:** e121

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

Chan GK, Liu ST & Yen TJ (2005) Kinetochore structure and function. *Trends Cell Biol.* **15:** 589–98

Charlesworth B, Sniegowski P & Stephan W (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371:** 215–220

Chen ES, Zhang K, Nicolas E, Cam HP, Zofall M & Grewal SIS (2008) Cell cycle control of centromeric repeat transcription and heterochromatin assembly. *Nature* **451:** 734–7

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, Piccolboni A, Bekiranov S, Bailey DK, Ganesh M, Ghosh S, Bell I, Gerhard DS & Gingeras TR (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308:** 1149–54

Chiodi I, Corioni M, Giordano M, Valgardsdottir R, Ghigna C, Cobianchi F, Xu RM, Riva S & Biamonti G (2004) RNA recognition motif 2 directs the recruitment of SF2/ASF to nuclear stress bodies. *Nucleic Acids Res.* **32:** 4127–36

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

Choo KHA (2001) Domain Organization at the Centromere and Neocentromere. *Dev. Cell* **1:** 165–177

Chueh AC, Northrop EL, Brettingham-Moore KH, Choo KHA & Wong LH (2009) LINE retrotransposon RNA is an essential structural and functional epigenetic component of a core neocentromeric chromatin. *PLoS Genet.* **5:** e1000354

Clemson C, McNeil J, Willard H & Lawrence J (1996) XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *J. Cell Biol.* **132:** 259–75

Cooper JL & Henikoff S (2004) Adaptive evolution of the histone fold domain in centromeric histones. *Mol. Biol. Evol.* **21:** 1712–8

Cramer P, Armache KJ, Baumli S, Benkert S, Brueckner F, Buchen C, Damsma G, Dengl S, Geiger S, Jasiak A, Jawhari A, Jennebach S, Kamenski T, Kettenberger H, Kuhn CD, Lehmann E, Leike K, Sydow J & Vannini A (2008) Structure of eukaryotic RNA polymerases. *Annu. Rev. Biophys.* **37:** 337–52

Cramer P, Bushnell DA & Kornberg RD (2001) Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* **292:** 1863–76

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

Delattre M, Spierer A, Jaquet Y & Spierer P (2004) Increased expression of Drosophila Su(var)3-7 triggers Su(var)3-9-dependent heterochromatin formation. *J. Cell Sci.* **117:** 6239–47

Dezélée S & Sentenac A (1974) Role of Deoxyribonucleic Acid-Ribonucleic Acid Hybrids in Eukaryotes. *J. Biol. Chem.* **249:** 5978–5953

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

Djupedal I, Portoso M, Spåhr H, Bonilla C, Gustafsson CM, Allshire RC & Ekwall K (2005) RNA Pol II subunit Rpb7 promotes centromeric transcription and RNAi-directed chromatin silencing. *Genes Dev.***:** 2301–2306

Doolittle WF (2013) Is junk DNA bunk? A critique of ENCODE. *Proc. Natl. Acad. Sci. U. S. A.* **110:** 5294–300

Doolittle WF & Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284:** 601–603

Dover G (2002) Molecular drive. *Trends Genet.* **18:** 587–589

Drouin G & de Sá MM (1995) The concerted evolution of 5S ribosomal genes linked to the repeat units of other multigene families. *Mol. Biol. Evol.* **12:** 481–93

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

Dunleavy EM, Roche D, Tagami H, Lacoste N, Ray-Gallet D, Nakamura Y, Daigo Y, Nakatani Y & Almouzni-Pettinotti G (2009) HJURP is a cell-cycle-dependent maintenance and deposition factor of CENP-A at centromeres. *Cell* **137:** 485–97

Durfy SJ & Willard HF (1990) Concerted evolution of primate alpha satellite DNA. *J. Mol. Biol.* **216:** 555–566

Earnshaw W, Bordwell B, Marino C & Rothfield N (1986) Three human chromosomal autoantigens are recognized by sera from patients with anti-centromere antibodies. *J. Clin. Invest.* **77:** 426–30

Earnshaw W & Rothfield N (1985) Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. *Chromosoma* **91:** 313–21

Earnshaw WC, Ratrie H & Stetten G (1989) Visualization of centromere proteins CENP-B and CENP-C on a stable dicentric chromosome in cytological spreads. *Chromosoma* **98:** 1–12

Eddy SR (2012) The C-value paradox, junk DNA and ENCODE. *Curr. Biol.* **22:** R898–9

Egloff S & Murphy S (2008) Cracking the RNA polymerase II CTD code. *Trends Genet.* **24:** 280–8

Ekwall K, Olsson T, Turner B, Cranston G & Allshire R (1997) Transient inhibition of histone deacetylation alters the structural and functional imprint at fission yeast centromeres. *Cell* **91:** 1021–32

Ellermeier C, Higuchi EC, Phadnis N, Holm L, Geelhood JL, Thon G & Smith GR (2010) RNAi and heterochromatin repress centromeric meiotic recombination. *Proc. Natl. Acad. Sci. U. S. A.* **107:** 8701–5

Epstein LM & Gall JG (1987) Self-cleaving transcripts of satellite DNA from the newt. *Cell* **48:** 535–543

Espinoza C, Allen T, Hieb A, Kugel JF & Goodrich JA (2004) B2 RNA binds directly to RNA polymerase II to repress transcript synthesis. *Nat. Struct. Mol. Biol.* **11:** 822–9

Eymery A, Callanan M & Vourc'h C (2009a) The secret message of heterochromatin: new insights into the mechanisms and function of centromeric and pericentric repeat sequence transcription. *Int. J. Dev. Biol.* **53:** 259–68

Eymery A, Horard B, El Atifi-Borel M, Fourel G, Berger F, Vitte AL, Van den Broeck A, Brambilla E, Fournier A, Callanan M, Gazzeri S, Khochbin S, Rousseaux S, Gilson E & Vourc'h C (2009b) A transcriptomic analysis of human centromeric and pericentric sequences in normal and tumor cells. *Nucleic Acids Res.* **37:** 6340–54

Eymery A, Souchier C, Vourc'h C & Jolly C (2010) Heat shock factor 1 binds to and transcribes satellite II and III sequences at several pericentromeric regions in heat-shocked cells. *Exp. Cell Res.* **316:** 1845–55

Fachinetti D, Diego Folco H, Nechemia-Arbely Y, Valente LP, Nguyen K, Wong AJ, Zhu Q, Holland AJ, Desai A, Jansen LET & Cleveland DW (2013) A two-step mechanism for epigenetic specification of centromere identity and function. *Nat. Cell Biol.* **15:** 1056–1066

Ferbeyre G, Smith J & Cedergren R (1998) Schistosome satellite DNA encodes active hammerhead ribozymes. *Mol. Cell. Biol.* **18:** 3880–8

Ferri F, Bouzinba-Segard H, Velasco G, Hubé F & Francastel C (2009) Non-coding murine centromeric transcripts associate with and potentiate Aurora B kinase. *Nucleic Acids Res.* **37:** 5071–80

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

Fischer T, Cui B, Dhakshnamoorthy J, Zhou M, Rubin C & Zofall M (2009) Diverse roles of HP1 proteins in heterochromatin assembly and functions in fission yest. *Proc. Natl. Acad. Sci. U. S. A.* **106:** 1–6

Fitzgerald D, Dryden G, Bronson E, Williams J & Anderson J (1994) Conserved patterns of bending in satellite and nucleosome positioning DNA. *J. Biol. Chem.* **269:** 21303–14

Folco HD, Pidoux AL, Urano T & Allshire RC (2008) Heterochromatin and RNAi are required to establish CENP-A chromatin at centromeres. *Science* **319:** 94–7

Foltz DR, Jansen LET, Black BE, Bailey AO, Yates JR & Cleveland DW (2006) The human CENP-A centromeric nucleosome-associated complex. *Nat. Cell Biol.* **8:** 458–69

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

Furuyama S & Biggins S (2007) Centromere identity is specified by a single centromeric nucleosome in budding yeast. *Proc. Natl. Acad. Sci. U. S. A.* **104:** 14706–11

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

Gascoigne KE, Takeuchi K, Suzuki A, Hori T, Fukagawa T & Cheeseman IM (2011) Induced ectopic kinetochore assembly bypasses the requirement for CENP-A nucleosomes. *Cell* **145:** 410–22

Gaubatz J & Cutler R (1990) Mouse satellite DNA is transcribed in senescent cardiac muscle. *J. Biol. Chem.* **265:** 17753–8

Glynn M, Kaczmarczyk A, Prendergast L, Quinn N & Sullivan KF (2010) Centromeres: assembling and propagating epigenetic function. *Subcell. Biochem.* **50:** 223–49

Gopalakrishnan S, Sullivan BA, Trazzi S, Della Valle G & Robertson KD (2009) DNMT3B interacts with constitutive centromere protein CENP-C to modulate DNA methylation and the histone code at centromeric regions. *Hum. Mol. Genet.* **18:** 3178–93

Haaf T, Mater AG, Wienberg J & Ward DC (1995) Presence and abundance of CENP-B box sequences in great ape subsets of primate-specific alpha-satellite DNA. *J. Mol. Evol.* **41:** 487–91

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

Hall LE, Mitchell SE & O'Neill RJ (2012) Pericentric and centromeric transcription: a perfect balance required. *Chromosome Res.* **20:** 535–46

Han JS, Szak ST & Boeke JD (2004) Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* **429:** 268–74

Hasson D, Panchenko T, Salimian KJ, Salman MU, Sekulic N, Alonso A, Warburton PE & Black BE (2013) The octamer is the major form of CENP-A nucleosomes at human centromeres. *Nat. Struct. Mol. Biol.* **20:** 687–95

Henikoff S, Ahmad K & Malik HS (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293:** 1098–102

Holstege F, Fiedler U & Timmers H (1997) Three transitions in the RNA polymerase II transcription complex during initiation. *EMBO J.* **16:** 7468–80

Van Hooser A, Ouspenski I, Gregson H, Starr D, Yen T, Goldberg M, Yokomori K, Earnshaw W, Sullivan K & Brinkley B (2001) Specification of kinetochore-forming chromatin by the histone H3 variant CENP-A. *J. Cell Sci.* **114:** 3529–42

Horard B, Eymery A, Fourel G, Vassetzky N, Puechberty J, Roizes G, Lebrigand K, Barbry P, Laugraud A, Gautier C, Simon E Ben, Devaux F, Magdinier F, Vourc'h C & Gilson E (2009) Global analysis of DNA methylation and transcription of human repetitive sequences. *Epigenetics* **4:** 339–50

Howman E, Fowler K, Newson A, Redward S, MacDonald A, Kalitsis P & Choo K (2000) Early disruption of centromeric chromatin organization in centromere protein A (Cenpa) null mice. *Proc. Natl. Acad. Sci. U. S. A.* **97:** 1148–53

Hudson D, Fowler K, Earle E, Saffery R, Kalitsis P, Trowell H, Hill J, Wreford N, de Kretser D, Cancilla M, Howman E, Hii L, Cutts S, Irvine D & Choo K (1998) Centromere protein B null mice are mitotically and meiotically normal but have lower body and testis weights. *J. Cell Biol.* **141:** 309–19

Hutchinson JN, Ensminger AW, Clemson CM, Lynch CR, Lawrence JB & Chess A (2007) A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* **8:**

Ikeno M, Masumoto H & Okazaki T (1994) Distribution of CENP-B boxes reflected in CREST centromere antigenic sites on long-range alpha-satellite DNA arrays of human chromosome 21. *Hum. Mol. Genet.* **3:** 1245–57

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

Izuta H, Ikeno M, Suzuki N, Tomonaga T, Nozaki N, Obuse C, Kisu Y, Goshima N, Nomura F, Nomura N & Yoda K (2006) Comprehensive analysis of the ICEN (Interphase Centromere Complex) components enriched in the CENP-A chromatin of human cells. *Genes Cells* **11:** 673–84

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

Jason de Koning AP, Gu W, Castoe TA, Batzer MA & Pollock DD (2011) Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* **7:** e1002384

Johnson TL & Chamberlin MJ (1994) Complexes of yeast RNA polymerase II and RNA are substrates for TFIIS-induced RNA cleavage. *Cell* **77:** 217–224

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

Jonstrup AT, Thomsen T, Wang Y, Knudsen BR, Koch J & Andersen AH (2008) Hairpin structures formed by alpha satellite DNA of human centromeres are cleaved by human topoisomerase II. *Nucleic Acids Res.* **36:** 6165–74

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

Karpen GH & Allshire RC (1997) The case for epigenetic effects on centromere identity and function. *Trends Genet.* **13:** 489–96

Kato H, Goto DB, Martienssen R, Urano T, Furukawa K & Murakami Y (2005) RNA polymerase II is required for RNAi-dependent heterochromatin assembly. *Science* **309:** 467–9

Kettenberger H, Eisenführ A, Brueckner F, Theis M, Famulok M & Cramer P (2006) Structure of an RNA polymerase II-RNA inhibitor complex elucidates transcription regulation by noncoding RNAs. *Nat. Struct. Mol. Biol.* **13:** 44–8

Kipling D & Warburton P (1997) Centromeres, CENP-B and Tiger too. *Trends Genet.* **13:** 1–5

Kiss T (2002) Small Nucleolar RNAs: An Abundant Group of Noncoding RNAs with diverse cellular functions. *Cell* **109:** 145–148

Koch J (2000) Neocentromeres and alpha satellite: a proposed structural code for functional human centromere DNA. *Hum. Mol. Genet.* **9:** 149–154

Kwon M, Hori T, Okada M & Fukagawa T (2007) CENP-C Is Involved in Chromosome Segregation, Mitotic Checkpoint Function, and Kinetochore Assembly. *Mol. Biol. Cell* **18:** 2155–2168

Van de Lagemaat LN, Landry J-R, Mager DL & Medstrand P (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.* **19:** 530–6

Lai MMC (2005) RNA Replication without RNA-Dependent RNA Polymerase: Surprises from Hepatitis Delta Virus. *J. Virol.* **79:** 7951–7958

Lam AL, Boivin CD, Bonney CF, Rudd MK & Sullivan BA (2006) Human centromeric chromatin is a dynamic chromosomal domain that can spread over noncentromeric DNA. *Proc. Natl. Acad. Sci. U. S. A.* **103:** 4186–91

Lander E, Linton L, Birren B, Nusbaum C, Zody M, Baldwin J & Devon K (2001) Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921

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

Lee JT (2012) Epigenetic regulation by long noncoding RNAs. *Science* **338:** 1435–9

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

Li YX & Kirby ML (2003) Coordinated and conserved expression of alphoid repeat and alphoid repeat-tagged coding sequences. *Dev. Dyn.* **228:** 72–81

Liao JY, Ma LM, Guo YH, Zhang YC, Zhou H, Shao P, Chen YQ & Qu LH (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

Lippman Z & Martienssen R (2004) The role of RNA interference in heterochromatic silencing. *Nature* **431:** 364–70

Litingtung Y, Lawler A, Sebald S, Lee E, Gearhart J, Westphal H & Corden J (1999) Growth retardation and neonatal lethality in mice with a homozygous deletion in the C-terminal domain of RNA polymerase II. *Mol. Gen. Genet.* **261:** 100–5

Lo AW, Magliano DJ, Sibson MC, Kalitsis P, Craig JM & Choo KHA (2001) A novel chromatin immunoprecipitation and array (CIA) analysis identifies a 460-kb CENP-A-binding neocentromere DNA. *Genome Res.* **11:** 448–457

Lu J & Gilbert DM (2007) Proliferation-dependent and cell cycle regulated transcription of mouse pericentric heterochromatin. *J. Cell Biol.* **179:** 411–21

Maida Y, Yasukawa M, Furuuchi M, Lassmann T, Okamoto N, Kasim V, Hayashizaki Y & Hahn WC (2010) An RNA dependent RNA polymerase formed by hTERT and the RNase MRP RNA. *Nature* **461:** 230–235

Maison C, Bailly D, Peters A, Quivy J-P, Roche D, Taddei A, Lachner M, Jenuwein T & Almouzni G (2002) Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component. *Nat. Genet.* **30:** 329–34

Mariner PD, Walters RD, Espinoza CA, Drullinger LF, Wagner SD, Kugel JF & Goodrich JA (2008) Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. *Mol. Cell* **29:** 499–509

Martianov I, Ramadass A, Serra Barros A, Chow N & Akoulitchev A (2007) Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* **445:** 666–70

Masumoto H, Masukata H, Muro Y, Nozaki N & Okazaki T (1989) A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. *J. Cell Biol.* **109:** 1963–73

Masumoto H, Nakano M & Ohzeki J-I (2004) The role of CENP-B and alpha-satellite DNA: de novo assembly and epigenetic maintenance of human centromeres. *Chromosome Res.***:** 543–56

Mattick JS (2004) RNA regulation: a new genetics? *Nature* **5:** 1662–1666

Mattick JS & Gagen MJ (2005) Accelerating Networks. *Science* **307:** 856–858

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

Metz A, Soret J, Vourc'h C, Tazi J & Jolly C (2004) A key role for stress-induced satellite III transcripts in the relocalization of splicing factors into nuclear stress granules. *J. Cell Sci.* **117:** 4551–8

Miell MD, Fuller CJ, Guse A, Barysz HM, Downes A, Owen-Hughes T, Rappsilber J, Straight AF & Allshire RC (2013) CENP-A confers a reduction in height on octameric nucleosomes. *Nat. Struct. Mol. Biol.* **20:** 763–5

Moazed D (2009) Small RNAs in transcriptional gene silencing and genome defence. *Nature* **457:** 413–20

Motamedi MR, Verdel A, Colmenares SU, Gerber SA, Gygi SP & Moazed D (2004) Two RNAi complexes, RITS and RDRC, physically interact and localize to noncoding centromeric RNAs. *Cell* **119:** 789–802

Mravinac B, Plohl M & Ugarkovic D (2005) Preservation and high sequence conservation of satellite DNAs suggest functional constraints. *J. Mol. Evol.* **61:** 542–50

Murakami K, Elmlund H, Kalisman N, Bushnell DA, Adams CM, Azubel M, Elmlund D, Levi-Kalisman Y, Liu X, Gibbons BJ, Levitt M & Kornberg RD (2013) Architecture of an RNA polymerase II transcription pre-initiation complex. *Science* **342:** 1238724

Murchison EP, Partridge JF, Tam OH, Cheloufi S & Hannon GJ (2005) Characterization of Dicer-deficient murine embryonic stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **102:** 12135–40

Muro Y, Masumoto H, Yoda K, Nozaki N, Ohashi M & Okazaki T (1992) Centromere protein B assembles human centromeric alpha-satellite DNA at the 17-bp sequence, CENP-B box. *J. Cell Biol.* **116:** 585–96

Nagano T, Mitchell JA, Sanz LA & Pauler FM (2008) The Air Noncoding RNA epigenetically silences transcription by targeting G91 to chromatin. *Science* **322:** 1717–1720

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

Nakayama J, Rice J, Strahl B, Allis C & Grewal S (2001) Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* **292:** 110–3

Ohkuni K & Kitagawa K (2011) Endogenous transcription at the centromere facilitates centromere activity in budding yeast. *Curr. Biol.* **21:** 1695–703

Ohno S (1972) So much "junk" DNA in our genome. *Brookhaven Symp. Biol.* **23:** 366–70

Ohzeki J, Nakano M, Okada T & Masumoto H (2002) CENP-B box is required for de novo centromere chromatin assembly on human alphoid DNA. *J. Cell Biol.* **159:** 765–75

Okada M, Cheeseman IM, Hori T, Okawa K, McLeod IX, Yates JR, Desai A & Fukagawa T (2006) The CENP-H-I complex is required for the efficient incorporation of newly synthesized CENP-A into centromeres. *Nat. Cell Biol.* **8:** 446–57

Okada M, Okawa K, Isobe T & Fukagawa T (2009) CENP-H– containing Complex Facilitates Centromere Deposition of CENP-A in Cooperation with FACT and CHD1. *Mol. Biol. Cell* **20:** 3986–3995

Okada T, Ohzeki J, Nakano M, Yoda K, Brinkley WR, Larionov V & Masumoto H (2007) CENP-B controls centromere formation depending on the chromatin context. *Cell* **131:** 1287–300

Orgel LE & Crick FHC (1980) Selfish DNA: the ultimate parasite. *Nature* **284:** 604–607

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

Pall GS & Hamilton AJ (2008) Improved northern blot method for enhanced detection of small RNA. *Nat. Protoc.* **3:** 1077–84

Palmer DK, Day KO, Trongt HL, Charbonneau H & Margolis RL (1991) Purification of the centromere-specific protein CENP-A and demonstration that it is a distinctive histone. *Proc. Natl. Acad. Sci. U. S. A.* **88:** 3734–3738

Partridge J, Borgstrøm B & Allshire R (2000) Distinct protein interaction domains and protein spreading in a complex centromere. *Genes Dev.* **14:** 783–91

Plohl M, Luchetti A, Mestrović N & Mantovani B (2008) Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. *Gene* **409:** 72–82

Polavarapu N, Mariño-Ramírez L, Landsman D, McDonald JF & Jordan IK (2008) Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. *BMC Genomics* **9:**

Prades C, Laurent AM, Puechberty J, Yurov Y & Roizés G (1996) SINE and LINE within human centromeres. *J. Mol. Evol.* **42:** 37–43

Probst AV, Okamoto I, Casanova M, El Marjou F, Le Baccon P & Almouzni G (2010) A strand-specific burst in transcription of pericentric satellites is required for chromocenter formation and early mouse development. *Dev. Cell* **19:** 625–38

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

Puechberty J, Laurent A, Gimenez S, Billault A, Brun-Laurent M, Calenda A, Marçais B, Prades C, Ioannou P, Yurov Y & Roizès G (1999) Genetic and physical analyses of the centromeric and pericentromeric regions of human chromosome 5: recombination across 5cen. *Genomics* **56:** 274–87

Rackwitz H, Rohde W & Sänger H (1981) DNA-dependent RNA polymerase II of plant origin transcribes viroid RNA into full-length copies. *Nature* **291:** 297–301

Regnier V, Vagnarelli P, Fukagawa T, Zerjal T, Burns E, Trouche D, Earnshaw W & Brown W (2005) CENP-A Is Required for Accurate Chromosome Segregation and Sustained Kinetochore Association of BubR1. *Mol. Cell. Biol.* **25:** 3967–3981

Reinhart BJ & Bartel DP (2002) Small RNAs correspond to centromere heterochromatic repeats. *Science* **297:** 1831

Rieder CL (1979) Ribonucleoprotein staining of centrioles and kinetochores in newt lung cell spindles. *J. Cell Biol.* **80:** 1–9

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E & Chang HY (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129:** 1311–23

Rizzi N, Denegri M, Chiodi I, Corioni M, Valgardsdottir R, Cobianchi F, Riva S & Biamonti G (2004) Transcriptional Activation of a Constitutive Heterochromatic Domain of the Human Genome in Response to Heat Shock. *Mol. Biol. Cell* **15:** 543–551

Rojas A, Vazquez-Tello A, Ferbeyre G, Venanzetti F, Bachmann L, Paquin B, Sbordoni V & Cedergren R (2000) Hammerhead-mediated processing of satellite pDo500 family transcripts from Dolichopoda cave crickets. *Nucleic Acids Res.* **28:** 4037–43

Romanova L, Deriagin G, Mashkova T, Tumeneva I, Mushegian A, Kisselev L & Alexandrov I (1996) Evidence for selection in evolution of alpha satellite DNA: the central role of CENP-B/pJ alpha binding region. *J. Mol. Biol.* **261:** 334–40

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

Rudert F, Bronner S, Garnier J & Dollé P (1995) Transcripts from opposite strands of gamma satellite DNA are differentially expressed during mouse development. *Mamm. Genome* **6:** 76–83

Sadaie M, Iida T, Urano T & Nakayama J-I (2004) A chromodomain protein, Chp1, is required for the establishment of heterochromatin in fission yeast. *EMBO J.* **23:** 3825–35

Sadeghi L, Siggens L, Svensson JP & Ekwall K (2014) Centromeric histone H2B monoubiquitination promotes noncoding transcription and chromatin integrity. *Nat. Struct. Mol. Biol.* **21:** 236–43

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

Saffery R, Wong LH, Irvine D V, Bateman M a, Griffiths B, Cutts SM, Cancilla MR, Cendron a C, Stafford a J & Choo KH (2001) Construction of neocentromere-based human minichromosomes by telomere-associated chromosomal truncation. *Proc. Natl. Acad. Sci. U. S. A.* **98:** 5705–10

Sandelin A, Alkema W, Engström P, Wasserman WW & Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32:** D91–4

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

Schmutz J, Wheeler J, Grimwood J, Dickson M, Yang J, Caoile C, Bajorek E, Black S, Chan YM, Denys M, Escobar J, Flowers D, Fotopulos D, Garcia C, Gomez M, Gonzales E, Haydu L, Lopez F, Ramirez L, Retterer J, et al (2004) Quality assessment of the human genome sequence. *Nature* **429:** 365–368

Schueler MG, Dunn JM, Bird CP, Ross MT, Viggiano L, Rocchi M, Willard HF & Green ED (2005) Progressive proximal expansion of the primate X chromosome centromere. *Proc. Natl. Acad. Sci. U. S. A.* **102:** 10563–8

Schueler MG, Higgins AW, Rudd MK, Gustashaw K & Willard HF (2001) Genomic and genetic definition of a functional human centromere. *Science* **294:** 109–15

Schumann GG, Gogvadze E V, Osanai-Futahashi M, Kuroki A, Münk C, Fujiwara H, Ivics Z & Buzdin A a (2010) Unique functions of repetitive transcriptomes. *Int. Rev. Cell Mol. Biol.* **285:** 115–88

Shamovsky I, Ivannikov M, Kandel ES, Gershon D & Nudler E (2006) RNA-mediated response to heat shock in mammalian cells. *Nature* **440:** 556–60

Shapiro JA & von Sternberg R (2005) Why repetitive DNA is essential to genome function. *Biol. Rev. Camb. Philos. Soc.* **80:** 227–50

Shelby RD, Vafa O & Sullivan KF (1997) Assembly of CENP-A into centromeric chromatin requires a cooperative array of nucleosomal DNA contact sites. *J. Cell Biol.* **136:** 501–13

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

Stimpson KM & Sullivan BA (2010) Epigenomics of centromere assembly and function. *Curr. Opin. Cell Biol.* **22:** 772–80

Struhl K (2007) Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.* **14:** 103–5

Sugiyama T, Cam HP, Sugiyama R, Noma K, Zofall M, Kobayashi R & Grewal SI (2007) SHREC, an effector complex for heterochromatic transcriptional silencing. *Cell* **128:** 491–504

Sullivan BA & Schwartz S (1995) Identification of centromeric antigens in dicentric Robertsonian translocations: CENP-C and CENP-E are necessary components of functional centromeres. *Hum. Mol. Genet.* **4:** 2189–97

Sullivan BA & Willard HF (1998) Stable dicentric X chromosomes with two functional centromeres. *Nat. Genet.* **20:** 227–228

Sullivan KF & Glass CA (1991) CENP-B is a highly conserved mammalian centromere protein with homology to the helix-loop-helix family of proteins. *Chromosoma* **100:** 360–70

Sunwoo H, Dinger ME, Wilusz JE, Amaral PP, Mattick JS & Spector DL (2009) MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res.* **19:** 347–59

Sverdlov E (1998) Perpetually mobile footprints of ancient infections in human genome. *FEBS Lett.* **428:** 1–6

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, Pheasant M & Mattick JS (2007) The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* **29:** 288–99

Taft RJ, Simons C, Nahkuri S, Oey H, Korbie DJ, Mercer TR, Holst J, Ritchie W, Wong JJL, 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

Tagarro I, Fernández-Peralta A & González-Aguilera J (1994) Chromosomal localization of human satellites 2 and 3 by a FISH method using oligonucleotides as probes. *Hum. Genet.* **93:** 383–388

Tal M, Shimron F & Yagil G (1994) Unwound regions in yeast centromere IV DNA. *J. Mol. Biol.* **243:** 179–89

Taylor MS, Kai C, Kawai J, Carninci P, Hayashizaki Y & Semple CA (2006) Heterotachy in mammalian promoter evolution. *PLoS Genet.* **2:** e30

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

Thon G & Verhein-Hansen J (2000) Four chromo-domain proteins of *S. pombe* differentially repress transcription at various chromosomal locations. *Genetics* **155:** 551–68

Thorsen M, Hansen H, Venturi M, Holmberg S & Thon G (2012) Mediator regulates non-coding RNA transcription at fission yeast centromeres. *Epigenetics Chromatin* **5:** 19

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* **331:** 593–6

Tittel-Elmer M, Bucher E, Broger L, Mathieu O, Paszkowski J & Vaillant I (2010) Stress-induced activation of heterochromatic transcription. *PLoS Genet.* **6:** e1001175

Tomonaga T, Matsushita K & Yamaguchi S (2003) Overexpression and Mistargeting of Centromere Protein-A in Human Primary Colorectal Cancer. *Cancer Res.* **63:** 3511–3516

Topp CN, Zhong CX & Dawe RK (2004) Centromere-encoded RNAs are integral components of the maize kinetochore. *Proc. Natl. Acad. Sci. U. S. A.* **101:** 15986–91

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

Tyler-Smith C, Gimelli G, Giglio S, Floridia G, Pandya A, Terzoli G, Warburton PE, Earnshaw WC & Zuffardi O (1999) Transmission of a fully functional human neocentromere through three generations. *Am. J. Hum. Genet.* **64:** 1440–4

Ugarkovic D (2008) Satellite DNA Libraries and Centromere Evolution. *Open Evol. J.* **2:** 1–6

Ugarkovic Đ (2005) Functional elements residing within satellite DNAs. *EMBO Rep.* **6:** 1035–9

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

Vafa O & Sullivan K (1997) Chromatin containing CENP-A and alpha-satellite DNA is a major component of the inner kinetochore plate. *Curr. Biol.* **7:** 897–900

Valgardsdottir R, Chiodi I, Giordano M, Cobianchi F, Riva S & Biamonti G (2005) Structural and Functional Characterization of Noncoding Repetitive RNAs Transcribed in Stressed Human Cells. *Mol. Biol. Cell* **16:** 2597–2604

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

Venter JC (2003) A part of the human genome sequence. *Science* **299:** 1183–4

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

Verdel A, Jia S, Gerber S, Sugiyama T, Gygi SP, Grewal SIS & Moazed D (2004) RNAi-mediated targeting of heterochromatin by the RITS complex. *Science* **303:** 672–676

Verdel A, Vavasseur A, Le Gorrec M & Touat-Todeschini L (2009) Common themes in siRNA-mediated epigenetic silencing pathways. *Int. J. Dev. Biol.* **53:** 245–57

Vissel B & Choo KA (1987) Human alpha satellite DNA-consensus sequence and conserved regions. *Nucleic Acids Res.* **15:** 6751–6752

Volpe TA, 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* **297:** 1833–7

Volpe T & Martienssen RA (2011) RNA interference and heterochromatin assembly. *Cold Spring Harb. Perspect. Biol.* **3:** a003731

Vourc'h C & Biamonti G (2011) Transcription of Satellite DNAs in Mammals. *Prog. Mol. Subcell. Biol.* **51:** 95–118

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

Waring M & Britten RJ (1966) Nucleotide sequence repetition: a rapidly reassociating fraction of mouse DNA. *Science* **154:** 791–4

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

Weber MJ (2006) Mammalian small nucleolar RNAs are mobile genetic elements. *PLoS Genet.* **2:** e205

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 JM (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

Willard HF & Waye JS (1987) Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends Genet.* **3:** 192–198

Willingham AT & Gingeras TR (2006) TUF love for “junk” DNA. *Cell* **125:** 1215–20

Wilusz JE, Freier SM & Spector DL (2008) 3’ end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* **135:** 919–32

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

Wong NC, Wong LH, Quach JM, Canham P, Craig JM, Song JZ, Clark SJ & Choo KHA (2006) Permissive transcriptional activity at the centromere through pockets of DNA hypomethylation. *PLoS Genet.* **2:** e17

Zhong CX, Marshall JB, Topp C, Mroczek R, Kato A, Nagaki K, Birchler JA, Jiang J & Dawe RK (2002) Centromeric Retroelements and Satellites Interact with Maize Kinetochore Protein CENH3. *Plant Cell* **14:** 2825–2836

Zhou H, Hu H & Lai M (2010) Non-coding RNAs and their epigenetic regulatory mechanisms. *Biol. Cell* **102:** 645–55

Zhu L, Chou S & Reid B (1996) A single G-to-C change causes human centromere TGGAA repeats to fold back into hairpins. *Proc. Natl. Acad. Sci. U. S. A.* **93:** 12159–64

Zhu Q, Pao GM, Huynh AM, Suh H, Tonnu N, Nederlof PM, Gage FH & Verma IM (2011) BRCA1 tumour suppression occurs via heterochromatin-mediated silencing. *Nature* **477:** 179–84

Appendix

6 Supplementary Tables

id	chromosome	genomic location		sequence
		start	end	
1	chr10	42399314	42399501	agatttgaaacactctttttgtggaattttcaagtggagatttcaatcgctttgaggccaattgtaggaaaggaaatatcttcttataaaaactagacaaaatcattctcagaaactactttgtgatgtgtgtgttcaactcacagagtttaacctttcttttcatagagcagtttggaaaccctct;
2	chr10	42400547	42400685	ctttgaggccaaaggaagaaaaggaaatatcttcgtataaaaactagacagaatcattctcagaaactactttgtgatgtgtgcgttcaactcacagagtttaagctttcttttcatagagcagtttggaaacactct;
3	chr10	42408981	42409269	tctttttgtagaatctgcaagtggatatttggacctctttgaggcctctgttggaaacaggtttcttcatatataagtagacagaagaattctcagaaacttctttgtgatgtgtgcattctactcacagcgttgaaccttcctttcaatagagcagttttgaaacactctttttgtagaatttgcaagtcgagacttaaagcgctttgtggccaatggtagaaaaggaaatatctttgtataaaaactagacagaatcattctcagaaactactttgtgatgtgtgc;
4	chr1	121484730	121484809	catgtaaggctagacagaagaattcccagtaacttccttgtgttgtgtgcattcaactcacagagttgaacgttccctt;
5	chr1	121484901	121485070	gaagaaatcccgtttccaacgaaggccacaagatgtcagaatatccacttacagactttacaaacagagcgtttcctaactgctctatgaacagaaaggttaaactctgtgagttgaacaaacacatcacaacgcagtttgtgggaatgattctgtctagttttgaaac;
6	chr1	121484901	121485087	gtttcaaaactagacagaatcattcccacaaactgcgttgtgatgtgtttgttcaactcacagagtttaacctttctgttcatagagcagttaggaaacgctctgtttgtaaagtctgtaagtggatattctgacatcttgtggccttcgttggaaacgggatttcttcctattctgctagacaga;
7	chr1	121485262	121485414	ttctcagaaactcctttgtgatgtgtgcgttcaactcacagagtttaacctttcttttcatagagcagttaggaaacactctgtttgtaaagtctgcaagtggatattcagacctccttgaggccttcgttggaaacgggatttcttcatat;
8	chr11	48893504	48893626	actctttttgtaggatctgcaagtggatatttgtaccgctttgaggcctttgttggaaattggaatatcttcacataaaaactagacagaagcattctcagaaacttctttgtgatgtgtgc;
9	chr11	55015157	55015283	gaaacactctttttgtagaatctgcaagtggacatttggagcgctttgaggcctatggtgaaaaaggaaatatcttcacataaaaactagacagaagcattctcagaaacttctttgtgatgtttg;
10	chr14	19011964	19012284	cattctcagaaacttctttgtgatgtgtgcattcaactcacagagttgtacccttcttttgatagagcagttttgaaacactctctttgtagaatctgcaagttgacattttgtgtgctttgaggactatggtgaaaaaggaaatatcttagcataaaaactagacagaagcattctcagaaacttctttgtgatgtgtgcagtcaacacacagagttgaagctttatttgacagagcgtttttaaacactctttcagtacaatctgcaagtggacatttagagcgctttgtggccttcgttggaaacgggaatatattc;
11	chr14	19012799	19012969	ttctgtctagtttttatatgaagatatttcctttttcaccataggcctcaaagagctccaaatgtccactggcagatactacaaaaagagtgtttcaaaactgctctatgaaaaggaatgttcaactctgtgagttgaatgcaaacattacaaagaagtttctgagaatg;
12	chr14	19034621	19034747	cacacatcacaaagaagtttctcagaattcttctgtctagtttttatgtgaagatatttccttttccaccacaggcctcaaagcgctccaaatgtccatttgcacattctacaaaaagagtgtttc;
13	chr15	20020527	20020815	ctctttttgtagaatctgcaagtggagatttagagtgctttgtggcctatggtagaaaaggaaataccttcacataaaatgtagacagaagtaatatgagaaaattctttgtgatatgtgcattcatctcacagtgttaaacattgcttttgaatgagcattttgaaactctgttttgtagaatctggaagtgtacatttggagcagtttgaggccaatgtggaaaaggaaatatcttcacataaaaactagacagaagaatactgagaaacttctttgtgatgtgtg;
14	chr15	20020556	20020815	ctctttttgtagaatctgcaagtggagatttagagtgctttgtggcctatggtagaaaaggaaataccttcacataaaatgtagacagaagtaatatgagaaaattctttgtgatatgtgcattcatctcacagtgttaaacattgcttttgaatgagcattttgaaactctgttttgtagaatctggaagtgtacatttggagcagtttgaggccaatgtggaaaaggaaatatcttcacataaaaactagacagaag;
15	chr15	20034261	20034387	gaaacactctttttgtagaatctgcaagtggacatttggagaactttgcggcctatagtggaaaaggaaatatcttcacataaaaactagacagaagaattctgagaaacttctttgtgatgtgtg;
16	chr18	15402789	15402915	aatcactctttttgtagaatgtgcaagtggacatttggagcgctttgcggactatggtagaaaaggaaatatcttcacattaaatctagacagaagcaatctgagaaatttctttgagatgtgtgc;
17	chr18	18516919	18517060	actcagctaagagagtggaacctttctttttacagagcagctttgatacactatttttgtagaatctgcaatttgatattttgattgctttaaagatatcgttggaaacaggaatatcttcatataaaatctagacagaag;
18	chr18	18516919	18517060	actcagctaagagagtggaacctttctttttacagagcagctttgatacactatttttgtagaatctgcaatttgatattttgattgctttaaagatatcgttggaaacaggaatatcttcatataaaatctagacagaag;
19	chr18	18516919	18517060	actcagctaagagagtggaacctttctttttacagagcagctttgatacactatttttgtagaatctgcaatttgatattttgattgctttaaagatatcgttggaaacaggaatatcttcatataaaatctagacagaag;
20	chr18	18516919	18517060	actcagctaagagagtggaacctttctttttacagagcagctttgatacactatttttgtagaatctgcaatttgatattttgattgctttaaagatatcgttggaaacaggaatatcttcatataaaatctagacagaag;
21	chr19	27731908	27732241	ttctgtctagtttttatacgaagatatttcctttctaccactgacctcaaagcggctgaaatctccacttacaaattccacaaaaagagtgtctcaaatctgctctgtgtaaagaaccgttcaactctgtgagttgaatacacacaacacaaggaagttactgagaattcttctgtctagcataatataaagaaatcccgtt

				tccaacgaaggcctcaaagaggtctgaatatccacttgtagactttacaaacagagtgtttcctaactgctctatgaaaagaaagttgaaactctgtgagttgaacgcacacatcacaaagcagtttctg;
22	chr19	27731908	27732241	ttctgtctagtttttatacgaagatatttccttttctaccactgacctcaaagcggctgaaatctccacttacaaattccacaaaaagagtgtctcaaatctgctctgtgtaaagaaccgttcaactctgtgagttgaatacacacaacacaaggaagttactgagaattcttctgtctagcataatataaagaaatcccgtttccaacgaaggcctcaaagaggtctgaatatccacttgtagactttacaaacagagtgtttcctaactgctctatgaaaagaaagttgaaactctgtgagttgaacgcacacatcacaaagcagtttctg;
23	chr19	27735610	27735714	gtttaaaaactagacaaaatcattcccagaaactgcgttgtgatgtgtgcgttcaactcaaaaagtttaacctttcttttcatagagccgtttggaaacactct;
24	chr19	27737478	27737557	catgtaaggctagacagaagaattctcagtaacttccttgtgttgtgtgtattcaactcacagatttgaacgttccttt;
25	chr19	27738784	27739112	agagtgtttccaaatggctctatgaaaagaaaggttaaacaaagtgagttcaacgcacacatcataacgcagtttgtgggaatgatcctgtctagtttttaaacgaagatattcccttttctgccattgaccttaaatcgcttgaaatctccacttgcaaattccacaaaaagcgtgtttcaaatctgctctgtctaaaggaacgttcaactctgtgagttgaatacacacaacacaaggaagttactgagaattcttctgtcgtgccttatatgaagaaatcccgtttccaacgaacgcctcaaggaggtcaaaatatccacttgca;
26	chr19_gl000208_random	47262	47331	ctagacagaagcattctcagaaacttctttgtgatgtttgcattcaactcacggagttgaaccttcctt;
27	chr21	14339069	14339194	gaaacactctttttgtagaatgtgcaagtggatatttggatagttttggggctttcgctggaaacgggaatatcttcacataaaaactagacagaagcattctcagaaacttctttcagatgtgt;
28	chr21	14340724	14340989	tatttggacggctttgtggccttcattggaaatgggaatatcttcacataaaaactagacagcagcattttcagaaactactttgtgatgtgtgcattccactcacagtgttgaagctttcttttgatagagcagcttgaaacactctttttataaaatctgcaagtggatatttggacggttttgaggacttcgttggaaaccgtaatatcttcacataaatattagacagaagcattctcagaaacttctttgtgatgtgtg;
29	chr21	14367680	14367977	gaaacactctttttgtagaatctgcaagtggacctttggaaggctttgaggtctatggtggaagaggaattatcttcgcataaaaactagacacaagcattctcagaaacttccttgtgatgtttgcactcaactcacagagttgaacacacgttttcatggagcagttttgacagattgtttttgtagaatctcctagtggatatttggactgctttgaggccttcgttggaaacgggaatgtcttcacataaaaactagacagatgcattctcagaaacttctttgtgatgtgtg;
30	chr2	132980486	132980735	ggcctatggtggaaaaggaaatatgttcacataaaaactagacagaagtattctcagaaacttctttctgatgtttgcattcaactcacagagttgaacataccttatgataaagcagttttgaaacactcttttagtagaaactgtaagtggatatttggaccgctttgaggccttcgttggaaacgggaatatatttacataaaaattagacagcagcattctcattaacttctttgcgatgtgtac;
31	chr2	132982734	132982827	actagacagaagctttctccgaaacttctttgttctgtgtgcattcagctcacagagttgatcctttcttttgatagagcaggtttgaaacac;
32	chr2	132998503	132998798	aaacactctttttgtagaatctgcaagtggatatttggatagctttgaggctttctttggaaatgggaatatcttcacataaaaactagacagaagcattctcagaaacttctttgtgatgctttcattcaactcacggagctgaacattccttttcatagagcagttttgtaacactctttttgtatatctgcaagtggaaacttggtgaacttagaagtctatggtgaaaaaagaaatatcttcccataaaaactagacagaagaattctcagaaacttctttgtgatgtgtg;
33	chr2	132998600	132998940	attctcagaaacttctttgtgatgctttcattcaactcacggagctgaacattccttttcatagagcagttttgtaacactctttttgtatatctgcaagtggaaacttggtgaacttagaagtctatggtgaaaaaagaaatatcttcccataaaaactagacagaagaattctcagaaacttctttgtgatgtgtgtactcaactcacagatttgaacttttcttttgatagagcagttttgagacactctttttgtacaatctgcaagtggatatttgggtagctttgaggattttgttggaaacgggtatatcttcacataaaaactagaaagaag;
34	chr2	92291455	92291625	cttctgtctagttttcaggggaagatatttcctttttcaccataggcctgaaagcgctccaaatgtccacatccagatactacaaaaagagtgtttcaaacctgctctatgaaagggaatgttcaactctgtgacttgaatgcaaacatcacaaagaagtttctgggaat;
35	chr2	92313758	92313928	attcccagaatcttctttgtgatgtttgcattcaagtcacagagttgaacattccctttcatagagcaggtttgaaacactctttttgtagtatctggatgtggacatttggagcgctttcaggcctatggtgaaaaaggaaatatcttcccctgaaaactagacagaag;
36	chr7	58041295	58041416	cactctttttgtagaatctgcaattggacatttggagtgctttgaggcctatggtggaaaatgtaatatcttcacataaaaactagacagaagacgctgagaaacttctttgtgatgtgtg;
37	chr7	61093893	61094019	aaacactctttttgtagaatctgcaagtggccatttggagagctttgaggcctatggtggaaagggaaatatcttcacatgaaaactagacagaagcatactcagaaacttctttgtgatgtgtgc;
38	chr7	61122771	61122897	aatcactctttttgtagaatacgcatttagatatttggagcgctttgaagacttcattggaatcgcgaataccttcacataaaaactagacagaaccattctcagaaacttctttgagatgtgtgc;
39	chr8	43770437	43770607	attctcagaaacttctttgtgatgtgtgcattcaactcacagatttgagccttccttttggtagaacagtttttgaaacactctttttgtggaatctgcaagtggatatttggagcgctttgaggccttcggtggaaatgggaatatcttcacataaaaactagacagaag;
40	chr8	43770437	43770607	cttctgtctagtttttatgtgaagatattcccatttccaccgaaggcctcaaagcgctccaaatatccacttgcagattccacaaaaagagtgtttcaaaactgttctaccaaaaggaaggctcaaatctgtgagttgaatgcacacatcacaaagaagtttctgagaat;
41	chr8	43774099	43774225	aaacactctttttgtagaatctgcaactggatatttggattactttgaggccttcggtggaaacgggaatatcttcacataaaaactagacagaagcattctcagaaacttctttgtgatgtgtgc;
42	chr8	43816356	43816694	ttctcagaaacttctttgtgatctctgcactccactcagagatttgaaacttccttttgatagagcagttttgaaacactatttttgtaggatttgaaagtgaatatttagagcgttttggagcctatgttggaaaagataatatcttcattcaaaaactacacagaagcattctcagaaactactttgatgtttgcattcaactcacagagttgaacattccttttgatagagcagttttgtaacactttctttgtagaatctgcaagtggatattttgacctctctgaggccttcgttggaaacgggaatttctacgtataaaaactagacagaag;
43	chr8	46885835	46885929	ctagacagaagcattctcagaaacttctttggatgtgtgcattcaactcacagagttgaacctttcttgggatagagcagttttgaaacactct;
44	chr9	66783055	66783322	acatttggagagctttgaggactatggtgggaaaggaaatatcttcatatcaaaactagacagaagcatactcagaaacttctttatgatgtttgcattaaactcacagagttgaactttcattttcatagacca

				gttttgaaacactcttttcatagtatctgcaagtggatatttggactgctttgaggacttcattggaaacgggtataacttcacataaacattagacagaagcattctcagaaacttctttgtgatgtgtgc;
45	chr9	66796476	66796593	ctttttgtagtatgtgcaagtagacatttggagcgcttgaggcctatggtgaaaaaggaaatattttcacataaaaactagacagaagcattctcagaaacttctttgtgatgtgtg;
46	chr9	66797767	66797924	ttctttgtgatgtgtgcattcaactcacagggttgaacaatctttttcatacagcagttttgaatctctcttttgtagaatctccaatggacatttggaatgctttggggccttcattcgaaacgggaatatcttcccataaaaactagacagaag;
47	chr9	66797767	66797924	ttctttgtgatgtgtgcattcaactcacagggttgaacaatctttttcatacagcagttttgaatctctcttttgtagaatctccaatggacatttggaatgctttggggccttcattcgaaacgggaatatcttcccataaaaactagacagaag;
48	chr9	66800824	66800993	attctcagaaacttctttgtgttgtgtgtattcaactcacagagttgaacctttattttgatagagcagattgaaacactctttttgtagaatgtgcaagtggatattgggatagttttgaggctttcgttggaaacgggaatatcttcacataaaaactagacagaa;
49	chr9	66803216	66803386	attgtcagaaacttctttgtgatgtgtgcatttaactcacagagttgaacccttcttttgatagagcagttttgaaacactctctttgtagaatctgcaagttgatatttggacagctttgaggcattcattggaaacgggaatatcttcacataaaatctagacagaag;
50	chr9	66970996	66971292	gaaacactctttttgtagaatctgcaagtggatatttggatagctttgaggatttcgttggaaacgggaatatcttcatataaaatctagacagaagcattctcagaaacatctctgtgatgtttgcattcaagtcacagagttgaacattccctttcatagagcaggtttgaaacactctttttgtagtatctggaagtgcacatttggagcgcattgaggcctaaggtgaaaaaggaaatatcttcccataaaaactagacagaagcattctcagaaacttgtttaggatgtgt;
51	chr9	66971880	66972049	cattctcagaaacctctttgtgatgtgtgtactcaactcacagagtttaacatttcttttgatacaccagtttgaaacagtctttttgtagtatctacaagtggatatttggatagcttggcagctttcattggaaacgggaatatcttcacataaaaactagacagaa;
52	chr9	69976769	69977036	actagacagaagcattctcagaaagttctttgtgatgtgtgcattcaactcacagagttgaacgttgcttttgatggagcagttttgacaaactctttttgtaaaatctgcaattggatatttgaagagctttgaggcctatggtcaaaaaggaaatatcttcacataaaagctacagagaagcattctcagaaacttcttggtgatgtgtgctttcaactcacagaattgaaccttgcttttgatagagcaggtttgaaacactct;
53	chr9	69992442	69992612	attctcagaaacttctctgtgatgtgtacattcaactcacagagttgaaaattctttttcatagagcagatttgaaacactccttttgtagaatctgcaagtggacatttggagcgctctgaggccttcgctcgaaatgggaatatcgtcacataaaaactagacagaag;
54	chr9_gl000199_random	129210	129328	caaacatcacaaagaagtttctgagaatgcttctgtctagattttatatgaagatatcccgtttccaaagaaatcctcaaaggtatccaaatatctacttccagattctacaaaaaga;
55	chr9_gl000199_random	33795	33921	caaacatcacaaagaagtttctgagaatgcttctgtctagcttttatgtgaagatattcccgtttccaacgaaagcctcaaagctatccaaatatccacttgaagattccacaaaaagagtgattc;
56	chr9_gl000199_random	33825	33995	cattctcagaaacttgtttgtcatgtatgtactcaactaacagagttgaacctttcttttgatagagcagttttgaatcactctttttgtggaatcttcaagtggatatttggatagctttgaggctttcgttggaaacgggaatatcttcacataaaagctagacagaa;
57	chrX	61811324	61811393	tagacagaagcattctcggaaacatctttgtgatgtgtgcactcaactcacagagttgaacctttcctt;

Table 8 Collection of αsatRNAs isolated from HeLa cells. Sequences of αsatRNAs isolated via RT-PCRs were confirmed by cloning, sequencing and mapped to the human assembly hg19 via BLAT. Only clones that showed a unique best mapping to α satellite arrays are reported in the table.

id	chromosome	PBE			α satellite			reciprocal orientation
		start	end	strand	start	end	strand	
1	chr1	121353516	121353558	+	121353489	121353655	-	RC
2	chr1	121354223	121354262	+	121354164	121354335	-	RC
3	chr1	121355745	121355790	+	121355682	121355849	-	RC
4	chr1	121355758	121355793	-	121355682	121355849	-	direct
5	chr1	121357097	121357136	+	121357041	121357209	-	RC
6	chr1	121359950	121359990	+	121359930	121360096	-	RC
7	chr1	121381129	121381175	-	121381121	121381290	-	direct
8	chr1	121386288	121386311	-	121386236	121386391	-	direct
9	chr1	121446433	121446479	-	121446425	121446594	-	direct
10	chr1	121450165	121450211	-	121450157	121450326	-	direct
11	chr1	121452031	121452077	-	121452023	121452192	-	direct
12	chr1	121453897	121453943	-	121453889	121454058	-	direct
13	chr1	121463226	121463275	-	121463218	121463387	-	direct
14	chr1	121465091	121465140	-	121465083	121465252	-	direct
15	chr1	121468823	121468872	-	121468815	121468984	-	direct
16	chr1	121474418	121474464	-	121474410	121474579	-	direct
17	chr1	121478839	121478878	+	121478782	121478951	-	RC
18	chr1	121483536	121483569	+	121483528	121483698	-	RC
19	chr10	42397868	42397916	-	42397751	42397888	+	RC
19	chr10	42397868	42397916	-	42397891	42398059	+	RC
20	chr10	42398348	42398379	+	42398230	42398398	+	direct
21	chr10	42399226	42399275	-	42399079	42399246	+	RC
21	chr10	42399226	42399275	-	42399248	42399415	+	RC
22	chr10	42402199	42402236	-	42402128	42402296	+	RC
23	chr10	42525989	42526032	-	42525851	42526019	-	direct
24	chr10	42527270	42527313	-	42527161	42527329	-	direct
25	chr10	42527729	42527762	-	42527672	42527837	-	direct
26	chr10	42527977	42528021	-	42527841	42528006	-	direct
26	chr10	42527977	42528021	-	42528008	42528176	-	direct
27	chr10	42528331	42528359	-	42528178	42528346	-	direct
27	chr10	42528331	42528359	-	42528348	42528516	-	direct
28	chr10	42530214	42530264	-	42530049	42530217	-	direct
28	chr10	42530214	42530264	-	42530221	42530387	-	direct
29	chr10	42530327	42530372	-	42530221	42530387	-	direct
30	chr10	42530405	42530447	-	42530390	42530555	-	direct
31	chr10	42530505	42530538	-	42530390	42530555	-	direct
32	chr10	42530629	42530669	+	42530560	42530726	-	RC
33	chr10	42530697	42530740	-	42530560	42530726	-	direct
33	chr10	42530697	42530740	-	42530728	42530896	-	direct
34	chr10	42531055	42531088	-	42530898	42531066	-	direct
34	chr10	42531055	42531088	-	42531068	42531234	-	direct
35	chr10	42532728	42532775	-	42532592	42532760	-	direct
35	chr10	42532728	42532775	-	42532762	42532930	-	direct
36	chr10	42535788	42535829	-	42535649	42535816	-	direct
36	chr10	42535788	42535829	-	42535818	42535986	-	direct
37	chr10	42535956	42535996	+	42535818	42535986	-	RC
37	chr10	42535956	42535996	+	42535988	42536146	-	RC
38	chr10	42537656	42537693	+	42537515	42537683	-	RC
38	chr10	42537656	42537693	+	42537685	42537843	-	RC
39	chr10	42539860	42539905	-	42539723	42539890	-	direct

39	chr10	42539860	42539905	-	42539892	42540060	-	direct
40	chr10	42542162	42542192	-	42542101	42542266	-	direct
41	chr10	42542923	42542962	-	42542781	42542948	-	direct
41	chr10	42542923	42542962	-	42542950	42543118	-	direct
42	chr10	42818031	42818080	-	42817900	42818068	+	RC
42	chr10	42818031	42818080	-	42818070	42818240	+	RC
43	chr11	48811858	48811905	-	48811727	48811895	-	direct
43	chr11	48811858	48811905	-	48811896	48812048	-	direct
44	chr11	48827630	48827681	+	48827484	48827653	+	direct
44	chr11	48827630	48827681	+	48827654	48827824	+	direct
45	chr11	48858791	48858834	+	48858640	48858808	+	direct
45	chr11	48858791	48858834	+	48858809	48858978	+	direct
46	chr11	48859358	48859408	+	48859320	48859489	+	direct
47	chr11	48887780	48887818	+	48887621	48887791	+	direct
47	chr11	48887780	48887818	+	48887792	48887962	+	direct
48	chr11	48895888	48895922	+	48895815	48895984	+	direct
49	chr11	48942367	48942391	+	48942321	48942489	-	RC
50	chr11	50718012	50718045	+	50717959	50718119	-	RC
51	chr11	50762433	50762459	-	50762344	50762517	+	RC
52	chr11	50765482	50765524	+	50765419	50765588	+	direct
53	chr11	50768736	50768778	+	50768586	50768756	+	direct
53	chr11	50768736	50768778	+	50768757	50768922	+	direct
54	chr11	50778569	50778604	+	50778536	50778702	+	direct
55	chr11	51129933	51129973	-	51129970	51130073	-	direct
56	chr11	51571379	51571413	+	51571343	51571512	-	RC
57	chr11	51572130	51572168	-	51572027	51572196	-	direct
58	chr11	51579453	51579496	-	51579315	51579482	-	direct
58	chr11	51579453	51579496	-	51579483	51579653	-	direct
59	chr11	51580848	51580874	-	51580846	51581015	-	direct
60	chr11	51582884	51582916	-	51582717	51582887	-	direct
60	chr11	51582884	51582916	-	51582888	51583056	-	direct
61	chr11	51584935	51584962	-	51584934	51585101	-	direct
62	chr12	34597982	34598010	-	34597932	34598091	-	direct
63	chr12	34847469	34847497	+	34847428	34847592	+	direct
64	chr12	34853865	34853909	+	34853706	34853874	+	direct
64	chr12	34853865	34853909	+	34853876	34854043	+	direct
65	chr12	37996258	37996289	-	37996090	37996258	+	RC
65	chr12	37996258	37996289	-	37996260	37996430	+	RC
66	chr12	38035100	38035150	-	38034965	38035135	-	direct
66	chr12	38035100	38035150	-	38035136	38035305	-	direct
67	chr14	19005180	19005218	-	19005129	19005299	-	direct
68	chr14	19007521	19007565	-	19007517	19007685	-	direct
69	chr14	19008920	19008967	-	19008880	19009048	-	direct
70	chr14	19024042	19024074	-	19023893	19024062	-	direct
70	chr14	19024042	19024074	-	19024064	19024234	-	direct
71	chr14	19035144	19035170	+	19034992	19035161	-	RC
71	chr14	19035144	19035170	+	19035162	19035324	-	RC
72	chr14	19042628	19042661	-	19042493	19042663	-	direct
73	chr16	33973731	33973773	+	33973713	33973881	+	direct
74	chr16	35235207	35235246	+	35235052	35235218	+	direct
74	chr16	35235207	35235246	+	35235222	35235392	+	direct
75	chr16	35260253	35260284	-	35260123	35260293	+	RC

76	chr17	22248229	22248265	-	22248100	22248270	+	RC
77	chr17	22249730	22249775	+	22249625	22249794	+	direct
78	chr17	22250438	22250464	-	22250308	22250476	+	RC
79	chr17	22250709	22250753	-	22250647	22250814	+	RC
80	chr17	22253179	22253217	-	22253026	22253193	+	RC
80	chr17	22253179	22253217	-	22253194	22253363	+	RC
81	chr17	22256879	22256903	-	22256759	22256928	+	RC
82	chr18	15401334	15401382	+	15401177	15401346	+	direct
82	chr18	15401334	15401382	+	15401348	15401516	+	direct
83	chr18	18511660	18511700	+	18511627	18511797	+	direct
84	chr18	18513755	18513780	-	18513672	18513841	+	RC
85	chr18	18514678	18514727	+	18514675	18514841	+	direct
86	chr18	18515672	18515710	+	18515522	18515691	+	direct
86	chr18	18515672	18515710	+	18515693	18515863	+	direct
87	chr18	18517304	18517344	-	18517232	18517396	+	RC
88	chr18	18518394	18518417	+	18518343	18518513	-	RC
89	chr18	18518913	18518948	+	18518856	18519026	-	RC
90	chr18	18519059	18519096	+	18519027	18519193	-	RC
91	chr18	18519990	18520033	-	18519878	18520048	-	direct
92	chr18	18520280	18520306	-	18520220	18520342	-	direct
93	chr19	24474665	24474693	+	24474665	24474832	+	direct
94	chr19	24526166	24526217	-	24526130	24526300	+	RC
95	chr19	24574360	24574391	-	24574285	24574455	+	RC
96	chr19	24615177	24615204	-	24615061	24615230	+	RC
97	chr19	27732228	27732276	-	27732073	27732241	+	RC
97	chr19	27732228	27732276	-	27732244	27732411	+	RC
98	chr19	27733067	27733113	+	27732910	27733078	+	direct
98	chr19	27733067	27733113	+	27733080	27733248	+	direct
99	chr19	27733234	27733279	-	27733080	27733248	+	RC
99	chr19	27733234	27733279	-	27733250	27733419	+	RC
100	chr19	27733399	27733447	+	27733250	27733419	+	direct
100	chr19	27733399	27733447	+	27733421	27733591	+	direct
101	chr19	27733613	27733654	+	27733597	27733760	+	direct
102	chr19	27733847	27733892	-	27733763	27733930	+	RC
103	chr19	27733853	27733889	+	27733763	27733930	+	direct
104	chr19	27734422	27734465	-	27734271	27734439	+	RC
104	chr19	27734422	27734465	-	27734441	27734609	+	RC
105	chr19	27735474	27735517	+	27735461	27735628	+	direct
106	chr19	27736462	27736508	-	27736311	27736479	+	RC
106	chr19	27736462	27736508	-	27736481	27736648	+	RC
107	chr19	27736468	27736514	+	27736311	27736479	+	direct
107	chr19	27736468	27736514	+	27736481	27736648	+	direct
108	chr19	27736904	27736938	+	27736822	27736989	+	direct
109	chr19	27737237	27737286	+	27737161	27737328	+	direct
110	chr19	27737967	27737999	-	27737839	27738007	+	RC
111	chr19	27738841	27738887	-	27738689	27738857	+	RC
111	chr19	27738841	27738887	-	27738859	27739025	+	RC
112	chr19	27739269	27739320	+	27739199	27739367	+	direct
113	chr19	27740023	27740056	-	27739875	27740043	+	RC
113	chr19	27740023	27740056	-	27740045	27740213	+	RC
114	chr19	27740198	27740242	-	27740045	27740213	+	RC
114	chr19	27740198	27740242	-	27740215	27740383	+	RC

115	chr19	27859285	27859324	+	27859172	27859341	-	RC
116	chr19	27870306	27870334	+	27870139	27870308	+	direct
116	chr19	27870306	27870334	+	27870311	27870478	+	direct
117	chr19	27989268	27989292	+	27989196	27989366	+	direct
118	chr19_gl000208_random	40236	40264	+	40185	40355	-	RC
119	chr19_gl000208_random	80683	80729	-	80571	80741	-	direct
120	chr19_gl000208_random	83233	83277	-	83126	83294	-	direct
121	chr2	92273403	92273436	+	92273366	92273535	+	direct
122	chr2	92281255	92281302	+	92281229	92281399	+	direct
123	chr2	92290959	92291005	+	92290947	92291116	+	direct
124	chr2	92297864	92297906	+	92297759	92297924	+	direct
125	chr2	92306084	92306127	+	92305949	92306115	+	direct
125	chr2	92306084	92306127	+	92306116	92306286	+	direct
126	chr2	92306216	92306255	-	92306116	92306286	+	RC
127	chr2	92306703	92306729	+	92306629	92306794	+	direct
128	chr2	92307184	92307232	+	92307137	92307307	+	direct
129	chr2	92313063	92313099	+	92312911	92313077	+	direct
129	chr2	92313063	92313099	+	92313078	92313248	+	direct
130	chr2	92315855	92315880	+	92315798	92315964	+	direct
131	chr2	92317991	92318020	+	92317835	92318000	+	direct
131	chr2	92317991	92318020	+	92318001	92318171	+	direct
132	chr2	92318075	92318116	-	92318001	92318171	+	RC
133	chr2	92318180	92318230	+	92318172	92318342	+	direct
134	chr2	92318416	92318458	-	92318343	92318513	+	RC
135	chr2	92318661	92318697	-	92318514	92318680	+	RC
135	chr2	92318661	92318697	-	92318681	92318851	+	RC
136	chr2	92319345	92319375	+	92319195	92319360	+	direct
136	chr2	92319345	92319375	+	92319361	92319531	+	direct
137	chr2	92321723	92321769	-	92321577	92321742	+	RC
137	chr2	92321723	92321769	-	92321743	92321913	+	RC
138	chr2	92321819	92321863	-	92321743	92321913	+	RC
139	chr2	132985117	132985162	+	132984973	132985143	+	direct
139	chr2	132985117	132985162	+	132985144	132985314	+	direct
140	chr2	132997504	132997540	-	132997406	132997575	+	RC
141	chr2	132999642	132999675	-	132999622	132999793	+	RC
142	chr20	26264528	26264564	-	26264464	26264633	+	RC
143	chr20	26269532	26269565	-	26269372	26269542	+	RC
143	chr20	26269532	26269565	-	26269543	26269713	+	RC
144	chr20	26286275	26286313	-	26286200	26286366	+	RC
145	chr21	10757972	10758002	-	10757906	10758073	-	direct
146	chr21	14359844	14359871	-	14359830	14359995	-	direct
147	chr21	14365294	14365344	-	14365150	14365320	-	direct
147	chr21	14365294	14365344	-	14365321	14365491	-	direct
148	chr3	90354498	90354535	-	90354389	90354557	-	direct
149	chr3	90454492	90454535	-	90454345	90454514	-	direct
149	chr3	90454492	90454535	-	90454515	90454685	-	direct
150	chr3	90457729	90457767	-	90457588	90457758	-	direct
150	chr3	90457729	90457767	-	90457759	90457929	-	direct
151	chr3	90467398	90467431	-	90467252	90467421	-	direct
151	chr3	90467398	90467431	-	90467425	90467592	-	direct
152	chr3	90467399	90467430	+	90467252	90467421	-	RC

152	chr3	90467399	90467430	+	90467425	90467592	-	RC
153	chr3	90474645	90474681	+	90474608	90474777	-	RC
154	chr3	90477698	90477743	-	90477561	90477726	-	direct
154	chr3	90477698	90477743	-	90477728	90477898	-	direct
155	chr4	52682090	52682116	-	52682091	52682256	-	direct
156	chr4	68264315	68264348	-	68264305	68264473	-	direct
157	chr4	68264343	68264382	+	68264305	68264473	-	RC
158	chr4	68264665	68264696	+	68264645	68264813	-	RC
159	chr4	68265210	68265246	-	68265154	68265322	-	direct
160	chr4	68265296	68265333	+	68265154	68265322	-	RC
160	chr4	68265296	68265333	+	68265324	68265492	-	RC
161	chr4	68265643	68265679	+	68265494	68265662	-	RC
161	chr4	68265643	68265679	+	68265664	68265832	-	RC
162	chr4	68265644	68265673	-	68265494	68265662	-	direct
162	chr4	68265644	68265673	-	68265664	68265832	-	direct
163	chr5	46357971	46358012	-	46357827	46357997	-	direct
163	chr5	46357971	46358012	-	46357998	46358167	-	direct
164	chr5	46366557	46366590	+	46366536	46366706	+	direct
165	chr5	46377517	46377545	+	46377522	46377691	+	direct
166	chr5	46392004	46392042	+	46391985	46392153	+	direct
167	chr5	49411728	49411778	+	49411681	49411851	-	RC
168	chr5	49533207	49533248	+	49533060	49533229	-	RC
168	chr5	49533207	49533248	+	49533230	49533399	-	RC
169	chr5	49534901	49534945	-	49534765	49534931	-	direct
169	chr5	49534901	49534945	-	49534933	49535102	-	direct
170	chr5	49535957	49535979	-	49535956	49536125	-	direct
171	chr6	58773643	58773693	-	58773613	58773708	+	RC
172	chr6	58774374	58774415	-	58774222	58774389	+	RC
172	chr6	58774374	58774415	-	58774391	58774561	+	RC
173	chr6	58776446	58776484	+	58776427	58776596	+	direct
174	chr6	58776549	58776592	+	58776427	58776596	+	direct
175	chr6	58776627	58776655	+	58776598	58776765	+	direct
176	chr6	58776746	58776782	+	58776598	58776765	+	direct
176	chr6	58776746	58776782	+	58776767	58776934	+	direct
177	chr6	58777638	58777679	-	58777616	58777784	+	RC
178	chr6	58778248	58778273	-	58778128	58778296	+	RC
179	chr6	58778954	58779003	+	58778808	58778976	+	direct
179	chr6	58778954	58779003	+	58778978	58779144	+	direct
180	chr6	58779011	58779054	-	58778978	58779144	+	RC
181	chr6	58779231	58779264	-	58779145	58779313	+	RC
182	chr6	58779410	58779450	-	58779315	58779482	+	RC
183	chr6	58779684	58779712	+	58779655	58779822	+	direct
184	chr6	61905226	61905272	+	61905088	61905258	-	RC
184	chr6	61905226	61905272	+	61905259	61905429	-	RC
185	chr6	61905226	61905273	-	61905088	61905258	-	direct
185	chr6	61905226	61905273	-	61905259	61905429	-	direct
186	chr6	61917736	61917770	+	61917603	61917774	-	RC
187	chr7	58008628	58008661	-	58008560	58008729	-	direct
188	chr7	58013857	58013904	-	58013778	58013948	-	direct
189	chr7	58021540	58021572	-	58021458	58021628	-	direct
190	chr7	58037807	58037840	+	58037736	58037905	-	RC
191	chr7	58050336	58050365	+	58050293	58050461	-	RC

192	chr7	61097551	61097587	+	61097406	61097576	+	direct
192	chr7	61097551	61097587	+	61097577	61097748	+	direct
193	chr7	61276584	61276612	-	61276426	61276596	+	RC
193	chr7	61276584	61276612	-	61276598	61276766	+	RC
194	chr7	61546378	61546414	+	61546316	61546485	-	RC
195	chr7	61638379	61638414	+	61638317	61638487	-	RC
196	chr7	61650819	61650867	+	61650759	61650929	-	RC
197	chr7	61847570	61847608	+	61847431	61847602	-	RC
198	chr7	61967424	61967467	-	61967330	61967498	+	RC
199	chr7	61967574	61967619	+	61967500	61967668	+	direct
200	chr7	61967576	61967624	-	61967500	61967668	+	RC
201	chr7	61968242	61968287	-	61968180	61968347	+	RC
202	chr7	61968334	61968377	+	61968180	61968347	+	direct
202	chr7	61968334	61968377	+	61968349	61968517	+	direct
203	chr7	61968442	61968473	-	61968349	61968517	+	RC
204	chr7	61968591	61968626	-	61968519	61968687	+	RC
205	chr7	61968783	61968814	-	61968689	61968857	+	RC
206	chr7	61969025	61969054	-	61968859	61969027	+	RC
206	chr7	61969025	61969054	-	61969029	61969197	+	RC
207	chr7	61969840	61969864	+	61969711	61969879	+	direct
208	chr7	61970233	61970279	+	61970221	61970389	+	direct
209	chr7	61970496	61970544	+	61970391	61970559	+	direct
210	chr7	61970907	61970956	+	61970896	61971060	+	direct
211	chr7	61971223	61971271	+	61971069	61971236	+	direct
211	chr7	61971223	61971271	+	61971238	61971406	+	direct
212	chr7	61972078	61972116	-	61971918	61972086	+	RC
212	chr7	61972078	61972116	-	61972089	61972256	+	RC
213	chr7	61972747	61972795	+	61972597	61972765	+	direct
213	chr7	61972747	61972795	+	61972767	61972934	+	direct
214	chr7	61973054	61973094	+	61972936	61973103	+	direct
215	chr7	61973436	61973465	+	61973277	61973444	+	direct
215	chr7	61973436	61973465	+	61973446	61973614	+	direct
216	chr7	61975470	61975505	-	61975318	61975478	+	RC
216	chr7	61975470	61975505	-	61975481	61975648	+	RC
217	chr7	61980512	61980549	-	61980404	61980572	+	RC
218	chr7	61980837	61980862	-	61980744	61980912	+	RC
219	chr7	61984918	61984961	-	61984826	61984995	+	RC
220	chr7	61986112	61986148	-	61986018	61986185	+	RC
221	chr7	61989413	61989443	-	61989254	61989423	+	RC
221	chr7	61989413	61989443	-	61989424	61989591	+	RC
222	chr7	61994011	61994039	+	61993855	61994027	+	direct
222	chr7	61994011	61994039	+	61994029	61994197	+	direct
223	chr7	62254717	62254766	+	62254705	62254873	+	direct
224	chr7	62365869	62365911	-	62365850	62365947	+	RC
225	chr8	43546965	43547010	-	43546889	43547046	+	RC
226	chr8	43794757	43794795	-	43794606	43794772	+	RC
226	chr8	43794757	43794795	-	43794774	43794940	+	RC
227	chr8	43822077	43822118	-	43821977	43822143	+	RC
228	chr8	43822094	43822118	+	43821977	43822143	+	direct
229	chr8	43824281	43824317	-	43824189	43824353	+	RC
230	chr8	43825527	43825577	+	43825383	43825544	+	direct
230	chr8	43825527	43825577	+	43825546	43825716	+	direct

231	chr8	43825618	43825661	-	43825546	43825716	+	RC
232	chr8	43826149	43826189	-	43826057	43826221	+	RC
233	chr8	43827409	43827446	+	43827251	43827411	+	direct
233	chr8	43827409	43827446	+	43827414	43827584	+	direct
234	chr8	43830161	43830187	+	43830129	43830297	+	direct
235	chr8	43831131	43831178	+	43830986	43831147	+	direct
235	chr8	43831131	43831178	+	43831149	43831319	+	direct
236	chr8	43831221	43831264	-	43831149	43831319	+	RC
237	chr8	43831583	43831619	-	43831487	43831657	+	RC
238	chr8	43831752	43831794	-	43831660	43831824	+	RC
239	chr8	43833008	43833040	+	43832854	43833014	+	direct
239	chr8	43833008	43833040	+	43833017	43833187	+	direct
240	chr8	43833089	43833132	-	43833017	43833187	+	RC
241	chr8	43833620	43833662	-	43833528	43833692	+	RC
242	chr8	43836735	43836773	+	43836590	43836751	+	direct
242	chr8	43836735	43836773	+	43836753	43836923	+	direct
243	chr8	43837081	43837111	-	43836924	43837090	+	RC
243	chr8	43837081	43837111	-	43837091	43837261	+	RC
244	chr8	43837356	43837392	-	43837264	43837428	+	RC
245	chr8	43838021	43838060	-	43837942	43838113	+	RC
246	chr8	43838602	43838648	+	43838459	43838620	+	direct
246	chr8	43838602	43838648	+	43838622	43838792	+	direct
247	chr8	46843635	46843658	-	46843531	46843697	+	RC
248	chr8	46846114	46846157	+	46846077	46846245	+	direct
249	chr8	46847076	46847126	+	46846933	46847095	+	direct
249	chr8	46847076	46847126	+	46847099	46847267	+	direct
250	chr8	46847756	46847805	+	46847608	46847769	+	direct
250	chr8	46847756	46847805	+	46847773	46847942	+	direct
251	chr8	46847982	46848027	+	46847944	46848114	+	direct
252	chr8	46848949	46848992	+	46848802	46848964	+	direct
252	chr8	46848949	46848992	+	46848966	46849136	+	direct
253	chr8	46850000	46850043	+	46849984	46850154	+	direct
254	chr8	46852685	46852734	+	46852539	46852701	+	direct
254	chr8	46852685	46852734	+	46852703	46852873	+	direct
255	chr8	46852773	46852814	+	46852703	46852873	+	direct
256	chr8	46853030	46853072	+	46852874	46853040	+	direct
256	chr8	46853030	46853072	+	46853041	46853211	+	direct
257	chr8	46857190	46857219	-	46857120	46857285	+	RC
258	chr8	46857441	46857485	-	46857286	46857452	+	RC
258	chr8	46857441	46857485	-	46857453	46857620	+	RC
259	chr8	47376030	47376055	+	47376007	47376168	-	RC
260	chr9	66824584	66824615	+	66824547	66824716	+	direct
261	chr9	66971314	66971358	-	66971196	66971366	-	direct
262	chr9	66985013	66985044	-	66984845	66985015	-	direct
262	chr9	66985013	66985044	-	66985016	66985185	-	direct
263	chr9	69964306	69964342	+	69964146	69964316	+	direct
263	chr9	69964306	69964342	+	69964317	69964487	+	direct
264	chr9_gl000199_random	43998	44032	+	43853	44023	-	RC
264	chr9_gl000199_random	43998	44032	+	44024	44193	-	RC
265	chr9_gl000199_random	47366	47411	-	47255	47421	-	direct
266	chr9_gl000199_r	76161	76211	-	76006	76172	-	direct

	andom							
266	chr9_gl000199_random	76161	76211	-	76173	76343	-	direct
267	chr9_gl000199_random	108156	108203	+	108146	108312	-	RC
268	chr9_gl000199_random	110248	110296	-	110185	110355	-	direct
269	chr9_gl000199_random	113394	113436	+	113249	113419	-	RC
269	chr9_gl000199_random	113394	113436	+	113420	113586	-	RC
270	chr9_gl000199_random	125081	125122	-	124987	125155	-	direct
271	chr9_gl000199_random	138106	138155	+	138083	138251	-	RC
272	chr9_gl000199_random	153874	153918	+	153727	153897	-	RC
272	chr9_gl000199_random	153874	153918	+	153898	154064	-	RC
273	chr9_gl000199_random	153877	153920	-	153727	153897	-	direct
273	chr9_gl000199_random	153877	153920	-	153898	154064	-	direct
274	chr9_gl000199_random	158359	158383	+	158320	158486	-	RC
275	chr9_gl000199_random	159477	159500	-	159339	159509	-	direct
276	chr9_gl000199_random	161298	161335	-	161205	161374	-	direct
277	chr9_gl000199_random	165197	165228	+	165119	165289	-	RC
278	chrUn_gl000226	200	231	-	124	293	+	RC
279	chrUn_gl000226	556	581	-	466	636	+	RC
280	chrUn_gl000226	1127	1177	+	1146	1316	+	direct
280	chrUn_gl000226	1127	1177	+	979	1145	+	direct
281	chrUn_gl000226	1619	1654	+	1488	1657	+	direct
282	chrUn_gl000226	1917	1951	-	1830	2000	+	RC
283	chrUn_gl000226	2477	2499	-	2343	2508	+	RC
284	chrUn_gl000226	2978	3021	+	2852	3021	+	direct
285	chrUn_gl000226	3863	3903	+	3707	3873	+	direct
285	chrUn_gl000226	3863	3903	+	3874	4044	+	direct
286	chrUn_gl000226	4353	4387	+	4216	4385	+	direct
286	chrUn_gl000226	4353	4387	+	4387	4557	+	direct
287	chrUn_gl000226	4420	4446	+	4387	4557	+	direct
288	chrUn_gl000226	4648	4671	-	4558	4728	+	RC
289	chrUn_gl000226	5223	5258	-	5071	5237	+	RC
289	chrUn_gl000226	5223	5258	-	5238	5408	+	RC
290	chrUn_gl000226	5667	5690	-	5580	5749	+	RC
291	chrUn_gl000226	7378	7424	-	7286	7456	+	RC
292	chrUn_gl000226	8740	8763	-	8650	8820	+	RC
293	chrUn_gl000226	9805	9839	+	9672	9841	+	direct
294	chrUn_gl000226	11177	11209	+	11036	11205	+	direct
294	chrUn_gl000226	11177	11209	+	11207	11377	+	direct
295	chrUn_gl000226	11468	11491	-	11378	11548	+	RC
296	chrUn_gl000226	12832	12875	-	12742	12912	+	RC
297	chrUn_gl000226	13277	13320	+	13255	13421	+	direct
298	chrUn_gl000226	13907	13929	+	13764	13933	+	direct
299	chrUn_gl000226	14186	14222	-	14106	14276	+	RC
300	chrUn_gl000226	14760	14807	+	14619	14785	+	direct
300	chrUn_gl000226	14760	14807	+	14786	14956	+	direct
301	chrX	58561382	58561415	-	58561379	58561549	-	direct

302	chrX	61685395	61685436	+	61685250	61685420	-	RC
302	chrX	61685395	61685436	+	61685422	61685589	-	RC
303	chrX	61694296	61694325	-	61694157	61694327	-	direct
304	chrX	61694694	61694745	+	61694670	61694836	-	RC
305	chrX	61708879	61708924	-	61708725	61708895	-	direct
305	chrX	61708879	61708924	-	61708896	61709062	-	direct
306	chrX	61717099	61717148	-	61716953	61717123	-	direct
306	chrX	61717099	61717148	-	61717124	61717290	-	direct
307	chrX	61719173	61719209	-	61719010	61719180	-	direct
307	chrX	61719173	61719209	-	61719181	61719346	-	direct
308	chrX	61760274	61760313	-	61760173	61760341	+	RC
309	chrX	61800909	61800955	+	61800903	61801073	+	direct
310	chrX	61843491	61843524	+	61843452	61843620	+	direct
311	chrX	61845327	61845353	+	61845161	61845331	+	direct
311	chrX	61845327	61845353	+	61845332	61845502	+	direct
312	chrX	61847378	61847405	-	61847296	61847464	+	RC
313	chrX	61868432	61868480	-	61868367	61868536	+	RC
314	chrY	9914984	9915008	+	9914848	9915006	+	direct

Table 9 Genomic coordinates of α satellites-derived aptamers isolated in Genomic SELEX against RNA polymerase II

Genomic coordinates of PBEs overlapping with DFAM annotated α satellites. Aptamers αsatPBE #111 and αsatPBE #276 are highlighted in grey.

7 Acknowledgments

Writing this dissertation was probably the most daunting task during my whole PhD education. Being at the bench and conducting experiments on αsatRNAs was very exciting and challenging. I was driven by curiosity and thrill from one experiment to another and I was attracted to various different ideas.

First and foremost, I would like to thank Renée Schroeder for guiding me during my graduate education. She introduced me to the RNA world, required to be independent, encouraged to scout what interests me and allowed wondering around to finally realize what is important and where my goals really are. I would also like to emphasize that besides the scientific support I have received lots of emotional support and warmth during my mourning time.

I would also like to thank Christina Waldsich. I really appreciate your kind support, willingness to help with RNA techniques and discussions over the αsatRNAs project.

I would like to acknowledge both former and current, lab colleagues, in no significant order: Rike, Stefan, Christina, Lukas, Sabine, Oliver, Doris and Jen, Bob, Ivana, Boris, Nadia, Meghan, Martina, Johanna, Adam, Max, Hakim for composing an exceptional, multicoloured team and creating helpful and friendly work environment. I would like to spot Rike, who supervised me at the very beginning and taught me cell culture. I am particularly grateful to Jen for fruitful discussions over many experimental setups, for helping me a lot, for being my hands in the hot-lab when I was pregnant and lastly for the friendship we shared. I would also like to acknowledge Meghan for discussing results, reading and correcting this dissertation. Finally, I also would like to thank people from VBC: Henriette from Mochizuki group and Gordana from Peters lab.

Naturally, I would not be out here without the constant emotional and personal support, help and inspiration from my dearest family and long-life friends.

8 Curriculum Vitae

Katarzyna Matylla-Kulinska

Personal Information

Date of birth	18 November 1981
Birthplace	Poznań, Poland
Nationality	Polish
Children	Ignacy (2009), Aleksy (2011)
Home Address	Schoenburgstrasse 2/13 1040 Wien Austria

Education

2006-present Dr.rer.nat with Renée Schroeder at Max F.Perutz Laboratories
University of Vienna, Vienna, Austria

2003-2006 Master of Science in Molecular Biology with Jadwiga Jaruzelska
at Institute of Human Genetics, Polish Academy of Sciences
Adam Mickiewicz University, Poznań, Poland

2004-2005 Master of Science in Human Genetics with Xavier Jeunemaitre
Université Denis Diderot Paris 7, France

2000-2003 Bachelor of Science in Molecular Biology with Hanna Kmita
Adam Mickiewicz University, Poznań, Poland

1996-2001 Fryderyk Chopin Secondary Music School in viola class, Poznań, Poland

1996-2000 Karol Marcinkowski High School, Poznań, Poland

1988-1996 Henryk Wieniawski Primary Music School in violin class, Poznań, Poland

Research Experience

2006-present **Doctoral Research:** Identification and characterization of transcripts derived from human alpha satellite repetitive arrays.

2003-2006 **Undergraduate Research:** Studying the role of RNA DEAD-box helicase VASA in the human germ line by looking for its interactions with several fertility proteins.

2004-2005 **Undergraduate Research:** Characterization of the cardiovascular development of *WNK1* knock-out embryos.

Teaching Experiences

2008-2009	Tutoring Fundamentals of Biochemistry
2008	Mentoring the rotation student Olga Liska

Honors and Awards

2004-2005	The Socrates- Erasmus Scholarship

Selected presentations

2013	RNA 2013 (oral presentation) 18th annual meeting of the RNA Society, Davos, Switzerland “Human alpha satellite derived transcript interact with the active core of RNA pol II”
2010	69th Harden Conference (poster presentation) RNAP2010, Hinxton, UK “RNA pol II binding RNA aptamers involved in transcriptional silencing”
2009	RNA 2009 (poster presentation) 14th annual meeting of the RNA Society, Madison Wisconsin, USA “Human alpha satellites are transcribed”
2008	Spetses Summer School (poster presentation) Summer School on Chromatin & Transcription, Spetses Island, Greece “Characterization of human alpha satellite-derived transcripts containing RNA pol II binding elements (PBEs)”
2008	RNA 2008 (poster presentation) 13th annual meeting of the RNA Society, Berlin, Germany “Characterization of human RNAs containing RNA pol II binding elements (PBEs)”

9 Publications

Jennifer L. Boots, Frederike von Pelchrzim, Adam Weiss, Bob Zimmermann, Maximilian Radtke, Johanna Stranner, Marek Zywicki, Doris Chen, KATARZYNA MATYLLA-KULINSKA, Florian Brueckner, Patrick Cramer and Renée Schroeder (2014) *submitted*
Genome-wide screen for Pol II-binding RNA elements reveals a novel type of transcriptional control

MATYLLA-KULINSKA K.*, Tafer H.*, Weiss A., Schroeder, R. (2013) WIREs RNA (* co-first authors), *accepted*
Functional repeat-derived RNAs often originate from retrotransposon-propagated ncRNAs

MATYLLA-KULINSKA K., Tafer H., Radtke M., Zimmermann B., Boots, JL., Schroeder, R. (2013) NAR*, under revision*
Human α satellite transcripts are substrates for RNA Pol II and contain remnants of snoRNAs

MATYLLA-KULINSKA K., Boots, JL., Zimmermann, B., Schroeder, R. (2012)
Finding aptamers and small ribozymes in unexpected places.
Wiley Interdiscip Rev RNA 3: 73-91

Boots, J., MATYLLA-KULINSKA, K., Zywicki, M., Zimmermann, B., and Schroeder, R. (2012) **Genomic SELEX** in "Handbook of RNA Biochemistry"
2nd Edition Edt. Hartmann R.K., Bindereif, A., Westhof, E. Wiley-VCH

Jennifer L. Boots, Frederike von Pelchrzim, Adam Weiss, Bob Zimmermann, Maximilian Radtke, Johanna Stranner, Marek Zywicki, Doris Chen, KATARZYNA MATYLLA-KULINSKA, Florian Brueckner, Patrick Cramer and Renée Schroeder (2014) *submitted*

Genome-wide screen for Pol II-binding RNA elements reveals a novel type of transcriptional control

My contribution to the RNA polymerase II elements (PBE) project was setting up the *in vitro* transcription system in HeLa Nuclear Extract to test PBE RNAs. I initiated the analysis of ACRO satellite expression. I characterized α satellite-derived PBEs. I was also taking part in discussions on *in cis* inhibition experiments and conferring about a model for the PBE-mediated RNA polymerase II regulation.

Genome-wide screen for Pol II-binding RNA elements reveals a novel type of transcriptional control

Jennifer L. Boots[1,4], Frederike von Pelchrzim[1,4], Adam Weiss[1,4], Bob Zimmermann[1,4], Maximilian Radtke[1], Johanna Stranner[1], Marek Zywicki[2], Doris Chen[1], Katarzyna Matylla-Kulinska[1], Florian Brueckner[3], Patrick Cramer[3] 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] Gene Center Munich and Department of Biochemistry, Ludwig-Maximilians-Universität München, Feodor-Lynen-Strasse 25, 81377 Munich, Germany, and Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany.

[4] These authors contributed equally to this work.

* Correspondence: renee.schroeder@univie.ac.at

Running Title: Pol II aptamers control their own transcription

Summary: Transcription is tightly regulated—not only by proteins but also by RNAs. To find RNAs that regulate transcription through direct interaction with RNA polymerase II (Pol II) we performed Genomic SELEX using Pol II as a bait and identified a variety of Pol II-binding elements (PBEs). PBEs are enriched in repeat elements like ACRO1 satellites and LINEs as well as in protein-coding genes. We show that single PBEs reduce transcriptional output in their endogenous context and multiple PBEs confer complete transcriptional silencing in a reporter gene expression assay. Our results suggest that ACRO1 satellites are self-regulatory elements that disrupt their own transcription *in cis*. We demonstrate that Genomic SELEX in combination with deep sequencing is a powerful tool to screen genomes for regulatory RNA elements, also within repeat-derived regions. We propose a novel *cis*-acting type of transcription regulation, wherein nascent RNA interferes with Pol II elongation.

Highlights:

Genomic SELEX identifies RNA aptamers of human RNA Pol II

Pol II-binding elements (PBEs) are enriched in repeat elements like ACRO1 and LINE1

PBEs co-transcriptionally inhibit expression a reporter gene *in vivo*

PBEs employ a novel type of *cis*-acting transcriptional control by the nascent RNA

Introduction

Several non-coding RNAs regulate the activity of RNA polymerase II (Pol II) in an indirect way by interacting with proteins involved in transcriptional control (Barrandon et al., 2008; Goodrich and Kugel, 2006; Wang et al., 2011). So far, only three naturally occurring RNAs have been reported to directly bind to RNA polymerase and inhibit transcription: 6S RNA (*Escheriechia coli*), B2 RNA (*Mus musculus*) and Alu RNA (*Homo sapiens*) (Espinoza et al., 2004; Mariner et al., 2008; Wassarman and Storz, 2000). In addition, an *in vitro* selected RNA, the FC aptamer, is able to inhibit transcription of yeast Pol II *in vitro* (Thomas et al., 1997) by binding to the active center cleft (Kettenberger et al., 2006). Certain RNAs are also able to serve as template for an ancient RNA-dependent RNA polymerase activity of Pol II (Lehmann et al., 2007). Thus RNA polymerases have the potential to bind a multitude of RNAs with different affinities (Wettich and Biebricher, 2001; Windbichler et al., 2008), but the consequences of these interactions have not been addressed in detail.

The bacterial 6S RNA is the best-studied example of a *trans*-acting RNA that regulates the activity of RNA polymerase. Upon entry into stationary phase, 6S RNA binds the active center of σ^{70}-containing holoenzyme and inhibits housekeeping transcription (Wassarman and Storz, 2000). 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 (Wassarman, 2007).

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 in stress (Liu et al., 1995) and downregulate initiation of Pol II transcription at promoters (Mariner et al., 2008). In addition, non-coding RNAs influence transcription indirectly by binding to transcription factors (Barrandon et al., 2008).

Besides *trans*-regulation of transcription by small RNAs, the nascent RNA itself can affect RNA polymerase *in cis*. Bacterial riboswitches, located in the 5' unstranslated regions of mRNAs, can dynamically refold in response to ligand binding or temperature shift and promote elongation or termination (Serganov and Nudler, 2013). Similarly, eukaryotic Pol II has been shown to be affected by secondary structure in the nascent RNA *in vitro*. By inhibiting backtracking, stable secondary structure elements prevent pausing and thereby increase the rate of transcription (Zamft et al., 2012). Sequences within nascent transcripts can also contain target sites for regulatory factors, such as the yeast termination factors Nrd1 and Nab3. Their recognition motifs are enriched in divergent transcripts but underrepresented in mRNAs, which ensures directionality of promoters (Schulz et al., 2013). In contrast to *trans*-acting RNAs, nascent RNA has never been observed to regulate Pol II by direct interaction.

In this work, we demonstrate that RNA can be a potent *cis*-regulator of transcription. We screened the human genome for RNAs that directly interact with Pol II using Genomic SELEX in combination with deep sequencing. This procedure allows a genome-wide functional analysis of all RNA elements encoded in the genome independent of their expression levels (Lorenz et al., 2006; Zimmermann et al., 2010). We obtained a collection of RNA aptamers with high affinity to Pol II and termed them Pol II-binding elements (PBEs). They are distributed thoughout the entire genome, in both genic and intergenic regions. In particular, PBEs are present in many repetitive elements, such as ACRO1 satellites and LINE retrotransposons. We show that PBEs attenuate, and ACRO1 satellites completely abolish, their own transcription. We thus propose a novel mode of regulation of transcription, in which nascent RNA prevents Pol II elongation.

Results

SELEX with human genomic RNA library identifies Pol II aptamers

We constructed an RNA library (Singer et al., 1997) 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 cannot be obtained in sufficient purity and quantity. Due to the high degree of conservation of the enzymes (Cramer et al., 2000), and the fact that murine B2 RNA is able to bind to the *S. cerevisiae* Pol II core (Kettenberger et al., 2006), we assumed that the binding sites for other RNAs might be conserved as well. During the selection procedure (Figure 1A), Pol II-binding RNA elements (PBEs) started to enrich in the 4th SELEX cycle (Figure 1B). 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 low nanomolar range. As a first test set for evaluating the selected RNAs, 200 clones from the 7th cycle were Sanger-sequenced, resulting in 74 individual RNAs. We validated the selection by showing that a set of exemplary RNAs from the 7th SELEX cycle are expressed and bind human Pol II *in vitro* and *in vivo* using band-shift assay and co-immunoprecipitation with an antibody against Pol II (Figure 1C and Figures. S1A and S1B). These RNAs were derived from repeat regions, such as LINE elements, SINEs and satellites (the corresponding nucleotide sequences are specified in Supplemental Experimental Procedures). These findings show that the successfully selected endogenous PBE-containing RNAs bind to Pol II in their natural context. Because binding of total RNA from the 7th cycle pool to purified human Pol II can be outcompeted by B2 RNA, a portion of PBEs presumably interact with the Pol II active site (Figure S1C).

PBEs are found throughout the human genome, most notably in repetitive regions

Although the selection procedure resulted in successful isolation of RNA aptamers with high affinity 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 more diverse sequences. We therefore subjected this enriched pool to deep sequencing and computational analysis (Figure S2). A database was established to better access the outcome of the selection (http://alu.abc.univie.ac.at/pbe), which links all sequences to their genomic regions displayed in a GBROWSE instance (Stein et al., 2002). PBEs were analyzed in two different ways according to whether they mapped uniquely or multiple times to the genome. The unique hits were enriched in genic and intergenic regions, sense as well as antisense relative to the coding strand. The most prominent, PBE 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). PBEs do not contain one single dominant sequence or structural motif, suggesting that Pol II can bind many diverse RNA molecules (Windbichler et al., 2008). Generally, PBEs are CA-rich (Figure S2D) and the highest enrichment score among the repeats was reached by $(\mathrm{CAC}^{\mathrm{A}}/_{\mathrm{T}}{}^{\mathrm{C}}/_{\mathrm{A}})_n$ 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 ACRO-containing repeat elements (Figure 2A and Figure S3A). ACREs are partially or fully conserved among all sequenced primates (Figure S3B), however no non-primate organisms were 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 (Figure 2B-C). Many ACRO1 satellites have been mapped by FISH to chromosome 3 and to the acrocentric chromosomes 13, 14, 15, and 22 (Warburton et al., 2008), however these regions have not yet been annotated, indicating that many, if not most, ACREs are not represented in the current build of the human genome. Figure 2D shows SELEX read stacks mapping to ACRO1 consensus unit defining the Pol II-binding aptamer. We were unable to detect stable transcripts derived from ACRO1 satellites in HeLa cells (data not shown). It has nevertheless been reported that ACRO1 is expressed to a very low level in several epithelial cancers (Ting et al., 2011).

PBEs disrupt transcription *in cis*

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 transcription by a sequence-specific but otherwise unknown mechanism (Han et al., 2004). There are multiple PBEs within the 4 kb LINE1 ORF2 sequence (Figure 3A) and elimination of flanking PBEs led to a partial recovery of transcription in an *in vivo* reporter system (Figures 3B-C). Encouraged by this observation, we used the same system to test whether the highly enriched PBEs, such as ACRO1 repeats and PBE 5765, could also lead to this type of *cis*-acting transcriptional disruption. Single PBEs inserted into the transcriptional unit had no or little effect on steady-state RNA levels (Figure S4A). However, multiple PBEs cloned in tandem severely disrupted transcription of the reporter (Figures 3D and E and Figure S4B) and the number of PBEs correlated with the extent of transcriptional repression (Figures S4C and D). This disruption was alleviated in control reverse-complement insertions, showing its sequence and/or structural specificity.

PBE-mediated regulation is co-transcriptional

We further found that transcriptional disruption by PBEs is promoter-independent (Figure S5A) and that it has no effect *in trans* on other loci within a cell (Figure S5B). To distinguish between post- and co-transcriptional regulation we monitored transcript levels upstream and downstream of the ACRO insertion by RT-qPCR (Figure 4A). The significant decrease of RNA levels downstream of the inserted ACRO sequence relative to the RNA levels upstream of the insertion indicates that RNA production is compromised at the ACRO locus. Moreover, this effect was lost in the Poly(A)+ fraction, but not in the Poly(A)- fraction of total RNA (Figures 4B and C and Figure S5C) suggesting that the regulation cannot take place once transcription is completed. These results show that the PBE-mediated inhibition is co-transcriptional and spatially restricted to the vicinity of the PBE template.

To test whether individual PBEs exert transcriptional repression in their endogenous context, we took the same approach to quantify transcript levels upstream and downstream of the PBE 5765 within *MARK4* gene intron 13 (Figure 4D). The results show a moderate decrease of downstream RNA indicating that even a single PBE can modulate transcriptional output in its natural context.

Discussion

The transcription machinery in humans is regulated by a multitude of protein factors and a growing number of non-coding RNAs. They act predominantly during initiation and the transition checkpoint associated with promoter-proximal pausing. Nevertheless, the elongation phase of transcription is also regulated, for example by chromatin state or modifications of the Pol II C-terminal domain (CTD). In this work, we present evidence that Pol II can "sense" the nature of certain transcripts and that some elements encoded in the human genome have the potential to interfere with their own transcription *in cis*. We thus propose a novel mechanism of transcriptional control in human cells, wherein the nascent RNA binds to the transcribing Pol II making it elongation-incompetent (Figure 4E). It has to be determined wether Pol II stalls on, or dissociates from, the template DNA, but the resulting RNA presumably lacks hallmarks of mature RNA, such as Poly(A) tail, and is eliminated from the cell.

Importantly, there are many different types of RNA that can be accommodated in the active site of Pol II pointing to a potentially wide range of regulatory motifs. Secondary structure of nascent RNA has recently been shown to affect the rate of Pol II transcription *in vitro* by inhibiting backtracking and thus escape from pausing (Zamft et al., 2012). Interaction between RNA Pol II CTD with mRNA has also been reported to suppress transcription-coupled 3'-end processing and a few of our aptamers contain the motif isolated in this random SELEX experiment (Kaneko and Manley, 2005). Very 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 (Zhang et al., 2013). Here we add another layer of transcriptional regulation that involves *cis*-acting sequences within the nascent transcript. This might be an essential self-regulatory strategy for repeat elements to stay silent enabling their survival in the genome during evolution. We characterize ACRO1 satellites as an example and we propose that this is also the case of LINE1 retrotransposons, whose self-regulatory properties were described previously (Han et al., 2004). We suggest that PBEs present in LINE1 ORF2 affect elongation similarly to ACRO1 repeats, as their partial elimination from the sequence slightly alleviated the repression (see Figure 3C). In addition, we hypothesize that PBE-mediated control of transcription plays a role in gene regulatory processes, which depend on the rate of Pol II progression, such as alternative splicing and termination (Mata et al., 2003). Indeed, several PBEs map downstream of alternative splice sites and alternative polyadenylation sites.

We further demonstrate that Genomic SELEX in combination with deep sequencing is a powerful tool to discover novel RNAs with specific properties especially within repetitive sequences, which are not amenable to classical genetic methods. In contrast to massive sequencing of total RNA, Genomic SELEX selects for RNAs with a defined binding property irrespective of their expression levels. We show that the human genome encodes many transcripts with high affinity to Pol II, suggesting that an unanticipated large number of RNAs have the potential to regulate their own transcription.

Experimental Procedures:

Library construction and Genomic SELEX

The genomic library was created as described in (Lorenz et al., 2006; Zimmermann et al., 2010), 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 (Lorenz et al., 2006). 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 % confluency 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 ribonucleosid 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 (Peritz et al., 2006). 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. For immunoprecipitations the antibodies were used in a concentration of 2 µl/ml.

Transfection, microscopy and RNA preparation

HeLa cells were grown to 70-90 % confluency and transfected with 0.4 μg 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) softwares. 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 μg 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 μg 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 Table S1. Transfection was controled 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.

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

Acknowledgments: The sequences presented in this work are filed in a specific portal containing the PBE database and the link to the human genome (http://alu.abc.univie.ac.at/pbe/). 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 PBEs. We thank the members of the Schroeder lab for helpful 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. P.C. was supported by the Deutsche Forschungsgemeinschaft (SFB646, SFB960, SFB1064, CIPSM, NIM, QBM, GRK1721), an ERC Advanced Grant, the Jung-Stiftung and the Vallee Foundation.

References:

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

Cramer, P., Bushnell, D.A., Fu, J., Gnatt, A.L., Maier-Davis, B., Thompson, N.E., Burgess, R.R., Edwards, A.M., David, P.R., and Kornberg, R.D. (2000). Architecture of RNA polymerase II and implications for the transcription mechanism. Science *288*, 640–649.

Espinoza, C.A., Allen, T.A., Hieb, A.R., Kugel, J.F., and Goodrich, J.A. (2004). B2 RNA binds directly to RNA polymerase II to repress transcript synthesis. Nat. Struct. Mol. Biol. *11*, 822–829.

Goodrich, J.A., and Kugel, J.F. (2006). Non-coding-RNA regulators of RNA polymerase II transcription. Nat Rev Mol Cell Biol *7*, 612–616.

Han, J.S., Szak, S.T., and Boeke, J.D. (2004). Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature *429*, 268–274.

Kaneko, S., and Manley, J.L. (2005). The mammalian RNA polymerase II C-terminal domain interacts with RNA to suppress transcription-coupled 3′ end formation. Mol. Cell *20*, 91–103.

Kettenberger, H., Eisenführ, A., Brueckner, F., Theis, M., Famulok, M., and Cramer, P. (2006). Structure of an RNA polymerase II-RNA inhibitor complex elucidates transcription regulation by noncoding RNAs. Nat. Struct. Mol. Biol. *13*, 44–48.

Lehmann, E., Brueckner, F., and Cramer, P. (2007). Molecular basis of RNA-dependent RNA polymerase II activity. Nature *450*, 445–449.

Liu, W.M., Chu, W.M., Choudary, P. V, and Schmid, C.W. (1995). Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. Nucleic Acids Res. *23*, 1758–1765.

Lorenz, C., von Pelchrzim, F., and 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*, 2204–2212.

Mariner, P.D., Walters, R.D., Espinoza, C.A., Drullinger, L.F., Wagner, S.D., Kugel, J.F., and Goodrich, J.A. (2008). Human Alu RNA Is a Modular Transacting Repressor of mRNA Transcription during Heat Shock. Mol. Cell *29*, 499–509.

Mata, M. De, Alonso, C.R., Fededa, J.P., Pelisch, F., Cramer, P., Bentley, D., and Kornblihtt, A.R. (2003). Affects Alternative Splicing In Vivo. *12*, 525–532.

Peritz, T., Zeng, F., Kannanayakal, T.J., Kilk, K., Eiríksdóttir, E., Langel, U., and Eberwine, J. (2006). Immunoprecipitation of mRNA-protein complexes. Nat. Protoc. *1*, 577–580.

Schulz, D., Schwalb, B., Kiesel, A., Baejen, C., Torkler, P., Gagneur, J., Soeding, J., and Cramer, P. (2013). Transcriptome surveillance by selective termination of noncoding RNA synthesis. Cell *155*, 1075–1087.

Serganov, A., and Nudler, E. (2013). A decade of riboswitches. Cell *152*, 17–24.

Singer, B.S., Shtatland, T., Brown, D., and Gold, L. (1997). Libraries for genomic SELEX. Nucleic Acids Res. *25*, 781–786.

Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., et al. (2002). The generic genome browser: a building block for a model organism system database. Genome Res. *12*, 1599–1610.

Thomas, M., Chédin, S., Carles, C., Riva, M., Famulok, M., and Sentenac, A. (1997). Selective targeting and inhibition of yeast RNA polymerase II by RNA aptamers. J. Biol. Chem. *272*, 27980–27986.

Ting, D.T., Lipson, D., Paul, S., Brannigan, B.W., Akhavanfard, S., Coffman, E.J., Contino, G., Deshpande, V., Iafrate, A.J., Letovsky, S., et al. (2011). Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. Science *331*, 593–596.

Wang, X., Song, X., Glass, C.K., and Rosenfeld, M.G. (2011). The long arm of long noncoding RNAs: roles as sensors regulating gene transcriptional programs. Cold Spring Harb. Perspect. Biol. *3*, a003756.

Warburton, P.E., Hasson, D., Guillem, F., Lescale, C., Jin, X., and Abrusan, G. (2008). Analysis of the largest tandemly repeated DNA families in the human genome. BMC Genomics *9*, 533.

Wassarman, K.M. (2007). 6S RNA: a small RNA regulator of transcription. Curr. Opin. Microbiol. *10*, 164–168.

Wassarman, K.M., and Storz, G. (2000). 6S RNA regulates E. coli RNA polymerase activity. Cell *101*, 613–623.

Wettich, A., and Biebricher, C.K. (2001). RNA species that replicate with DNA-dependent RNA polymerase from Escherichia coli. Biochemistry *40*, 3308–3315.

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

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

Zhang, Y., Zhang, X.O., Chen, T., Xiang, J.F., Yin, Q.F., Xing, Y.H., Zhu, S., Yang, L., and Chen, L.L. (2013). Circular Intronic Long Noncoding RNAs. Mol. Cell *51*, 792–806.

Zimmermann, B., Bilusic, I., Lorenz, C., and Schroeder, R. (2010). Genomic SELEX: A discovery tool for genomic aptamers. Methods *52*, 125–132.

Figures

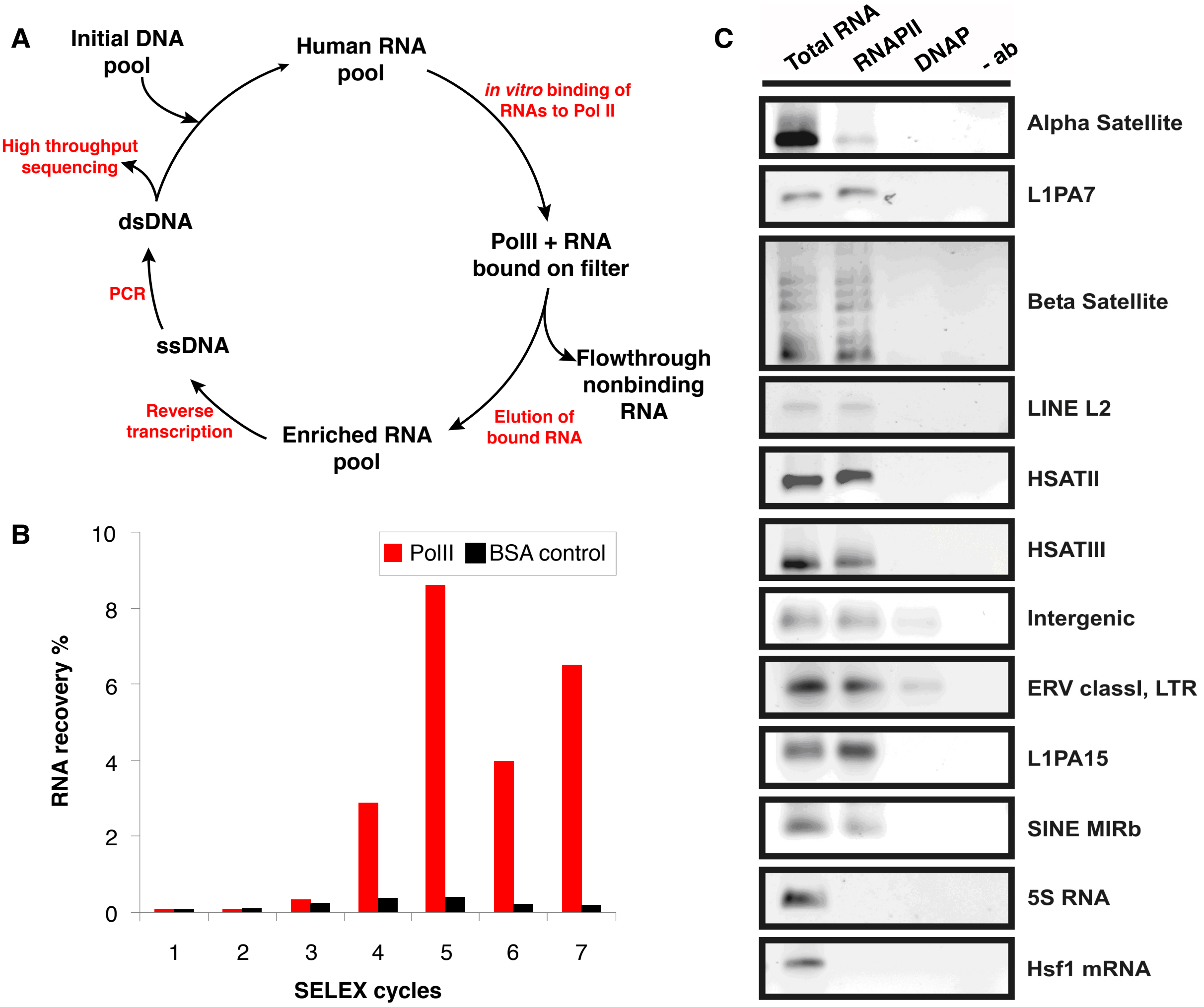


Figure 1. Genomic SELEX for RNA polymerase II-binding elements (PBEs). (A) The initial human DNA library was *in vitro* transcribed and the resulting RNA pool was bound to the highly purified Pol II. Protein-bound RNAs were retained on the filter and non-binding RNAs were discarded. Selected RNAs were 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 (Lorenz et al., 2006). **(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. Lane “c” indicates the control RT-PCR on total RNA. 5S and Hsf1 are abundant cellular RNAs used as control that were not enriched by SELEX. See also Figures S1 and S2.

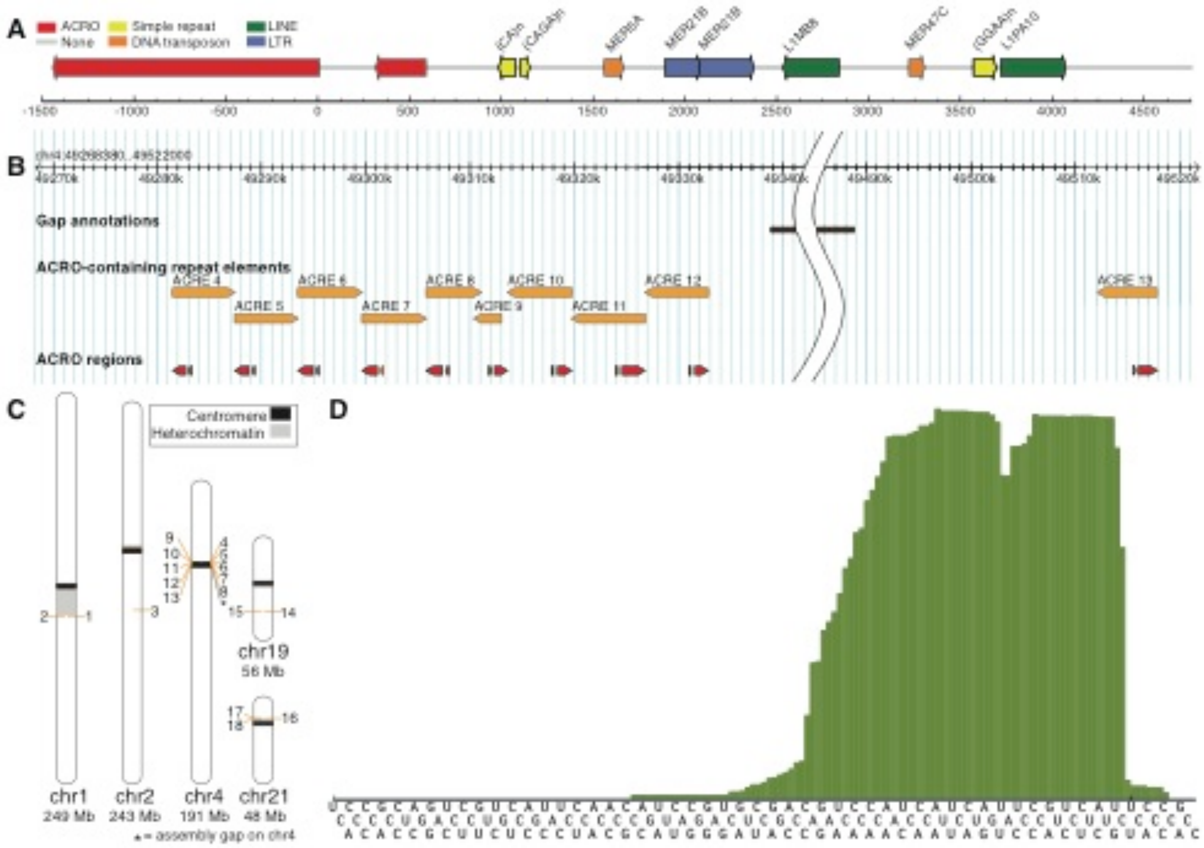


Figure 2. The structure and distribution of ACRO1 satellites. (A) ACRE (ACRO-containing repeat element) is a higher order repeat structure of 6 kb harboring the ACRO satellite array. **(B)** Organization of the ACRE cluster in the pericentromeric region of chromosome 4, the densest region of sequenced ACREs. **(C)** ACREs were found on chromosomes 1, 2, 4, 19 and 21. **(D)** Sequence of ACRO1 consensus repeat unit and its SELEX enrichment profile. See also Figure S3.

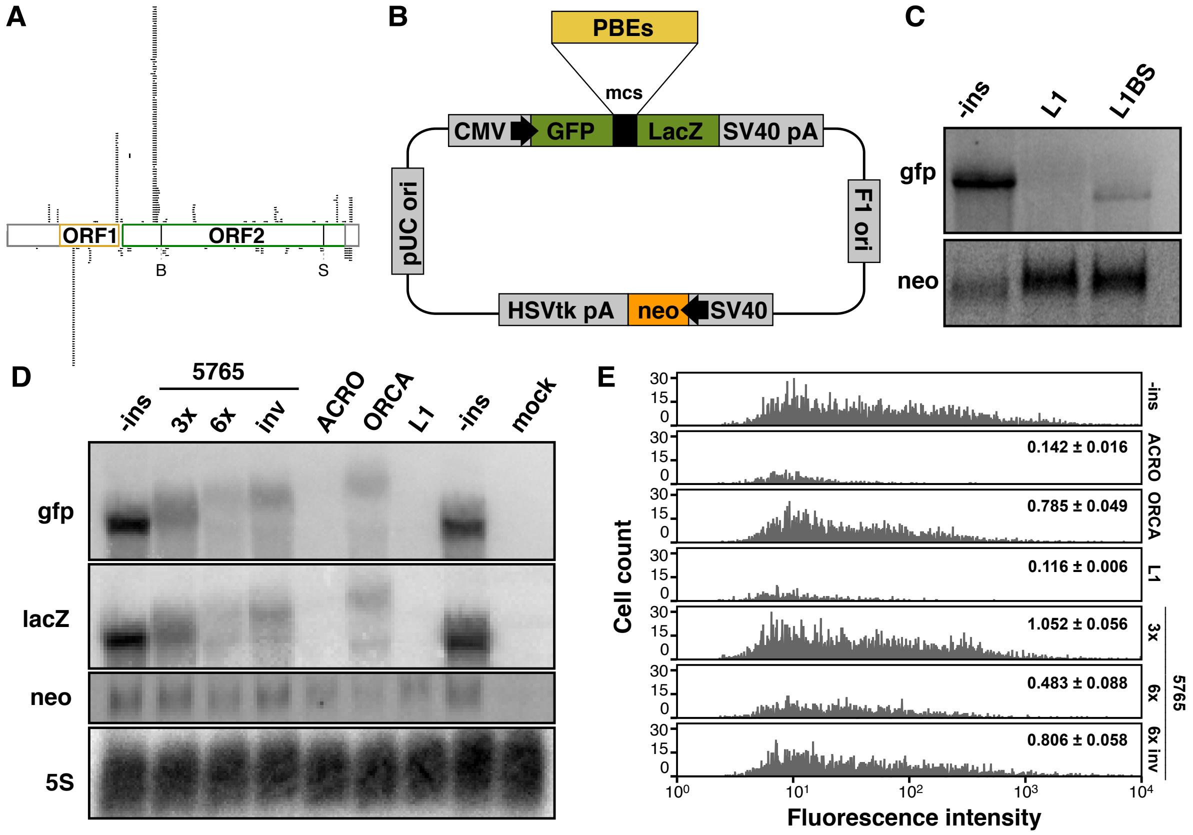


Figure 3: PBEs induce transcriptional silencing. (A) The LINE1 retrotransposon is illustrated here with the restriction sites "B" and "S" indicated (Han et al., 2004). LINE1-associated PBEs from the 7th SELEX cycle were mapped to the consensus with at least 80 % identity. **(B)** Vector used to monitor *in vivo* expression of the reporter cassette (adapted from ref.18). PBEs were cloned between the GFP and the LacZ sequences or, in case of L1 and L1BS, in place of LacZ gene. **(C), (D)** Northern blot analyses of total RNA extracted from HeLa cells transfected with various PBE-containing reporters. Probes detect regions within GFP, LacZ, neo as a transfection control and 5S rRNA as a loading control. The reporter cassettes contained empty GFP-LacZ fusion (-ins), LINE1 ORF2 (L1), its shortened version trimmed to the region between the "B" and "S" sites (L1BS), PBE 5765 cloned in tandem three times (3x), six times (6x) and six times in reverse complement (inv), ACRO1 1.4 kb element (ACRO), and its reverse complement (ORCA). **(E)** Quantification of GFP expression by flow cytometry. 10,000 cells from each sample were analysed 24 h post-transfection and their fluorescence levels were determined on a 1024-channel scale (fluorescence intensity). Only GFP-positive cells (determined by comparison with mock-treated sample) are plotted (count). Total number of GFP-positive cells relative to "–ins" (-ins = 1) is indicated ± SEM of five experiments. See also Figure S4

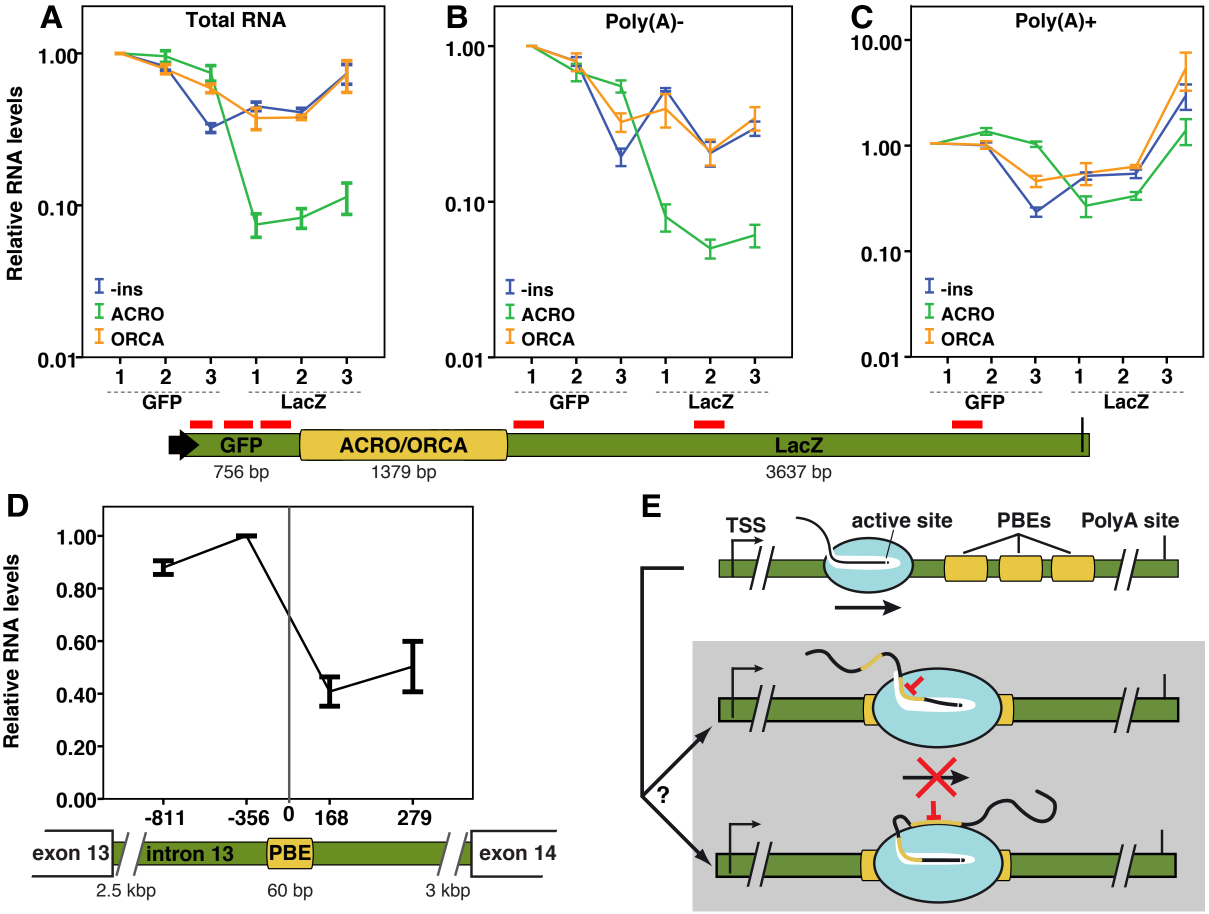


Figure 4: Autoregulation of PBEs is co-transcriptional. (A-C) RT-qPCR quantification of six different amplicons along the reporter transcript. Total RNA was isolated from HeLa cells 24 h after transfection with vectors carrying no insert (blue lines), ACRO (green lines) or its reverse complement ORCA (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 PBE 5765. Distance of the amplicon from the PBE (in bp) is indicated. Values are relative to amplicon -356. Error bars represent SEM of three experiments. **(E)** Model of transcriptional inhibition by PBEs. Pol II initiates at transcription start site (TSS) and continues into productive elongation. When PBEs 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 PBEs might be needed for efficient regulation. See also Figure S5.

Tables

Table 1. Top unique PBEs

PBE ID	Gene	Chromosome	Read count	Length (nt)	Orientation[a]
5765	Microtubule affinity-regulating kinase 4 (MARK4)	19	948	113	sense
141	Histone deacetylase 1 (HDAC1)	1	674	51	anti
858	Microtubule affinity-regulating kinase 1 (MARK1)	1	594	42	anti
10384	Guanylyl cyclase-activating protein 1 (GUCA1A)	6	535	58	anti
933	Probable saccharopine dehydrogenase (SCCPDH)	1	401	105	sense
891	*Intergenic*	1	316	88	-
2312	Voltage-dependent L-type calcium channel subunit alpha-1C (CACNA1C)	12	261	91	anti
122	Sodium/hydrogen exchanger 1 (SLC 9A1)	1	247	60	anti
6885	Disintegrin and metalloproteinase domain containing protein 33 (ADAM33)	20	244	40	sense
9515	*Intergenic*	5	176	92	-
1990	Homo sapiens olfactory receptor, family 9, subfamily Q, member 1 (OR9Q1)	11	117	91	sense
5920	Hippocalcin like protein 1 (HPCAL1)	2	92	34	sense
90	Immunoglobulin superfamily member 21 (IGSF21)	1	86	58	anti

[a] relative to mRNA strand

Table 2. Top repeat-derived PBEs

		Reads		
Repeat type	Repeat family	Sense	Antisense	Fold enrichment[a]
Simple repeat	(CACAC)n	1267	93	4442
Simple repeat	(CACTA)n	111	16	2665
Simple repeat	(CACAA)n	197	7	2217
Satellite	ACRO1	1029	1	2141
Simple repeat	(CACCAT)n	2020	20	1876
LINE1	L1HAL-2a MD	2498	8	1293
…				
	(CA)n	5561	831	231
…				
LINE	L1HS[b]	206	29	7

[a] enrichment of the more prominent strand normalized to the abundance in the genome

[b] L1HS is listed here because of its regulatory properties described previously (see text)

MATYLLA-KULINSKA K.*, Tafer H.*, Weiss A., Schroeder, R. (2013) WIREs RNA (* co-first authors), *accepted*

Functional repeat-derived RNAs often originate from retrotransposon-propagated ncRNAs

I contributed to this publication in discussing the conception, writing the introductory section and “snoRNAs are ancestors of α sat RNAs” chapter, as well as joining paragraphs written by other authors and handling the manuscript.

Functional repeat-derived RNAs often originate from retrotransposon-propagated ncRNAs

[A1]

First authors: Full name and affiliation; plus email address if corresponding author

[Katarzyna Matylla-Kulinska, Department of Biochemistry and Cell biology, Max F. Perutz Laboratories, University of Vienna]

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

Second author: Full name and affiliation; plus email address if corresponding author

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

Third author: Full name and affiliation; plus email address if corresponding author

[Renée Schroeder*, Department of Biochemistry and Cell biology, Max F. Perutz Laboratories, University of Vienna; renee.schroeder@univie.ac.at][A2]

Abstract

[The human genome is scattered with repetitive sequences and the ENCODE project revealed that 60-70 % of the genomic DNA is transcribed into RNA. As a consequence, the human transcriptome contains a large portion of repeat-derived RNAs (repRNAs). Here we present a hypothesis for the evolution of novel functional repeat-derived RNAs from non-coding RNAs (ncRNAs) by retrotransposition. Upon amplification, the ncRNAs can diversify in sequence and subsequently evolve new activities, which can result in novel functions. Non-coding transcripts derived from highly repetitive regions can therefore serve as a reservoir for the evolution of novel functional RNAs. We base our hypothetical model on observations reported for SINEs derived from 7SL RNA and tRNAs, α satellites derived from snoRNAs and SL RNAs derived from U1 snRNA. Further, we present novel putative human repeat-derived ncRNAs obtained by the comparison of the Dfam and Rfam databases, as well as several examples in other species. We hypothesize that novel functional ncRNAs can derive also from other repetitive regions and propose Genomic SELEX as a tool for their identification.][A3]

[The repetitive genome

The human genome is composed of approximately 3.3 billion base pairs. Canonical genes occupy 30 %, but only an estimated 1.5 % of the genomic content has protein-coding capacity. Repeats make up at least 51 % of the genome[1,2] (**Figure 1**) and can be classified by sequence similarity, dispersal patterns or by function. Most of the repetitive DNA consists of interspersed transposable elements (TEs), often referred to as parasitic DNA. About 45 % of the human genome falls into this class and even more is proposed to be transposon-derived[2].

TEs are either DNA transposons, which are mobilized by a cut-and-paste mechanism, or retrotransposons, which propagate in the host genome via RNA intermediates in a copy-and-paste manner. Retrotransposons constitute a large fraction of the DNA in many eukaryotes and some of them are still actively retrotransposing, *e.g.* Alu's germline transposition rate is estimated as 1 per 20 births[3]. There are three types of mammalian retrotransposons: i) long interspersed nuclear elements (LINEs) that transpose autonomously and account for 20.4 % of the genomic sequence; ii) short interspersed nuclear elements (SINEs) that make up 13.1 % of the genome and their transposition depends on other TEs, such as LINEs, since they lack a functional reverse transcriptase; and iii) long terminal repeats (LTRs) that account for 8.3 % of the human genome.

Although transposition events can cause damage to the host, there is also substantial evidence that TEs have been important for the evolution and function of genes and genomes[4–7]. It has been suggested that mobile DNA can serve as a dynamic reservoir for new cellular functions because TEs can evolve new genes that are beneficial to the host[8]. In an analogous way, small RNA-derived retroelements can also give rise to novel RNA-coding genes. The primate BC200 non-coding RNA (ncRNA) is the first known example of an Alu element that evolved into a novel functional small RNA-coding gene[9].

Another class of genomic repetitive sequences consists of arrays of high-copy-number tandem repeats known as satellite DNA. It accounts for about 8 % of the human genome[10] and is classified into macro-, mini- and microsatellites. Macrosatellites, or satellites, span up to hundreds of kilobases within the constitutive heterochromatin. They differ substantially from the rest of the genome in nucleotide content and hence can be separated by buoyant density gradient centrifugation, as satellite bands[11]. An example of a macrosatellite element is the alpha satellite family discussed below. Minisatellite arrays are somewhat shorter. For example, telomeric repeats with a short hexanucleotide repeat unit located at chromosomal ends span 10-15 kilobases in humans. Microsatellites are the smallest tandem repeats, often not longer than 4 bp, and among the most variable DNA sequences[12]. The most common CA/TG dinucleotide tandem repeats constitute 0.5 % of the human genome.]

[Repeat-derived ncRNAs, repRNAs

Rapid advances in next-generation sequencing allowed a deep insight into transcriptomes, and the ENCODE consortium reported that highly repetitive genomic regions are also transcribed in humans. These reports opened a lively debate about potential functions of these transcripts. The widespread transcription of repetitive DNA could be either i) producing functional, active non-coding RNAs, ii) important *per se* to set the chromatin state or to interfere with transcription of other genes, or iii) simply an insignificant background process. There is no straightforward way to distinguish between meaningful transcripts and transcriptional noise. So far evolutionary conservation served as a good indication of RNA function. However, recently this correlation has been under debate[13–15]. At this moment, only the analysis of individual RNAs can yield data on their functionality.

The impact of repeats on the evolution of genomes and protein-coding genes has been described elsewhere[16,17]. Here we summarize what is known about the evolution and function of several ncRNAs expressed from repetitive DNA. We coin the term repRNAs (repeat-derived RNAs) for non-coding transcripts with a distinct activity that are expressed from repetitive elements. We present a hypothesis that functional repRNAs can originate from retrotransposon-propagated ncRNAs. By acquiring the ability to retrotranspose, ncRNAs can become highly amplified and spread throughout the genome. Some of the new copies escape previous evolutionary constraints, accumulate mutations and as a result lose their original function and might acquire novel activities. Therefore, transcripts derived from highly repetitive regions can be a rich reservoir for the evolution of novel functional RNAs (**Figure 2**). It has to be kept in mind that even if a repRNA evolves new activities, it does not necessarily bring about a functional change in the cell. Only if the novel activity leads to a downstream cellular event can we clearly attribute a function to these novel ncRNAs**]**[A5]

[Examples of repRNAs evolved from non-coding RNAs]

[Signal recognition particle 7SL RNA as the ancestor of Alu elements

Alu repeats are a primate-specific SINE family. They are approximately 300 bp in length and originated from a ncRNA, the signal recognition particle component 7SL RNA, through processing and duplication[18,19]. Alu and its rodent counterpart B1 RNA evolved from 7SL in a common ancestor of primates and rodents around 100 million years ago[20,21]. There are approximately 10^6 copies of Alu elements making up 10.7 % of the human genome. Similarly, there can be up to 10^6 B1 elements in rodent genomes[22]. The 7SL RNA is the first representative for our model of retrotransposon-mediated evolution of novel RNAs: the 7SL RNA was retrotransposed, then propagated to a very high copy number to eventually give rise to ncRNAs with novel activities as well as several RNA domains that impact on gene evolution and expression.

Because SINEs contain an original RNA polymerase III promoter, Alu elements can be transcribed into individual RNAs. They have been shown to be induced in stress conditions, such as heat shock or cycloheximide treatment[23], and to inhibit transcription of RNA polymerase II *in trans*[24]. It has been proposed that direct interaction of Alu and RNA polymerase II at promoters leads to down-regulation of housekeeping transcription, presumably as a part of complex cellular stress response[24]. If this novel activity of Alu ncRNA has a functional relevance for the cell, still needs to be demonstrated.

Alu sequences are also present as domains embedded in many transcripts of protein-coding genes, as well. The Alu consensus sequence contains up to 10 potential 5' donor splice sites and up to 13 potential 3' acceptor sites[25]. As a consequence of many Alu insertions into genes, 5 % of all alternatively spliced exons within protein-coding regions contain Alu sites. Thus Alu sequences are elements that play an important role in the evolution of novel genes. An interesting example was reported where an Alu element gave rise to a novel 5' exon in the human tumor necrosis factor type 2 gene (*p75TNFR*) providing a novel N-terminal protein domain resulting in a novel receptor isoform[26]. In addition, gene-integrated Alus can be a source of promoters, enhancers, silencers, insulators and influence mRNA stability[27].

Thus 7SL is a prominent example of an ncRNA that has evolved diverse functions upon retrotransposition and amplification. The second lineage of SINEs derived from 7SL, the B1 elements, is much less studied than the Alu elements but there is evidence that it has also evolved regulatory functions in rodents[28].]

[tRNA-derived ncRNAs

LINE-1 Reverse Transcriptase (RT) is thought to recognize LINE-1 mRNA partially by a sequence-specific fashion and partially by a mechanism called *cis*-preference. While the RT is being translated, the nascent protein simply binds the nearest RNA, which most often is the mRNA that encodes it[29]. In order for SINEs to exploit *cis*-preference and serve as template for LINE-1 RT, they have to be able to come close to the translating ribosome[27]. Therefore it comes as no surprise that the vast majority (96 %) of SINE families originate from tRNAs[30,31].

tRNAs have evolved diverse functions after retrotransposition and amplification. Rodent-specific neuronal BC1 RNA is a translational repressor that specifically targets eIF4A and strongly impedes its helicase activity[32]. BC1 is 152-nucleotide long, twice the length of tRNA^{Ala}. While the sequence similarity of mouse tRNA^{Ala} and the BC1 5' region amounts to 80 %, the secondary structure is a stable hairpin instead of a cloverleaf-like structure. The BC1 gene was generated by retrotransposition of tRNA^{Ala} and arose after the mammalian radiation but before the diversification of Rodentia. The cDNA copy of tRNA^{Ala} was integrated in a locus that is expressed specifically in neurons[33,34].

Another example of tRNA SINE-derived functional RNA is B2, which is present on average in 10^5 copies throughout rodent genomes[35]. The heat shock-induced B2 is transcribed by RNA polymerase III into RNAs of variable sizes from 200 to 600 nucleotides[36]. B2 consists of the 5' tRNA-like sequence[37] followed by a polyadenylated 3' tail[38]. Rodent B2, like human Alu, was proposed to be a specific inhibitor of RNA polymerase II, binding an RNA-docking site in the core polymerase complex and, as a consequence, preventing the formation of an active closed complex[39,40]. Espinoza et al[41] further showed that a 51 nucleotide sequence of the B2 3' region was responsible for repressing RNA polymerase II activity.]

[snoRNAs are ancestors of α sat RNAs

The primate-specific α satellites belong to long tandem repeats and consist of 171 bp long units organized in a head-to-tail manner. Human α satellites are annotated at 44,058 loci covering 0.1 % of the genome. Each human centromere contains a chromosome-specific higher-order array of α satellites[42] that are positioned tandemly to span 3-5 Mb. Typically, the units within the higher-order repeats are highly similar (95-100 % identity)[43,44] due to sequence homogenization. In the pericentromeric regions α satellites occur as monomers that are often intermingled by other repeats, like SINEs, LINEs, LTRs or β satellites. Interestingly, the sequence similarity shared by those monomers is much lower than that of the units within higher-order repeats. In addition, comparative sequence analyses reveal that the sequence of α satellite paralogues within higher-order repeats differs substantially less than α satellite orthologues among primates[45]. All of those observations, together with the fact that centromeres of "lower" primates consist of α satellite monomers, are the basis for the hypothesis that initial higher-order arrays of α satellites originated from the progenitor monomeric sequence, that was transposed and propagated in chromosomes of "higher" primates forming functional centromeres[45,46].

We have proposed snoRNAs as ancestors of human α satellites (Matylla-Kulinska *et al.*, unpublished). The predicted secondary structure of the consensus sequences of human α satellite families retrieved from the Dfam database[47], resembles the structure of H/ACA-snoRNAs. It contains 2 stems joined by an unstructured linker enclosing degenerated H- and ACA-boxes (Matylla-Kulinska *et al*., unpublished). The evolutionary most distant homologues to human α satellites were identified in marmosets[48]. The structure analysis of marmoset alphoid sequences revealed degenerated a snoRNA-like structure. Interestingly, the consensus fold comprises a 3' flank region similar to the one previously characterized in marsupial snoRNA-derived retrotransposon, snoRTEs[49]. SnoRTEs including H/ACA snoRNA combined with retrotransposon-like non-LTR transposable elements (RTEs) were reported to have an ability to insert into new genomic loci. In addition, dyskerin, which is a centromere binding factor 5 (Cbfp5) homologue and a core member of H/ACA snoRNPs, seems to be also involved in mitotic spindle formation and the spindle assembly checkpoint[50]. Our structural bioinformatic data together with above-mentioned observations points to snoRNAs as primary sequence origin for primate α satellites.

In the course of mutation accumulation, segment duplications and sequence conversion, α satellites lost a snoRNA-related function, but their centromeric location allowed them to acquire some new functions instead. It is well established that the centromere and the underlying DNA is important for: i) recognition and pairing of homologous chromosomes, ii) coupling of the sister chromatids during nuclear division, then either releasing the joint (during mitosis and second meiotic division) or retaining it (first part of meiosis), as well as iii) the spindle formation[51–53]. Moreover, α satellites function also on the RNA level, as the α satellite transcripts are crucial for proper localization of centromere-specific proteins CENP-C1 and INCENP[54]. Results obtained in our laboratory (Matylla-Kulinska *et al*., unpublished) indicate that α satellite-derived aptamers not only can bind to Pol II but can also serve as templates for RNA-dependent RNA polymerization and/or 3' extension, both catalyzed by RNA polymerase II. However, the function of this interaction needs to be further elucidated.]

[U1 snRNA evolved into spliced leader RNA multiple times

In addition to *cis*-splicing, *i.e.* the removal of introns from pre-mRNAs, some phylogenetically distant organisms employ *trans*-splicing during mRNA biogenesis. In *trans*-splicing, the 5' portion of a pre-mRNA is substituted with a spliced leader RNA (SL RNA), which is transcribed from a distinct genomic locus. As a consequence, many mRNAs (in some organisms all mRNAs) share a common 5' end (reviewed in[55]). *Trans*-splicing can have a multitude of functions, for example, processing of polycistronic pre-mRNAs into individual mature mRNAs, providing 5' cap structure and thereby stabilizing the transcript, and providing initiator AUG codon[55,56].

There is evidence that SL RNAs evolved from the repetitive spliceosomal U1 small nuclear RNAs (snRNAs). Both RNA classes possess a trimethylguanosine cap structure and Sm-binding site, they are often dispersed in arrays of 5S rDNA and the *trans*-splicing machinery utilizes other snRNA components of the major spliceosome except U1. Indeed, it has been shown that SL RNA can complement U1 loss in an *in vitro* splicing system[57]. These similarities made it possible for SL RNAs to evolve independently several times in distant eukaryotic species[58,59].

U1 and other snRNAs behave like transposable elements giving rise to large families of pseudogenes[60]. It has been suggested that some of the pseudogene families are in fact the ancestral form of U1 indicating that U1 itself is a ncRNA derived from repeat elements[61]. During the evolution of eukaryotes, some of the U1 elements invaded the 5S rDNA repeat unit and became a part of a large array[62,63]. SL RNAs might have evolved from these 5S rDNA-linked U1 elements but perhaps they retained the capability to transpose, since they have been found dispersed at other genomic loci as well. We envision that SL RNAs and U1 snRNAs still have the ability to give rise to functionally distinct RNAs, as some U1 paralogues have been shown to be differentially expressed and are reported to have tissue- and developmental stage-specific functions[64,65].]

[Cross-analysis between Dfam and Rfam implies many more examples of repRNAs

In order to investigate whether there are other ncRNAs derived from repeat elements, we took a systematic approach to assess sequence similarity between the repeat families found in Dfam[47] and the ncRNA families found in Rfam[66]. To this end Hidden Markov Models (HMMs) were generated from the seed alignments of the corresponding Dfam/Rfam entries as well as the MirBase miRNAs with the help of the HMMER packages[67]. These HMMs were then compared based on an own implementation of the algorithm published in[68] taking special interest in RNAs. The HMM-HMM comparison can be conceptualized as an alignment of HMM states. The corresponding scoring function takes into account the transition probabilities of the HMMs and the emission probabilities along the HMMs at the same time (see **Figure 3A**). This approach was chosen to improve the sensitivity and speed of the search as well as to facilitate the homology-scoring by returning a single score and significance-value for each HMM comparison. In order to assess the significance of the HMM comparison, a score distribution was computed for each Dfam HMM model. This was done by approximate dinucleotide shuffling 10 times the seed alignments used to generate the HMMs and generating the HMMs for each of the shuffled alignments, leading to a total of 11,320 HMMs. For each Dfam HMM, the score distribution was then fitted by a Gumbel extreme value distribution in order to compute the significance value directly from the HMM-HMM comparison score.

The outcome of our cross-analysis unambiguously shows that the strong similarity between ncRNAs and RNAs derived from human repeats is predominantly seen for miRNAs. From the 1,433 ncRNAs having a *p*-value smaller than 10^{-5}, a threshold that corresponds to the previously reported sequence similarity between the mir-325 family and the L2 repeats[69], 87 % (1,248) were related to miRNAs. The vast majority of the miRNAs are homologous to Alu elements (SINEs), followed by LINEs, DNA transposons and LTR as reviewed in[70]. Furthermore, we found a complete overlap between the 3' end of LFSINE_vert and uc_338 (ultraconserved element) confirming a previous report from[71,72], and high similarity between the central region of Plat_L3 and imprinted long ncRNA, KCNQ1DN. Our analysis also confirms reports on other homologies, such as BC200 and 7SL. Next, we scanned the genomes of mouse, platypus and chicken with a similar approach. In order to generate the repeat-HMMs, the RepeatMasker annotation of the corresponding genomes was downloaded from the UCSC genome browser[73] and used to generate alignment for each repeat family. These alignments were passed to HMMER in order to generate the repeat HMM. Similarly to results of the human analysis, the majority of the repRNAs from mouse, platypus and chicken are miRNAs derived from DNA repeats and LINE elements. In contrast, no similarity between SINE elements and uc_338 could be found. In the lizard *Anolis carolinensis*, however, similarity between uc_338 and LFSINE_vert was detected. We also identified mir-7641 as a derivative of rRNA repeats, as well as mir-763 and mir-1641, which derive from DNA repeats. For complete results, see http://alu.abc.univie.ac.at/reprna.]

[Searching for functions of repRNAs

The protein-coding parts of genomes are thoroughly investigated but very little attention is brought to the large quantity of sequences that are not unique and do not belong to the conventional concept of a gene. Poor interest in repetitive arrays arises in part from the following two reasons: they are considered to be “junk” or non-functional, and their repetitive nature hampers the computational annotation and analysis of those parts of the genome. Canonical genetic and biochemical methods cannot easily be applied to address the function of highly repetitive elements. Yet it became obvious that repeat regions are not silent, but differentially expressed in various states of the cells[74].

In order to look for repRNA functions, biochemical and bioinformatic approaches are necessary. We recently employed Genomic SELEX combined with deep sequencing as an unbiased approach to screen entire genomes for short functional RNA motifs that bind to specific ligands of choice[75]. It is feasible to examine whole genomes because RNA libraries used for this approach are transcribed *in vitro* from genomic DNA and hence contain all potentially functional domains encoded in a genome regardless of their expression levels. Importantly, in these genomic libraries the repeat-derived sequences are equally represented compared to genic sequences, making the approach especially suitable for the analysis of repRNAs. The limitation of SELEX screens is the choice of baits that are used to isolate the target RNAs. On the other hand, once a protein-RNA interaction is detected, the protein will deliver first hints on the functionality of the RNA.**]**

Conclusion

[We showed that repRNAs, derived from ncRNAs by retrotransposition and amplification, are a potent source of new functional RNAs. We illustrated the phenomenon with four examples but it is likely that there are more ncRNAs that evolved new functions after retrotransposition. Sequence conservation across species may suggest function. Thus additional repRNAs might be derived, for instance, from conserved SINE descendants, $4.5S_I$ and 4.5SH RNAs[76,77]. Similarly, interaction of ncRNA with a cellular protein might imply function, as can be the case of snaR family[78].

It is important to note that repRNAs (and thus the evolutionary reservoir) can arise by different mechanisms as exemplified by telomeric TERRA RNA. TERRA transcripts are products of RNA polymerase II, but the telomeric loci are produced by the telomerase enzyme, which solves the end-replication problem. Telomerases extend telomeric 3' ends through reverse transcription using short telomere RNA as template[79,80]. This template contains a short sequence, which is copied in a repetitive fashion leading to an array containing many short tandem repeats. Telomerase-like reverse transcription is an example of how long tandem repeats can originate.

Similarly, not only origins of repRNAs are diverse, so are newly evolved functions and mechanisms of action, which do not necessarily remain on the RNA level. For example, RNA polymerase III-transcribed genes are generally repetitive[81] and in many loci of various genomes, the coding sequence has been lost and "orphan" RNA polymerase III promoter elements play a role, for instance, in the regulation of RNA polymerase II transcription[82] and possibly also in chromosome organization[83]. Similarly, tRNA genes, a class of RNA polymerase III transcripts, have been shown to regulate expression of neighbouring RNA polymerase II genes[84] or act as chromatin insulators[85].

The evolution of new functions of repRNAs can be hindered by a process of concerted evolution in which gene conversion or unequal crossover lead to overwriting of a repeat with the sequence of its paralogue and the repeats are thereby homogenized in a given genome. The phenomenon is documented in repeats arranged in arrays, for instance in rDNA and α satellites[86,87], and is beneficial when a gene product is needed in great abundance, as is the case of rRNAs and histone mRNAs[88]. Nevertheless, whether other gene families undergo concerted evolution is questionable[89] and many of them clearly diverged to the point where gene conversion is no longer possible.

Repeat elements have long been ignored in genomic annotation and high-throughput data analyses. Nevertheless, this is changing due to the recognition of their importance for genomes and transcriptomes. We can therefore expect that many more functional repRNAs will be discovered in future research.] [A7]

Acknowledgments

[We wish to thank all members of the Schroeder lab for critically reviewing the manuscript. This work was funded by the Austrian Science Fund FWF grants number F4301 and F4308 and the GenAU Programme from the Autrian Ministry of Science.][A8]

References

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

2. Jason de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* 2011;7:e1002384.

3. Cordaux R, Hedges DJ, Herke SW, Batzer M. Estimating the retrotransposition rate of human Alu elements. *Gene* 2006;373:134–7.

4. McDonald JF. Transposable elements: possible catalysts of organismic evolution. *Trends Ecol Evol* 1995;10:123–6.

5. Kazazian HH. Mobile elements: drivers of genome evolution. *Science* 2004;303:1626–32.

6. Makałowski W. Genomic scrap yard: how genomes utilize all that junk. *Gene* 2000;259:61–7.

7. Kramerov DA, Vassetzky NS. SINEs. *Wiley Interdiscip Rev RNA* 2011;2:772–86.

8. Volff JN. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *BioEssays* 2006;28:913–22.

9. Brosius J. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* 1999;238:115–34.

10. Lander E, Linton L, Birren B, Nusbaum C, Zody M, Baldwin J, Devon K. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860–921.

11. Waring M, Britten RJ. Nucleotide sequence repetition: a rapidly reassociating fraction of mouse DNA. *Science* 1966;154:791–4.

12. Weber JL. Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. *Genomics* 1990;7:524–30.

13. Pheasant M, Mattick JS. Raising the estimate of functional human sequences. *Genome Res* 2007;17:1245–53.

14. Doolittle WF. Is junk DNA bunk? A critique of ENCODE. *Proc Natl Acad Sci U S A* 2013;110:5294–300.

15. Mattick JS, Dinger ME. The extent of functionality in the human genome. *Hugo J* 2013;7:2.

16. McDonald JF. Transposable elements: possible catalysts of organismic evolution. *Trends Ecol Evol* 1995;10:123–6.

17. Kazazian HH. Mobile elements: drivers of genome evolution. *Science (80-)* 2004;303:1626–32.

18. Ullu E, Tschudi C. Alu sequences are processed 7SL RNA genes. *Nature* 1984;312:171–2.

19. Quentin Y. Fusion of monomer a free left Alu monomer and a free right Alu at the origin of the Alu family in the primate. *Nucleic Acids Res* 1992;20:487–93.

20. Quentin Y. A master sequence related to a free left Alu monomer (FLAM) at the origin of the B1 family in rodent genomes. *Nucleic Acids Res* 1994;22:2222–7.

21. Jurka J. Evolutionary impact of human Alu repetitive elements. *Curr Opin Genet Dev* 2004;14:603–8.

22. Veniaminova NA, Vassetzky NS, Kramerov DA. B1 SINEs in different rodent families. *Genomics* 2007;89:678–86.

23. Liu WM, Chu WM, Choudary P V, Schmid CW. Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. *Nucleic Acids Res* 1995;23:1758–65.

24. Mariner PD, Walters RD, Espinoza C a, Drullinger LF, Wagner SD, Kugel JF, Goodrich J a. Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. *Mol Cell* 2008;29:499–509.

25. Sorek R, Ast G, Graur D. Alu -Containing Exons are Alternatively Spliced. *Genome Res* 2002;12:1060–7.

26. Singer SS, Männel DN, Hehlgans T, Brosius J, Schmitz J. From "junk" to gene: curriculum vitae of a primate receptor isoform gene. *J Mol Biol* 2004;341:883–6.

27. Kramerov D a, Vassetzky NS. Origin and evolution of SINEs in eukaryotic genomes. *Heredity (Edinb)* 2011;107:487–95.

28. Tsirigos A, Rigoutsos I. Alu and b1 repeats have been selectively retained in the upstream and intronic regions of genes of specific functional classes. *PLoS Comput Biol* 2009;5:e1000610.

29. Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran J V. Human L1 Retrotransposition: cis preference versus trans complementation. *Mol Cell Biol* 2001;21:1429–39.

30. Vassetzky NS, Kramerov D a. SINEBase: a database and tool for SINE analysis. *Nucleic Acids Res* 2013;41:D83–D89.

31. Okada N. SINEs. *Curr Opin Genet Dev* 1991;1:498–504.

32. Lin D, Pestova T V, Hellen CUT, Tiedge H. Translational control by a small RNA: dendritic BC1 RNA targets the eukaryotic initiation factor 4A helicase mechanism. *Mol Cell Biol* 2008;28:3008–19.

33. Taylor BA, Navin A, Skryabin B V, Brosius J. Localization of the mouse gene (Bc1) encoding neural BC1 RNA near the fibroblast growth factor 3 locus (Fgf3) on distal chromosome 7. *Genomics* 1997;44:153–4.

34. Martignetti JA, Brosius J. BC1 RNA: transcriptional analysis of a neural cell-specific RNA polymerase III transcript. *Mol Cell Biol* 1995;15:1642–50.

35. Kramerov DA, Grigoryan A, Ryskov A, Georgiev G. Long double-stranded sequences (dsRNA-B) of nuclear pre-mRNA consist of a few highly abundant classes of sequences: evidence from DNA cloning experiments. *Nucleic Acids Res* 1979;6:697–713.

36. Fornace AJ, Mitchell JB. Induction of B2 RNA polymerase III transcription by heat shock: enrichment for heat shock induced sequences in rodent cells by hybridization subtraction. *Nucleic Acids Res* 1986;14:5793–811.

37. Daniels GR, Deininger PL. Repeat sequence families derived from mammalian tRNA genes. *Nature* 1985;317:819–22.

38. Kramerov DA, Tillib S, Ryskov A, Georgiev G. Nucleotide sequence of small polyadenylated B2 RNA. *Nucleic Acids Res* 1985;13:6423–37.

39. Espinoza C, Allen T, Hieb A, Kugel JF, Goodrich JA. B2 RNA binds directly to RNA polymerase II to repress transcript synthesis. *Nat Struct Mol Biol* 2004;11:822–9.

40. Yakovchuk P, Goodrich JA, Kugel JF. B2 RNA and Alu RNA repress transcription by disrupting contacts between RNA polymerase II and promoter DNA within assembled complexes. *Proc Natl Acad Sci U S A* 2009;106:5569–74.

41. Espinoza CA, Goodrich JA, Kugel JF. Characterization of the structure , function , and mechanism of B2 RNA , an ncRNA repressor of RNA polymerase II transcription. *RNA* 2007;13:583–96.

42. Willard HF. Chromosome-specific organization of human alpha satellite DNA. *Am J Hum Genet* 1985;37:524–32.

43. Rudd MK, Schueler MG, Willard HF. Sequence Organization and Functional Annotation of Human Centromeres. *Cold Spring Harb Symp Quant Biol* 2003;68:141–50.

44. Schindelhauer D, Schwarz T. Evidence for a Fast , Intrachromosomal Conversion Mechanism From Mapping of Nucleotide Variants Within a Homogeneous -Satellite DNA Array. *Genome Res* 2002;12:1815–26.

45. Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF. Genomic and genetic definition of a functional human centromere. *Science (80-)* 2001;294:109–15.

46. Kazakov AE, Shepelev VA, Tumeneva IG, Alexandrov A, Yurov YB, Alexandrov IA. Interspersed repeats are found predominantly in the “old” α satellite families. *Genomics* 2003;82:619–27.

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

48. Shepelev VA, Alexandrov AA, Yurov YB, Alexandrov IA. 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* 2009;5:e1000641.

49. Schmitz J, Zemann A, Churakov G, Kuhl H, Grützner F, Reinhardt R, Brosius J. Retroposed SNOfall — A mammalian-wide comparison of platypus snoRNAs. *Genome Res* 2008;18:1005–10.

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

51. Pidoux AL, Allshire RC. Centromeres: getting a grip of chromosomes. *Curr Opin Cell Biol* 2000;12:308–19.

52. Csink A, Henikoff S. Something from nothing: the evolution and utility of satellite repeats. *Trends Genet* 1998;14:200–4.

53. Karpen GH, Allshire RC. The case for epigenetic effects on centromere identity and function. *Trends Genet* 1997;13:489–96.

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

55. Hastings KEM. SL trans-splicing: easy come or easy go? *Trends Genet* 2005;21:240–7.

56. Cheng G, Cohen L, Ndegwa D, Davis RE. The flatworm spliced leader 3'-terminal AUG as a translation initiator methionine. *J Biol Chem* 2006;281:733–43.

57. Bruzik JP, Steitz JA. Spliced leader RNA sequences can substitute for the essential 5′ end of U1 RNA during splicing in a mammalian in vitro system. *Cell* 1990;62:889–99.

58. Derelle R, Momose T, Manuel M, Da Silva C, Wincker P, Houliston E. Convergent origins and rapid evolution of spliced leader trans -splicing in Metazoa: Insights from the Ctenophora and Hydrozoa. *RNA* 2010;16:696–707.

59. Douris V, Telford MJ, Averof M. Evidence for multiple independent origins of trans-splicing in Metazoa. *Mol Biol Evol* 2010;27:684–93.

60. Marz M, Kirsten T, Stadler PF. Evolution of spliceosomal snRNA genes in metazoan animals. *J Mol Evol* 2008;67:594–607.

61. Bernstein LB, Manser T, Weiner AM. Human U1 small nuclear RNA genes: extensive conservation of flanking sequences suggests cycles of gene amplification and transposition. *Mol Cell Biol* 1985;5:2159–71.

62. Pelliccia F, Barzotti R, Bucciarelli E, Rocchi A. 5S ribosomal and *U1* small nuclear RNA genes: A new linkage type in the genome of a crustacean that has three different tandemly repeated units containing 5S ribosomal DNA sequences. *Genome* 2001;44:331–5.

63. Manchado M, Zuasti E, Cross I, Merlo A, Infante C, Rebordinos L. Molecular characterization and chromosomal mapping of the 5S rRNA gene in Solea

senegalensis : a new linkage to the U1 , U2 , and U5 small nuclear RNA genes. *Genome* 2006;49:79–86.

64. Sierra-Montes JM, Pereira-Simon S, Smail SS, Herrera RJ. The silk moth Bombyx mori U1 and U2 snRNA variants are differentially expressed. *Gene* 2005;352:127–36.

65. Kyriakopoulou C, Larsson P, Liu L, Schuster J, Soderbom F, Kirsebom LA, Virtanen A. U1-like snRNAs lacking complementarity to canonical 5' splice sites. *RNA* 2006;12:1603–11.

66. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. *Nucleic Acids Res* 2003;31:439–41.

67. Eddy SR. A new generation of homology search tools based on probabilistic inference. *Genome Inform* 2009;23:205–11.

68. Söding J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* 2005;21:951–60.

69. Smalheiser NR, Torvik VI. Mammalian microRNAs derived from genomic repeats. *Trends Genet* 2005;21:318–22.

70. Hadjiargyrou M, Delihas N. The Intertwining of Transposable Elements and Non-Coding RNAs. *Int J Mol Sci* 2013;14:13307–28.

71. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* 2006;441:87–90.

72. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. Ultraconserved elements in the human genome. *Science* 2004;304:1321–5.

73. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler AD. The Human Genome Browser at UCSC. *Genome Res* 2002;12:996–1006.

74. Ting DT, Lipson D, Paul S, Brannigan BW, Akhavanfard S, Coffman EJ, Contino G, Deshpande V, Iafrate AJ, Letovsky S, Rivera MN, Bardeesy N, et al. Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science* 2011;331:593–6.

75. Zimmermann B, Bilusic I, Lorenz C, Schroeder R. Genomic SELEX: a discovery tool for genomic aptamers. *Methods* 2010;52:125–32.

76. Gogolevskaya IK, Kramerov DA. Evolutionary History of 4.5SI RNA and Indication That It Is Functional. *J Mol Evol* 2002;54:354–64.

77. Gogolevskaya IK, Koval AP, Kramerov D a. Evolutionary history of 4.5SH RNA. *Mol Biol Evol* 2005;22:1546–54.

78. Parrott AM, Mathews MB. snaR genes: recent descendants of Alu involved in the evolution of chorionic gonadotropins. *Cold Spring Harb Symp Quant Biol* 2009;74:363–73.

79. Cech TR. Beginning to understand the end of the chromosome. *Cell* 2004;116:273–9.

80. Blackburn EH, Greider CW, Szostak JW. Telomeres and telomerase: the path from maize, Tetrahymena and yeast to human cancer and aging. *Nat Med* 2006;12:1133–8.

81. Canella D, Praz V, Reina JH, Cousin P, Hernandez N. Defining the RNA polymerase III transcriptome: Genome-wide localization of the RNA polymerase III transcription machinery in human cells. *Genome Res* 2010;20:710–21.

82. Kleinschmidt RA, LeBlanc KE, Donze D. Autoregulation of an RNA polymerase II promoter by the RNA polymerase III transcription factor III C (TF(III)C) complex. *Proc Natl Acad Sci U S A* 2011;108:8385–9.

83. Moqtaderi Z, Wang J, Raha D, White RoJ, Snyder M, Weng Z, Struhl K. Genomic binding profiles of functionally distinct RNA Polymerase III Transcription Complexes in Human Cells. *Nat Struct Mol Biol* 2010;17:635–40.

84. Hull MW, Erickson J, Johnston M, Engelke DR. tRNA Genes as Transcriptional Repressor Elements. *Mol Cell Biol* 1994;14:1266–77.

85. Raab JR, Chiu J, Zhu J, Katzman S, Kurukuti S, Wade P a, Haussler D, Kamakaka RT. Human tRNA genes function as chromatin insulators. *EMBO J* 2012;31:330–50.

86. Drouin G, de Sá MM. The concerted evolution of 5S ribosomal genes linked to the repeat units of other multigene families. *Mol Biol Evol* 1995;12:481–93.

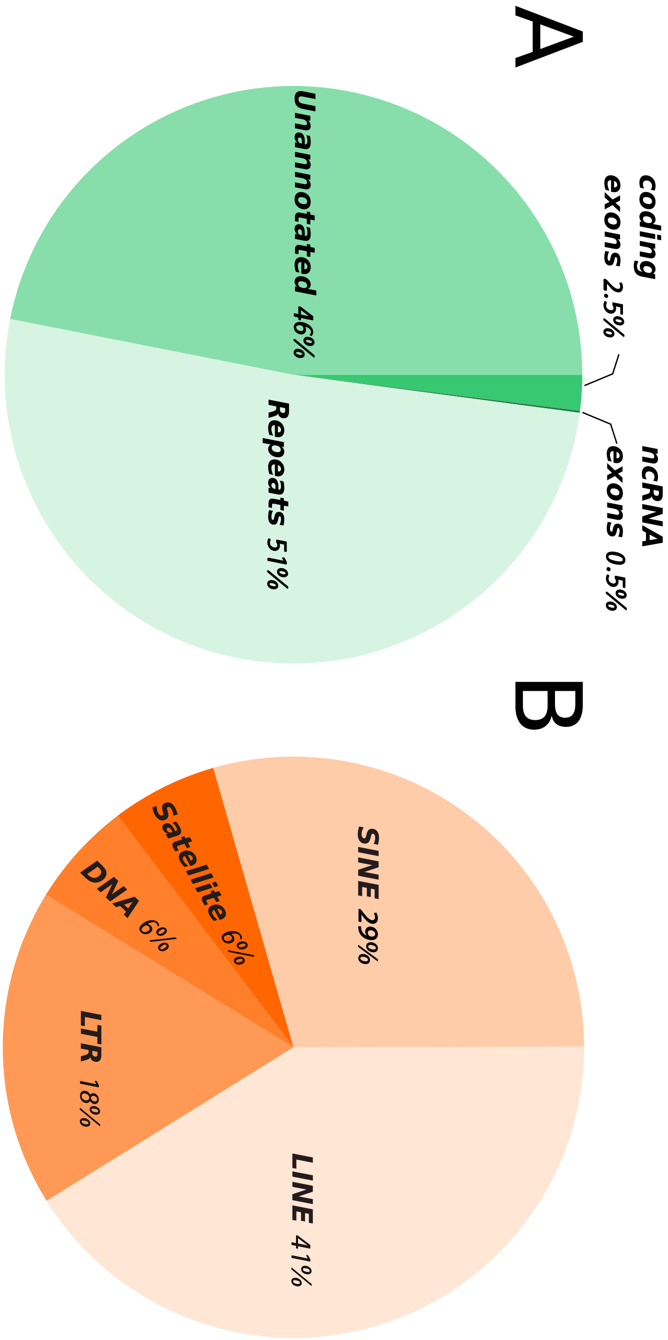
87. Durfy SJ, Willard HF. Concerted evolution of primate alpha satellite DNA. *J Mol Biol* 1990;216:555–66.

88. Innan H. Population genetic models of duplicated genes. *Genetica* 2009;137:19–37.

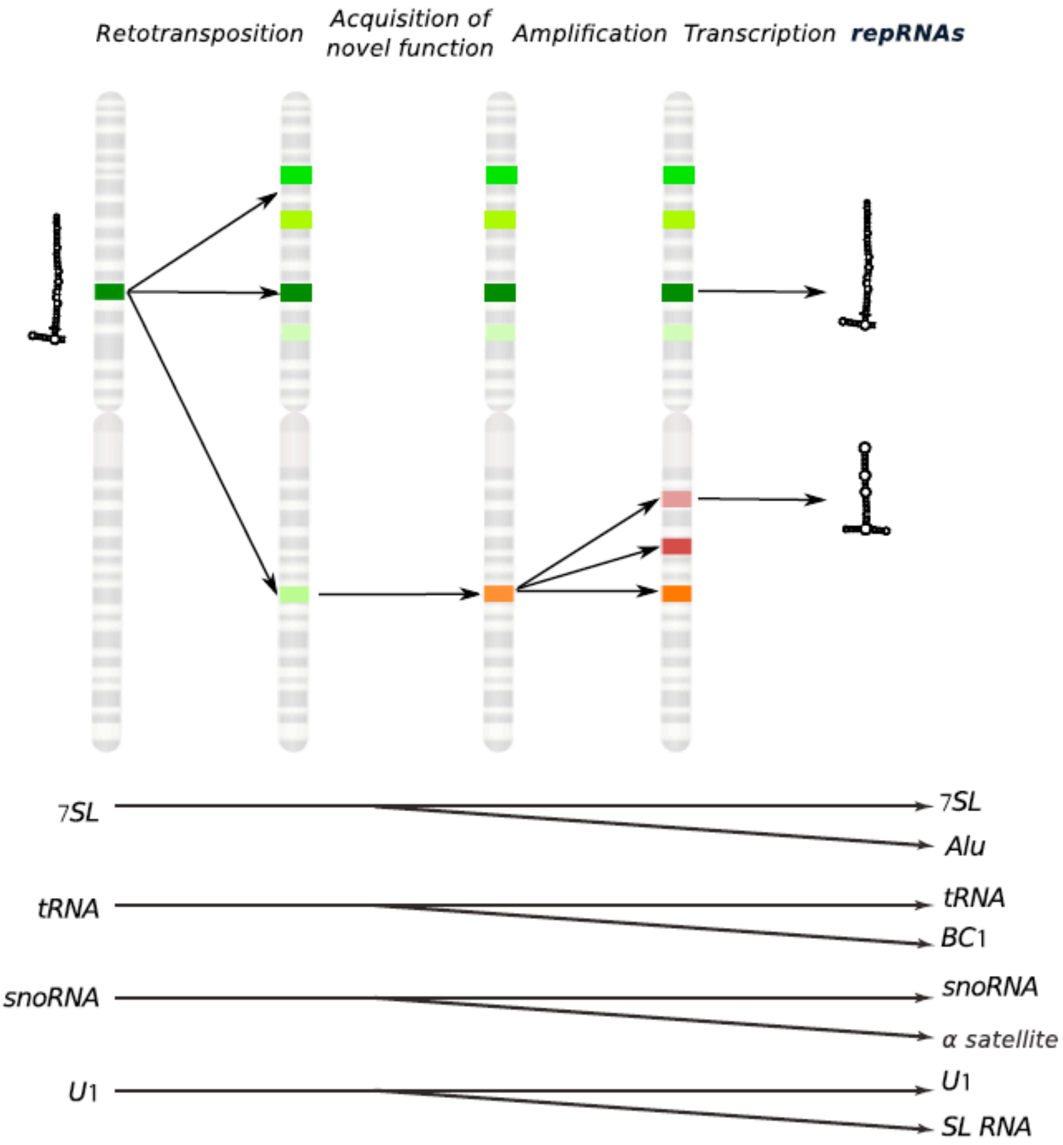
89. Nei M, Rooney AP. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet* 2005;39:121–52.

][A9]

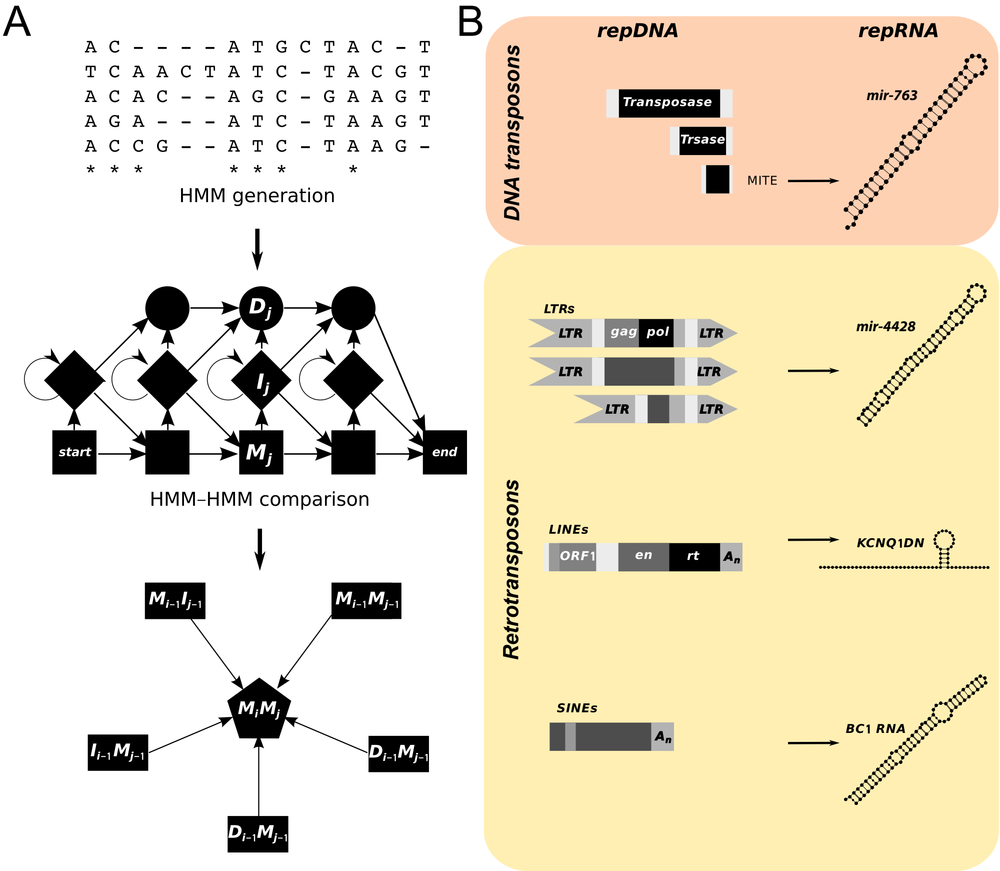
Figure captions



[**Figure 1**. **Human genome is repetitive. A,** Composition of the human genome. 2.5 % and 0.5 % of the human genome is covered with coding exons and ncRNA exons, respectively. Repeats represent 51 % of the genome while the unannotated regions amount to 46 % of the genome. **B,** Composition of the repetitive portion of the human genome. Repeats with the largest genome coverage are LINEs (41 %), followed by SINEs (29 %), LTRs (18 %), DNA transposons (6 %) and satellite repeats (6 %).] [A10]



[**Figure 2**. **repRNAs often originate from retrotransposed ncRNAs**. **Top**, Upon retrotransposition, ncRNAs are highly amplified and as they spread throughout the genome, they diversify in sequence (depicted as bands of different shade of the same color). Some copies evolve new functions (depicted as a band with a changed color) giving rise to new classes of repRNAs. Therefore non-coding transcripts derived from highly repetitive regions can be a rich reservoir for the evolution of novel functional RNAs. **Bottom**, Examples of repRNAs and their corresponding ancestor mastergenes. For detailed discussion, see text.]
[A11]



[**Figure 3.** Comparison of Dfam with Rfam reveals new relationships between repeat elements and ncRNAs. **A**, For each repeat and ncRNA family found in Dfam and Rfam, respectively, an HMM was constructed based on the corresponding seed alignments. These HMMs were then compared by literally aligning the states of both HMMs using dynamic programming. The best state alignment ending with the alignment of match state M_i and M_j can be obtained either from $M_{i-1}M_{j-1}$, $D_{i-1}M_{j-1}$, $M_{i-1}D_{j-1}$, $M_{i-1}I_{j-1}$ or $I_{i-1}M_{j-1}$. **B**, Examples of novel relationships between repeat elements and ncRNAs. mir-763 shows strong similarity with a MITE, mir-4428 derives from LTR repeats. KCNQ1DN ncRNA is highly homologous to LINE elements.] [A12]

MATYLLA-KULINSKA K., Boots, JL., Zimmermann, B., Schroeder, R. (2012)
Finding aptamers and small ribozymes in unexpected places.
Wiley Interdiscip Rev RNA 3: 73-91

I contributed to this publication in discussing the content as well as in writing the following sections: Discovery of aptamers, discovery of riboswitches.

Advanced Review

Finding aptamers and small ribozymes in unexpected places

Katarzyna Matylla-Kulinska, Jennifer L. Boots, Bob Zimmermann and Renée Schroeder*

The discovery of the catalytic properties of RNAs was a milestone for our view of how life emerged and forced us to reformulate many of our dogmas. The urge to grasp the whole spectrum of potential activities of RNA molecules stimulated two decades of fervent research resulting in a deep understanding of RNA-based phenomena. Most ribozymes were discovered by serendipity during the analysis of chemical processes, whereas RNA aptamers were identified through meticulous design and selection even before their discovery in nature. The desire to obtain aptamers led to the development of sophisticated technology and the design of efficient strategies. With the new notion that transcriptomes cover a major part of genomes and determine the identity of cells, it is reasonable to speculate that many more aptamers and ribozymes are awaiting their discovery in unexpected places. Now, in the genomic era with the development of powerful bioinformatics and sequencing methods, we are overwhelmed with tools for studying the genomes of all living and possibly even extinct organisms. Genomic SELEX (systematic evolution of ligands by exponential enrichment) coupled with deep sequencing and sophisticated computational analysis not only gives access to unexplored parts of sequenced genomes but also allows screening metagenomes in an unbiased manner. © 2011 John Wiley & Sons, Ltd. *WIREs RNA* 2011 DOI: 10.1002/wrna.105

INTRODUCTION

The repertoire of nonprotein coding RNAs with different functions is growing steadily and has surprised many of us in the last decades. As transcriptome analyses are hitting transcripts derived from almost every part of the genome, the notion that most if not all regions are transcribed is more appreciated.[1] This immediately leads us to the question of whether these transcripts are just ‘noise’ or ‘junk’ or whether they have functions.[2] These RNAs may differ in many aspects from what we know until now, and therefore, it is not easy to search for their function in a systematic way.

Ribozymes are non-coding RNAs that catalyze chemical reactions and they are extremely diverse in size, sequences, and shape.[3] They can be as complex as the large ribosomal subunit RNA that catalyzes peptide bond formation or as simple as the artificial leadzyme. Only very few naturally occurring catalytic RNAs are known, but there is growing evidence that the genomes are full of ribozymes awaiting discovery. The repertoire of known ribozymes was significantly enriched by the synthetic ribozymes isolated via SELEX (systematic evolution of ligands by exponential enrichment).[4] The aim of these efforts was to obtain all ribozymes necessary to sustain an RNA-based metabolism during the RNA world. A different situation occurred for RNA aptamers. They were invented and applied by scientists before they were discovered in nature.[5,6] They are defined as short RNA sequences that can fold into specific structures forming pockets that can accommodate specific ligands. The notion that RNA can fold into selective pockets arose with the discovery that guanosine was a cofactor for group I intron splicing.[7] Soon after this observation it was found that many other small molecules can be bound by RNAs, like arginine or antibiotics.[8,9] Many aptamers were subsequently selected for binding small metabolites and cofactors.[10] Only much later, in 2002, was it realized that nature had evolved many aptamers binding small metabolites to regulate gene expression. These aptamers, termed

*Correspondence to: renee.schroeder@univie.ac.at

Department of Biochemistry and Molecular Cell Biology, Max F. Perutz Laboratories, University of Vienna, Vienna, Austria

DOI: 10.1002/wrna.105

riboswitches, are highly abundant in several bacterial genomes.[11–13]

We assume that we know a small percentage of all living organisms and for only very few have the genomes been sequenced. With new technologies evolving at an incredible speed, the spectrum of functional RNAs will become more accessible. To mine all this new information, alternative, unbiased strategies will be needed. In our point of view, genomic SELEX coupled with high-throughput sequencing and bioinformatic analysis are very powerful tools to explore this still untouched genomic space. In this review, we want to recapitulate how two special classes of noncoding RNAs have been discovered, aptamers and small ribozymes, and how more of those classes are being detected.

DISCOVERY OF ARTIFICIAL APTAMERS

Aptamers are small single-stranded DNA or RNA molecules with high affinity to their targets. They bind a wide variety of molecules such as peptides,[14] proteins,[15] nucleic acids,[16] inorganic components,[17] small organic compounds,[18] antibiotics[19] as well as viral particles[20] or entire cells.[21] Dissociation constants (K_d) for aptamer complexes lie in a range of 10^{-11}–10^{-9} M for proteins, or 10^{-7}–10^{-6} M for small molecules. Aptamers interact specifically with their target molecule. As examples, the aptamer for the reverse transcriptase of human immunodeficiency virus 1 (HIV-1) is able to distinguish protein partners that differ by a single amino acid substitution[22] and the theophylline aptamer discriminates between theophylline and caffeine, which differ by a single methyl group.[23] Aptamers—derived from the Latin word: aptus-adjusted, and Greek: meros-particle—fold into unique tertiary structures, in which the ligand often becomes an intrinsic part of the RNA architecture. Interaction with the target results mainly from base aromatic rings stacking interactions and hydrogen bonding (for review see Ref 10).

Aptamers were originally identified using SELEX[24,25] (Figure 1). Briefly, SELEX involves the incubation of a synthetic single-stranded DNA or RNA library pool with the target molecule of interest. After partitioning of complexes from non-bound

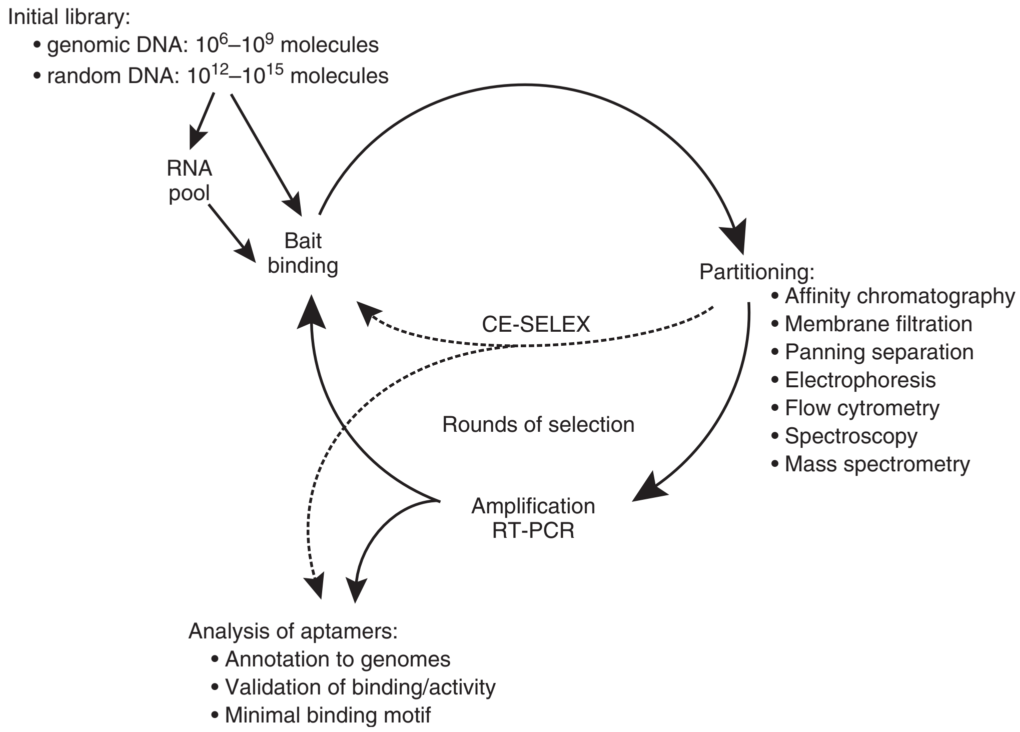


FIGURE 1 | Varieties of systematic evolution of ligands by exponential enrichment (SELEX). At its core, SELEX denotes a cyclical evolutionary screen for sequences conferring a specific activity, typically binding, as is in the case for aptamers. Choice of initial library depends on the application. When searching for artificial aptamers, a short, random section of nucleotides are flanked by fixed sequences. Genomic SELEX enables the screening of genomes for aptamers, and has the added benefit of reducing the complexity of the library. Several different methods can be used for partitioning, and it is informative to use multiple methods and perform technical duplicates in parallel. SELEX for RNAs usually requires rounds of transcription and amplification for pool maintenance. However, capillary electrophoresis SELEX bypasses these extra steps. High-throughput sequencing and bioinformatic analysis follow once sufficient enrichment is detected.

aptamers, the remaining sequences are amplified. Repeating those steps leads to a gradual enrichment of the library with sequences that specifically recognize the target. When the pool converges on a collection of sequences with high affinity for the target (usually after 8–15 cycles), it is cloned and sequenced. This approach is suitable not only for screening for high-affinity aptamers but also for catalytic activity (ribozymes and DNAzymes). We define aptamers derived from synthetic libraries of random sequences as artificial aptamers and ones obtained from genomic libraries as natural genomic aptamers.

SELEX Libraries

The starting library is an important determinant for the selection procedure. During selection of artificial aptamers, in classical SELEX, it is a chemically synthesized, combinatorial DNA pool, which can be transcribed into RNA.[24,25] The promoter sequence for T7 bacteriophage RNA polymerase is within the 5′ end fixed region of the single-stranded DNA library, and the pool is *in vitro* transcribed during every round of selection. The content of a classical, random library is on average 10^{13}–10^{15} molecules. There are also only partially randomized libraries based on known sequence constraints. The selection from those so-called doped libraries is carried out to discriminate nucleotides crucial for the recognition and binding to the target.[26] In a tailored version of SELEX that aims to minimize the aptamer size, libraries are free of fixed regions. The random sequences are flanked by a few fixed bases that are removed just after each amplification round.[27]

Partitioning or Selection

Efficient partitioning of complexes from unbound oligonucleotides is a key step in every SELEX procedure. Recovery of aptamer–bait complexes can be achieved mainly by performing affinity chromatography, membrane filtration, panning separation, or electrophoresis. Characteristics of commonly used partitioning methods with references are summarized in Table 1.

TABLE 1 | Characteristics of Commonly Used Partitioning Methods

Partitioning Method	Characteristic	References
Affinity chromatography	The target is immobilized on a sorbent	
Affinity columns (sepharose, agarose)	• requires large amount of target	24,28,29
Affinity beads	• requires only small amounts of target • convenient to handle	16,17,19
Titer plates (His-tag, GST, biotin etc)	• covalent cross-linking surface	30
Membrane filtration	The target-aptamer complex is formed in the solution	
Retention on nitrocellulose membrane	• nitrocellulose filter interacts nonspecifically with proteins • available filters with distinct molecular weight cutoff • suitable for RNA targets • quick and simple	15,20,25,26,31
Panning separation	When used targets are macromolecules (such as cells, viruses)	21
Electrophoresis		
Gel electrophoresis	• mobility shift assay under non-denaturing condition, • nucleic acid aptamer is recovered using crush- and- soak method	32
Capillary electrophoresis	• suitable for smaller targets • high-resolution capacity • minimal sample dilution	28,33
Flow cytometry	The complex is separated based on its fluorescence	
Fluorescence-activated cell sorting	• applied in cell-SELEX • target cell is fluorescently labeled • simultaneous isolation of cells of interests and removal of unbound aptamers	34
Spectroscopy		
Surface plasmon resonance	• allows real-time monitoring of complexes • provide binding efficiency information	35

Initial rounds of selection require less stringent conditions and longer incubation times. However, to obtain high-affinity binders, the selection of later cycles is usually performed under higher stringency, such as changing buffer conditions, the addition of mono- or divalent ions, increasing or decreasing aptamers to target ratio, or supplementing with nonspecific competitors. It is also possible to use ultraviolet (UV) cross-linking during selection. The cross-linking not only stabilizes the RNA/protein complex and makes it more resistant to extensive washes, but also reduces the nonspecific complex formation. For the RNA SELEX against HIV-1 Rev, the RNA pool was transcribed in the presence of the chromophore 5-iodouracil. The pool of photoreactive aptamers was incubated with the protein Rev as bait, and irradiated with UV. Stable complexes were then recovered using nitrocellulose retention.[31]

Amplification

After every round of selection, the binding oligonucleotides are amplified in order to maintain sufficient amounts of material. It is very important to note that the amplification steps [polymerase chain reaction (PCR), *in vitro* transcription, reverse transcription] may introduce some bias that leads to an imprecise selection. In our laboratory, we proposed a parallel control to SELEX that evaluates the amplification impact on the initial library,[36] whereby the selection steps are omitted from the SELEX cycle. Each round of so-called neutral SELEX performed was sequenced. The average sequence had a less stable structure as the cycles progressed. This effect is most likely caused by the reverse transcriptase having difficulty denaturing highly structured RNAs. While the effect is clear when no selection step is present, the selective pressures of binding can still lead to the isolation of a highly structured RNA. An example can be seen in the SELEX-derived streptomycin aptamer that was later crystallized and shown to have a stable structure even outside of the binding domain.[37] Other characteristic biases tested via the neutral SELEX such as length, nucleotide content, and divergence from the initial library were only mildly affected; however, this can vary depending on the features of the initial library. Therefore, it is advisable to perform a neutral SELEX control to any genomic SELEX in order to analyze the background signal.

Noncyclic *In Vitro* Selections

Typically, to get tightly binding aptamers, 8–15 amplification rounds with conventional partitioning methods are required. High-resolution separation enables fewer SELEX cycles resulting in less pronounced amplification bias, greater heterogeneity of the pool and reduction of time required for efficient pool enrichment.[33] Capillary electrophoresis is a high-resolution separation method[38] that takes advantage from a mobility shift after complex formation due to the changes in size and charge. Tang and colleagues isolated aptamers for ricin in parallel experiments using conventional SELEX with affinity chromatography and SELEX coupled with capillary electrophoresis (CE-SELEX). After four rounds of CE-SELEX 87.2% of the pool bound to ricin, whereas after nine cycles of conventional SELEX only 38.5% of oligonucleotides did. This work clearly proves that CE-SELEX is a very efficient method for aptamer isolation.[28]

Another approach to aptamer selection is the use of microarrays (Box 1).[39] Owing to their low capacity, the libraries are much less complex than that of SELEX. Chushak and Stone developed an *in silico* method to preselect aptamers for microarray analysis, proving the principle by verifying it on a set of six known aptamer–ligand complexes.[40] This approach has the advantage that, unlike SELEX, it favors structurally more stable sequences, however, it has yet to be used on novel aptamers.

SERENDIPITOUS DISCOVERY OF NATURAL APTAMERS–RIBOSWITCHES

For a long time, the mechanism for regulation of operons from many bacterial metabolic pathways had remained a mystery. In 1998, in the process of elucidating this regulation, Grundy and Henkin discovered the S-box domain, which proved to be a significant finding. They observed that this highly conserved motif within the 5′ end of a leader sequence of genes is involved in the biosynthesis of methionine and cysteine in *Bacillus subtilis* genome. After assessing the covariance (Box 2) in the conserved residues, Grundy and Henkin proposed that the leader sequence could fold into two alternative, mutually exclusive structures: a terminator and an antiterminator. Additionally, the upstream anti-antiterminator stem-loop structure was predicted to compete with the antiterminator formation. The S-box was hypothesized to serve as a binding-site for regulatory factors that respond to methionine levels within the cell and stabilize the anti-antiterminator structure.[43] Many groups were unsuccessfully trying to identify protein factors recognizing specific metabolites and thereby regulating expression of genes from this specific pathway. Finally, growing evidence that RNAs can serve as allosteric binders of specific effector molecules led to the idea that mRNAs might bind metabolites directly and affect their own regulation.

BOX 1

HIGH-THROUGHPUT METHODS

1. A *microarray* is a multiplex screening method based on the hybridization between the arrayed series of probes (DNA or RNA oligonucleotides representing part of known genes or transcripts) immobilized on to a solid surface and the labeled target molecules under high-stringency conditions. Hybridization signal is detected and quantified usually by fluorophore-, silver- or chemiluminescence.
2. A *tiling array* is a high-resolution subtype of the microarray chips using the same principle. The difference lies within the used probes that in case of the tiling array are designed to cover the entire genome or contiguous part of it.
3. *High-throughput sequencing* is the use of any number of newer methods to perform massively parallel sequencing of a population of nucleic acid molecules. The technology has been progressing at a staggering rate. Currently, it is quite inexpensive to sequence tens of millions of DNA sequences between 36 and 100 bases long, as well as perform *in situ* reversal to sequence both ends. RNA sequencing is currently on the horizon and promises to eliminate some of biases inherent in generating cDNA. For a historical perspective, see Ref 41.
4. *RNA genomics (RNomics)* According to the ENCODE project the vast majority of the genome is transcribed.[42] Most of the transcripts are nonprotein coding. RNomics is a combined experimental and computational approach to understand function of non-coding RNAs and their interactions at a genomic level. Experimental approaches encompass a broad range of RNA and cDNA preparation followed by high-throughput sequencing. As the procedure starts with RNA isolation, the outcome reflects the transcriptional output of a defined cell state.

In 2002, both the Breaker[13] and Nudler[11] laboratories reported the discovery of riboswitches. Using in-line probing, Nahvi and coworkers showed that *btuB* mRNA from *Escherichia coli* binds coenzyme B_{12} (adenosylcobalamin—AdoCbl) and stabilizes a conformational change within the mRNA. Moreover, they performed an equilibrium dialysis assay with labeled ^{3}H-AdoCbl and found that an excess of AdoCbl analogs could not compete for binding. Therefore, they concluded that the *btuB* forms a selective binding pocket for AdoCbl.[13] Independently, Mironov and colleagues described a transcriptional attenuation mechanism that controls riboflavin and thiamin synthesis in *B. subtilis*. After secondary structure analysis, they noticed that the riboflavin operon (*rib*) leader sequence consists of two stable, mutually exclusive structures: the classical hairpin terminator and the antiterminator that has a potential to base pair with the terminator hairpin. Observations that the *rib*-leader sequence is well conserved among bacteria, and that all mutations described to increase riboflavin production localize to this conserved region, strengthen the hypothesis that the termination/antitermination switch regulates *rib* expression. Mironov and coworkers constructed chromosomal *rib*-leader-*lacZ* transcriptional fusions to monitor the effect of the flavin mononucleotide (FMN) on *rib* expression. Comparison of the β-Gal activity between the wild type and mutated strains showed that FMN suppresses the *rib* operon expression. Deletion of the putative *rib*-leader terminator resulted in 20-fold increase in the operon expression. These results clearly indicate that flavins regulate *rib* operon by influencing the termination process. Next, by tracing premature termination in a single round of run-off transcription on wild type and mutated *rib*-leader templates Mironov and coworkers showed that FMN is a specific termination factor that interacts with the *rib*-leader sequence and thus stabilizes the structure that induces termination. In the same paper, Mironov and colleagues describe a similar transcription attenuation mechanism for the *B. subtilis* thiamin operon and speculate that regulation of transcription by the anti-antiterminator/antiterminator structure formation upon metabolite binding is a common mechanism.[11]

These serendipitous discoveries were followed by many reports of other riboswitches. The majority of examples come from bacteria, but they have also been described in archaea, fungi and plants.[48] Typically, riboswitches consist of an aptamer domain that folds into a stable structure upon ligand binding and an expression platform located outside the aptamer. Commonly, riboswitches respond to one metabolite. However, there are examples for tandem regulators, such as the 5′ UTR of *metE* from *Bacillus clausii* containing riboswitches for *S*-adenosylmethionine and the coenzyme B_{12} that can independently repress *metE*,[49] and the glycine riboswitch that binds cooperatively two ligands to regulate a single expression platform.[50] Most riboswitches are located at the 5′ UTR of

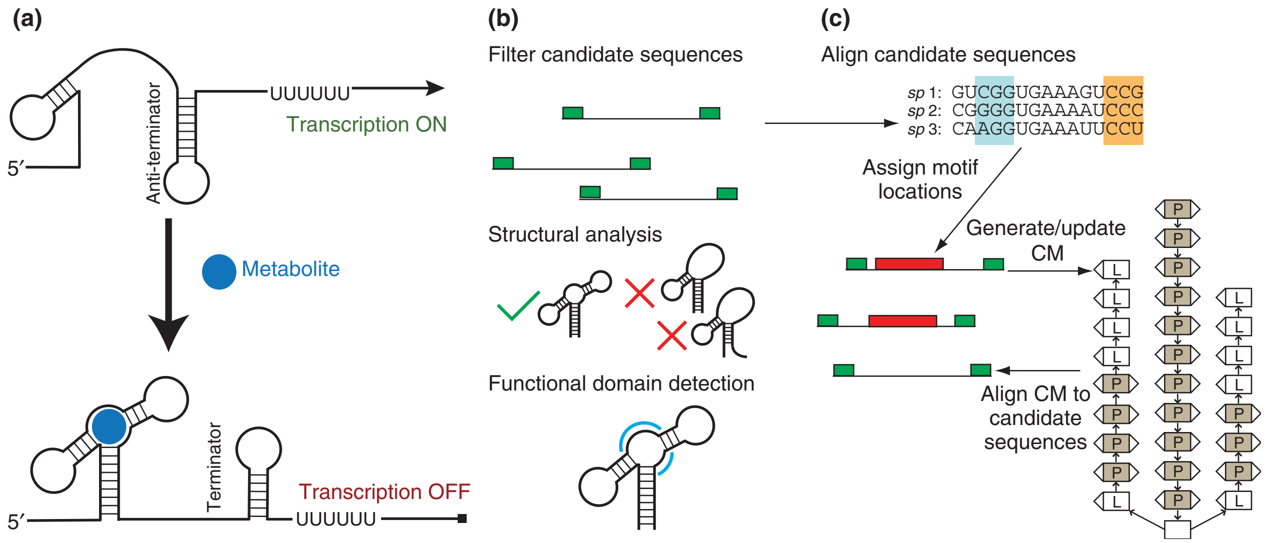


FIGURE 2 | Screening for novel riboswitches. Riboswitches are the elements of RNAs that undergo conformational change upon interaction with a specific ligand. They can be involved in transcriptional regulation as shown in (a). The binding of a metabolite induces a conformational change that allows the RNA to form a terminator structure, thereby aborting transcription short of the downstream open reading frame. Transcriptional regulation is one of the many modes of activity for riboswitches. (b) Many computational approaches to identify riboswitches are tailored specifically to detect features of known riboswitches. A reduced set of candidates is chosen based on the locations relative to an open reading frame (e.g., 5′ UTR, intergenic region on polycistronic gene) and targeted gene function (e.g., metabolite biosynthesis). Next, the candidates are selected for the necessary structure, in this case, a three-way junction. Finally, the functional domain (blue) is identified based on the most highly conserved nucleotides of riboswitch being matched. In some approaches, the functional region is detected before examining the structure. These approaches require knowledge of previously discovered riboswitches, whereas (c) detecting novel riboswitches, as is done with the CMFinder pipeline,[44–47] involves the iterative refinement of alignments of homologous regions. Initially, candidate sequences are aligned, followed by assignment of the motif locations. A consensus secondary structure among the sequences is then predicted and used to generate a covariance model (CM). This CM motif profile contains states which model paired regions (‘P’) and single nucleotide regions (‘L’eft and ‘R’ight). Insertions and deletions in the form of bulges are more easily allowed when considering structure in the alignment. The profile can then be aligned back to the candidates to generate a structure-aware estimate of the motif positions, which can then be used to improve the model itself. Iterative updates to the model halt after the updates ‘converge’, or show no sign of major changes.

mRNAs of genes involved in biosynthesis of the sensed metabolite. However, they were also discovered in the 3′ UTRs or within introns in fungi and plants.[51–54] The large collection of examples from three domains of life led to description of different modes of riboswitch function. The most common mechanisms include: (1) formation of the hairpin structure that leads to RNA polymerase stalling and premature transcription termination, (2) base-pairing between Shine-Dalgarno and anti-Shine-Dalgarno sequence that blocks translation, and (3) changing the splice sites.[53,55] There are also some very interesting rare mechanisms. An interesting example is the *glmS* riboswitch that combines ligand binding with a ribozyme activity[56] (discussed in Section *Discovery of Ribozymes that Control Gene Expression)*. It has been speculated that metabolite sensing RNAs are relics from an ancient RNA world, before proteins evolved.[57] The fact of pervasiveness of this mechanism speaks for it. Mandal and Breaker[58] estimated that >2% of all *B. subtilis* genes are regulated by riboswitches.

Biochemical methods for finding new riboswitches are limited. SELEX with immobilized ligands, a protocol well established for aptamer screening, is often not applicable for riboswitches because of difficulties in the immobilization of small ligands. Many structural studies have shown that metabolites are entirely engulfed by the RNA structure, thus it is not possible to immobilize the ligand via a linker. SELEX in solution may be an alternative choice. Sophisticated methods for SELEX in solution still need to be developed. Dilution approaches with oil drops or arrays coupled with mass spectrometry using stable isotopes for detection of the ligand–aptamer complex might be a promising approach.

Some riboswitches were also found unexpectedly, while doing RNomics in *Staphylococcus aureus*[59] or by tiling array on the total transcriptome from *Listeria monocytogenes*.[60,61] Bohn and colleagues noticed a group of prematurely terminated transcripts that were likely to be regulated by riboswitches. Structural probing revealed that they are *S*-adenosyl-methionine (SAM) riboswitches.[59] In a whole transcriptome analysis, Toledo-Arana and colleagues serendipitously identified SAM riboswitch elements that were standing out because of their different stability.[60]

Bioinformatic Approaches to Riboswitch Discovery

As riboswitch features are highly conserved, the predominant way to identify new members of known classes is through comparative sequence analysis on sequenced genomes and metagenomes (Figure 2). The common structural motifs in families of riboswitches have been exploited to search genomes for new ones. Elements near homologous riboswitch-regulated genes can be input to algorithms such as Riboswitch Finder[62] and RiboSW.[63] These programs take an RNA sequence as the input and match t.he consensus of riboswitch families to candidate sequences, followed by secondary structure analysis to predict riboswitch homologs. In particular, RiboSW which analyzes input against 12 families of known riboswitches, was shown to have very high sensitivity and is available as a web service. A more general approach to describing families of RNAs is the use of CMs (Box 2). Many riboswitch families are also described in the form of CMs,[64] and are available on Rfam.[65] The INFERNAL package[66] can be used to match these descriptions to novel, putative riboswitches, as well as create new CMs.[48]

A more difficult bioinformatic problem is the identification of novel riboswitch structural motifs. Because RNA structure prediction and comparison are time consuming to compute, the genomic features must be taken into account to filter for regions that are likely to contain riboswitches. RibEx uses a combination of the motif finders MEME and MAST, which are typically used for the identification of transcription factor binding sites, on filtered data in an iterative refinement to find novel riboswitches.[67,68] In another approach, Yao and colleagues developed the CMFinder algorithm, which refines alignments of RNAs to find a significant CM in a group of putatively related sequences.[44] The method discovers novel motifs by refining the CMs and resulting alignments progressively. As it is time consuming to compute, a pipeline was developed to filter sequences based on genomic context and sequence similarity seeds. The method was applied to 5′ UTR sequences of conserved sequences in bacteria, and successfully identified two confirmed and four putative novel riboswitch candidates.[45,46]

In some instances, bioinformatics is limited in its ability to predict riboswitches. This applies, e.g., to AdoCbl riboswitches that upon binding of metabolite form pseudoknot structures. Pseudoknots are particularly difficult to predict computationally, because the number of possible structures including pseudoknots is exponentially greater than nested secondary structures, and determining an optimal pseudoknotted fold of any given structure is computationally infeasible.[69]

BOX 2

COVARIATION AND COVARIANCE MODELS (CMs)

Functional RNA molecules often are more constrained in the conservation of secondary structure than primary sequence. Secondary structure alignments need to be based on a strong pairwise correlation of Watson-Crick base pair complementarity. This fact is taken into account in algorithms using *covariation* analysis, which is finding bases with variation between species that preserve secondary structure. Incidence of covariation is strong evidence for the correct structure.

Covariance models (CMs) are probabilistic models that enable a structurally motivated alignment of an RNA to a profile (for an example, see Figure 2). The profile (the CM itself) describes the consensus sequence and structure, which is used to guide the alignment and detection of covariation. The alignment algorithm is thus structure-aware, allowing the scoring for insertions and deletions to depend on where they fall within the proposed structure. For example, a double-stranded bulge or extension of a stem can be scored as an insertion in both strands simultaneously, even though several bases may separate the two strands in the primary sequence. Additionally, covariance itself can be described in each of the paired components of the structure. The alignment takes much more time to compute than traditional alignments, and cannot be used directly for scanning genomes in a reasonable amount of time without the use of heuristics (informed strategies) to reduce the number of candidate sequences.

Although many computationally predicted riboswitches have been validated through biochemical analysis, there are still some ‘orphan riboswitches’ missing their corresponding ligands. Ligands for candidate riboswitches are typically deduced according to the mRNA located downstream to the identified aptamer domain and confirmed in a series of *in vitro* binding studies and functional assays. The challenges of validation and ligand identification for riboswitch candidates were well discussed by Meyer and colleagues.[70]

SYSTEMATIC DISCOVERY OF GENOMIC APTAMERS

A systematic search for novel natural aptamers is possible through a potent, genomic SELEX procedure. Genomic SELEX enriches endogenously encoded sequences representing a bait-binding domain—the genomic aptamer—within the putative transcript. When combined with massive sequencing and the availability of genome sequences it enables discovery of new components of the RNA–protein interaction network. It not only allows for screening new, unexplored genomes, but also makes it feasible to discover high-affinity binders in unexpected regions of known genomes. It does so by the fact that the selection is performed irrespectively of the expression profile of the aptamer. This means that genomic SELEX has a potential to identify RNA aptamers that are either expressed at a very low level or only transiently in a certain time window or derives from a silent heterochromatic region within a genome.

The general genomic SELEX procedure does not deviate much from the one described in the classical SELEX section (Figure 1). However, the major difference between those protocols lies in the starting pool used in the selection, as discussed below.[71]

Genomic SELEX Libraries

The initial library for genomic SELEX derives from the total DNA of the organism of interest and comprises of all potential structures encoded in the genome irrespective of the expression of the sequence. Genomic DNA is randomly primed and in case of the RNA genomic SELEX, transcribed into RNA molecules.[16,72] Random priming of the genome of interest ensures its complete coverage and its even amplification. In order to enable mapping of selected aptamers after high-throughput sequencing to the correct genomic location, it is important to use DNA of an organism or strain whose sequence is fully annotated.

A challenging variant of genomic SELEX may be screening for natural aptamers in genetic material recovered from environmental samples or metagenomes. These are from organisms or strains that cannot be cultured in a laboratory or organisms obtained directly from their natural environment. A metagenome is all the genetic material taken from a single environmental sample, which would include many different species. We speculate that exploring metagenomes by constructing a metagenomic SELEX library from, for example, sequenced microbial flora of human digestive system,[73,74] can have useful medical implications.

Nowadays, it is becoming more achievable to recover and sequence genomic material from extinct organisms.[75,76] We speculate that constructing a genomic library from an ancient DNA library to screen for genomic aptamers may be very informative with respect to evolution and conservation of the RNA–protein interaction and regulatory networks.

Choice of Bait

The choice of bait is a crucial determinant for a successful genomic SELEX. Small molecules, peptides, or proteins are all used as bait in SELEX procedures. However, to exploit the full potential of genomic SELEX, it is advisable to pick the nucleic acid-binding protein carefully. The bait should possess many potential partners with diverse specificities and affinities. We expect that most RNA-binding proteins involved in the regulation of heterochromatin formation, transcription, RNA processing, stability, and degradation have a large spectrum of target sequences with varying affinities. It will be necessary to isolate all these targets to define the RNA regulatory network in cells. Nonspecific proteins, like BSA, or those that interact with nucleic acids only transiently, like the bacterial RNA chaperone StpA, would not serve as good baits for genomic aptamer selection because no specific sequences can be enriched.[15] It is worth mentioning that the selection can be carried out with the full-length protein as well as its mutated or truncated forms.

Genomic SELEX Combined with High-Throughput Sequencing

Genomic SELEX combined with high-throughput sequencing is a valuable alternative and complementary approach to RNomics and computational predictions for the discovery of active non-coding RNAs, which contain recognition elements for specific ligands. If only a small number of targets for your bait of choice is expected, conventional cloning and sequencing will be appropriate. However, if the bait is a global regulator protein, genomic SELEX will uncover a large range of potential targets and also identify the binding motif(s),[15] and high-throughput sequencing will be necessary.

RNomics methods currently yield cDNA sequences of fragments or ends of *bona fide* transcripts, and the bioinformatic challenge to the analysis is to reconstruct the transcripts that were isolated from the experiment. Genomic SELEX and RNomics have the shared aim of the discovery of novel, active RNA transcripts; however, genomic SELEX sequences are not recovered from *bona fide* transcripts. A genomic aptamer locus instead represents a region that, when transcribed, binds the selected target with high affinity. The individual sequences from genomic SELEX have varying length and position when mapped to the genome, and therefore the bioinformatic challenge is to uncover the regions of the genome conferring the binding activity out of a mixture of related sequences. We have found that the diversity and coverage of these regions necessitate high-throughput sequencing analysis in order to obtain a complete picture of the pool.

Sequences in genomic SELEX result from direct recovery from a binding assay and are amplified through linker sequences that flank the aptamer. This affords two major advantages over RNomics methods: (1) the direct recovery and amplification makes it possible to more reliably analyze the abundances between aptamers within a single pool and (2) strand information is built into the sequences based on the fact that the T7 promoter is always upstream of the known linker sequences. This and the fact that the molecules are size selected prior to the screen also make fragmentation of the sequences unnecessary, and thus the molecules used for sequencing are full length. In order to take advantage of this, a method sequencing both ends of or through the molecule should be used.

The process of genomic SELEX is a progressive homogenization from an initial library containing fragments of the genome with even coverage to a final pool enriched in fragments from the initial library that confer binding of the bait protein. Because some regions of the genome will confer more binding activity than others, the mapped sequences will cluster into these regions in highly varying abundances, depending on their binding fitness. In our Hfq genomic SELEX screen we observed more than 300-fold dynamic range of enrichment. We expect this would increase with deeper sequencing, as only about 10,000 sequences were analyzed.[15]

Even more interestingly, the tiled start and end points of the mapped reads within the clusters resulted in highly differential patterns of enrichment at the nucleotide level. When viewed as a signal map, the highest point of these 'hills' of enrichment often contained the motif which was independently discovered as a binding domain of Hfq aptamers, and additionally discovered in a *de novo* motif search of the entire pool using a computational motif search. As shown in Figure 3, one hill correlated with protected nucleotides in a dimethyl sulfate (DMS) modification protection assay. We envision that analyzing the shapes of these hills, coupled with content and secondary structure analysis will be the key to uncovering the location and type of binding domain from the experiment.

When dealing with genomes with higher variability and repetitiveness, it is helpful to sequence the initial genomic library. This is especially important when assessing the enrichment of difficult-to-map sequences such as repeats and centromeric elements. As copy number variation in individuals can bias the genomic size of annotated repeats, it is more informative to compare the enriched levels of a repeat element to the initial library rather than the genome.

We envision the application of genomic SELEX to study the machineries that regulate and modulate the transcriptome of a cell. For example, our genomic SELEX screen for aptamers binding *E. coli* global regulator Hfq led to the discovery that most Hfq binding domains occur in lowly expressed antisense intergenic regions of polycistronic genes. In general, all components that recognize and recruit RNAs are candidate baits, such as transcription factors, polymerases, chromatin remodelers, histones, splicing factors, components of the exosome, the degradosome, the spliceosome, and the ribosome to mention just the most prominent. We propose that genomic SELEX will identify many genomic aptamers that are *cis*-acting regulators of an RNA's transcription, localization and stability, as well as *trans*-regulating components of the RNA–protein interaction network.

DISCOVERY OF NATURAL SMALL RIBOZYMES

Thirty years ago, the first ribozyme was discovered in nature. Since then, the quest to discover unique ribozymes has resulted in very few naturally occurring ones. The first ribozyme reported was the 413-nt group I intron from *Tetrahymena thermophilia*, and it was shown to be a self-splicing RNA in the presence of a guanosine cofactor.[77] An exogenous guanosine nucleotide becomes covalently attached to the 5′ end of the intron, and therefore, it was easy to discover other group I intron ribozymes by using a radiolabeled guanosine.[78–82] An important class of ribozymes, subsequently discovered, is the small self-cleaving ribozymes. Self-cleaving ribozymes perform a reversible phosphodiester cleavage reaction. Several of these small self-cleaving ribozymes have been discovered in RNA plant viruses and are utilized in the replication of the virus by first self-cleaving after rolling circle replication into monomers and then self-ligating into a circle again.[83] These small ribozymes include the hairpin,[84,85] hammerhead,[86]

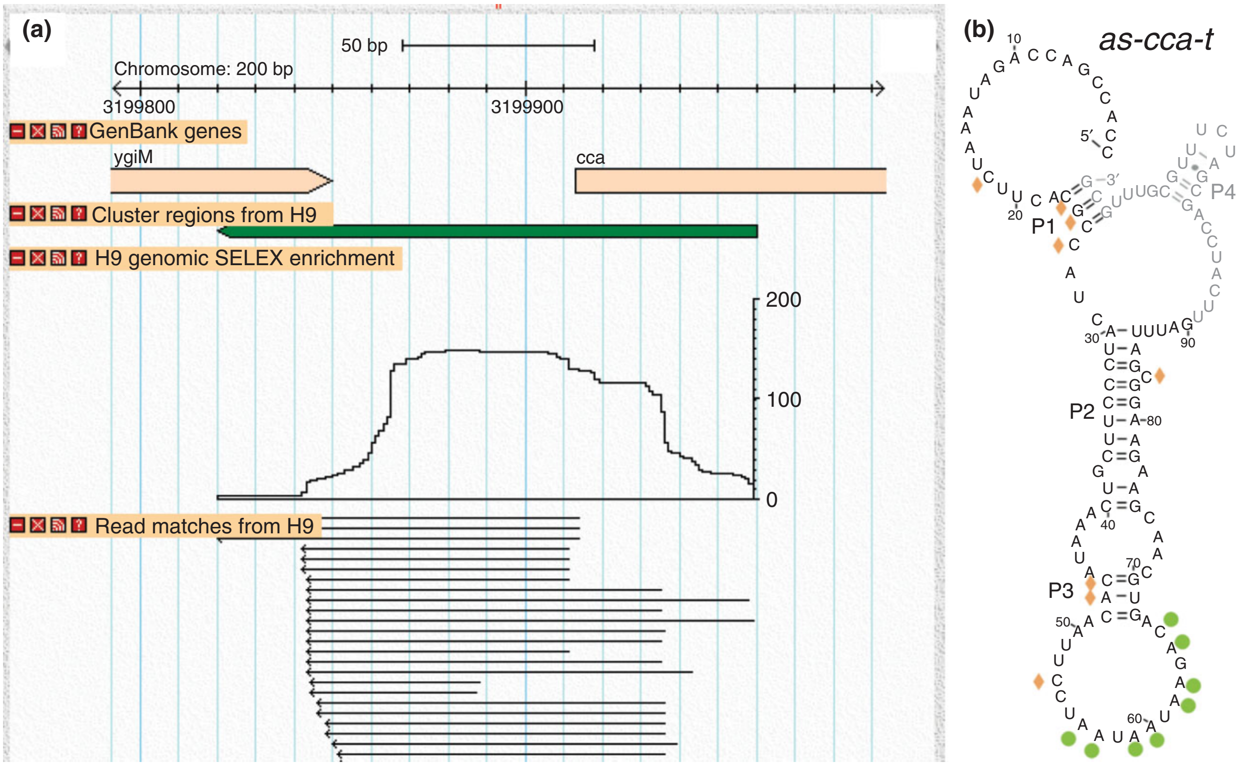


FIGURE 3 | Genomic systematic evolution of ligands by exponential enrichment (SELEX) high-throughput analysis. (a) Unlike RNomics, genomic SELEX regions are not *bona fide* transcripts. Instead, the reads cluster into regions where binding fitness is enhanced. The aligned reads (bottom left, not all are pictured) can be represented in a signal map showing the number of sequences recovered at each region. This nucleotide-level enrichment clarifies which parts of the sequence are most involved in the binding. The enriched part of this Hfq aptamer was selected for DMS footprinting analysis (b). The green circles indicate nucleotides protected upon Hfq binding, and orange diamonds indicate the nucleotides that were hydrolyzed in the presence of Hfq and not in the absence, indicating a conformational change. All the Hfq-protected bases are in the region of most enrichment as indicated using the high-throughput analysis.

hepatitis delta virus (HDV),[87] Varkud satellite (VS),[88] and the *glmS* ribozymes[56] (Figure 4). Although initially they were detected in viruses, they are now being discovered in the genomes of many different organisms but the function of these ribozymes in higher organisms is still being elucidated.[89]

Although the active sites of these ribozymes are highly conserved, the flanking stems and loops vary widely from species to species. Therefore, the discovery of small ribozymes within higher organisms has been limited. For example, the hammerhead ribozyme consists of three stems that do not have any sequence conservation between species, and a highly conserved 13-nucleotide active center at the junction of the stems.[91,92] An interaction between loops in stems I and II has been shown to be crucial for ribozyme activity under physiological conditions[92–94]; however, the sequences of these loops highly vary (for review on the differences in the structures and mechanisms of these ribozymes see Ref 90*)*. Although the active sites of these ribozymes have high sequence conservation, it is the overall structural architecture that is crucial for activity.

Discovery of Ribozymes through Genome Searches

Since many more genomes are being sequenced, it is possible to use alignment programs to search for new ribozymes. For example, searches for the hammerhead ribozyme typically consist of aligning the conserved core three-way junction to a genome, coupled with secondary structure analysis and manual inspection. Additionally, covariation analysis of the structure is used to determine if base-pairing interactions are conserved among the stems and if kissing loops are compatible. (For details see Section *Bioinformatic Methods for Discovering Ribozymes*).

Recently, two new hammerhead ribozymes were found in the *Arabidopsis* genome[95] and both are located in the antisense direction either at the 3′ end of an ORF or between two ORFs. In another study by de la Pena and coworkers, the authors performed

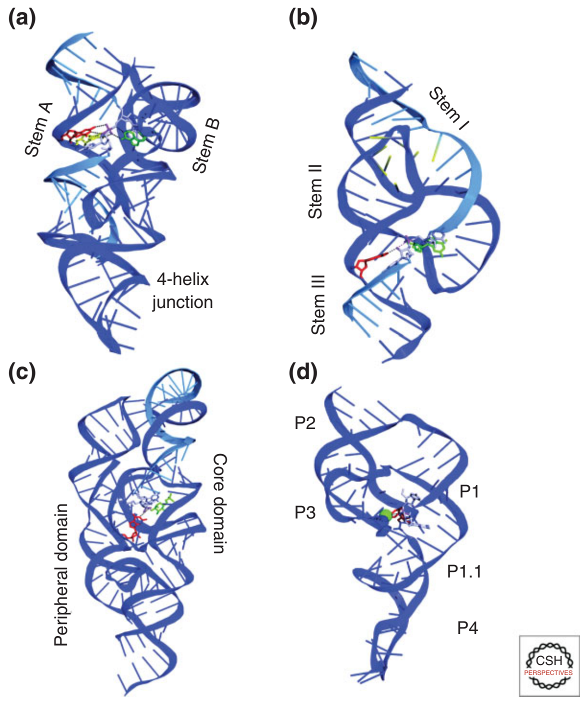


FIGURE 4 | Structures of four self-cleaving ribozymes. (a) The hairpin ribozyme, (b) the hammerhead ribozyme (c) the *glmS* ribozyme-riboswitch, and (d) the hepatitis delta virus (HDV) ribozyme. Active site residues are colored in red, green, and magenta. (Reprinted with permission from Ref 90. Copyright 2010 Cold Spring Harbor Laboratory Press) see reference for more information.

thorough searches on a large set of genomes looking specifically for hammerhead ribozymes.[96] They found hundreds of hammerhead motifs associated with retrotransposon elements. In addition, the ribozymes map to intronic regions and many seem to be ultraconserved. The ultraconservation suggests an essential biological role other than just retrotransposition. Owing to the low efficiency of the ribozyme cleavage, the authors speculate that these ribozymes contribute to alternative splicing of genes. Finding sequences that look like ribozymes within genomes only suggests that these ribozymes are functional. Further biochemical analysis is required to verify their biological role.

Searching databases for novel ribozymes within genomes can be a difficult process because, as stated before, the flanking sequences are not conserved, but are crucial for the proper folding and activity of ribozymes *in vivo*. Because it is known that ribozymes can function as modules consisting of an enzymatic module and a substrate module, and because RNA sequences far removed from each other in a single molecule can fold and form complexes, it is possible to search for 'discontinuous' ribozymes within the genome. This opens up the possibility for searching for ribozymes within genomes where it was previously thought there was none. Discontinuous hammerhead ribozymes were recently found within a mammalian genome.[97] Using bioinformatic searches, Martick and colleagues allowed up to 5000 bases to exist between stems I and III and discovered three 'discontinuous' hammerhead ribozymes within the 3′ UTRs of rodent C-type lectin type II mRNAs. To assess the biological relevance of the ribozymes in the 3′ UTR of an mRNA, they first tested and found that the ribozymes are active *in vitro*. Using a dual luciferase assay, they showed that the artificial ribozymes reduce reporter expression when incorporated into the 3′ UTR. Previously it was shown that ribozymes reduce the expression of genes by self-cleavage leading to unstable transcripts and their rapid destruction.[98] Therefore, it seems reasonable to search for more ribozymes in the 3′ UTRs of lowly expressed genes.

Discovery of Ribozymes that Control Gene Expression

The ability of RNA to form alternative stable structures gives it an advantage to be an excellent molecule for gene regulation because it can easily switch between an on and off conformation.[99] However, the cleavage activity of ribozymes involved in gene expression must be tightly regulated in order to turn on or off genes at the appropriate time. Hammerhead ribozymes have been artificially inserted into the 3′ and 5′ UTRs and intronic regions of genes and it was found that these ribozymes lead to the destabilization and degradation of transcripts.[98] Ribozymes inserted into the 5′ UTR of eukaryotic genes had the largest effect on the down-regulation of transcript expression, probably because of the removal of the 5′ cap. Although these ribozymes were artificially inserted into the transcript, these are mechanisms by which naturally occurring ribozymes may function.

One prominent example of a ribozyme controlling gene expression is that of the *glmS* ribozyme in bacteria, which was discovered through bioinformatic searches for riboswitches.[56] When the amount of glucosamine-6-phosphate (GlcN6p) is too high, it binds the riboswitch with the ribozyme activity, and thereby stabilizes the active conformation, cleaving the GlcN6p synthase mRNA. Additionally, an allosteric group I self-splicing ribozyme was recently discovered through bioinformatic searches and found to be sensitive to the levels of cyclic di-guanosyl-5′-monophosphate (c-di-GMP).[100] The ribozyme changes the alternative splicing of the genes involved in c-di-GMP production, degradation, and signaling depending on the levels of c-di-GMP in the cell.

Another example of a ribozyme controlling gene expression is one found in the *CPEB3* gene in the

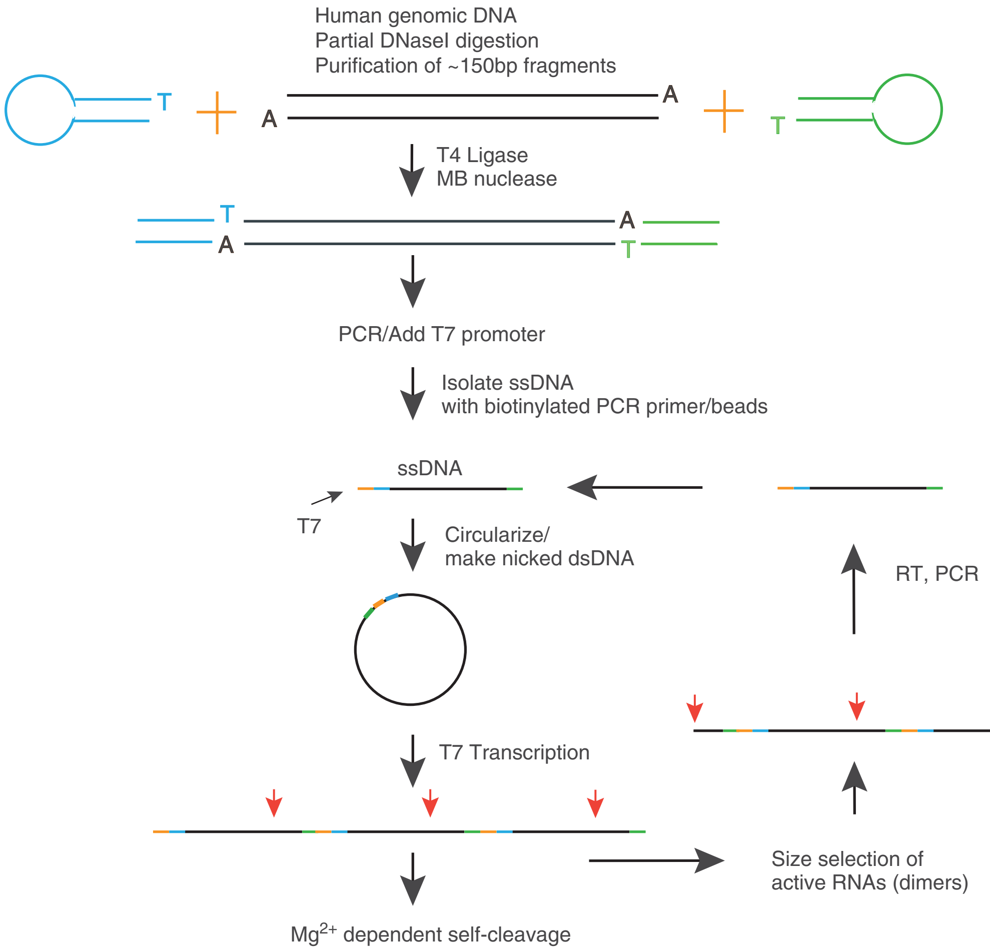


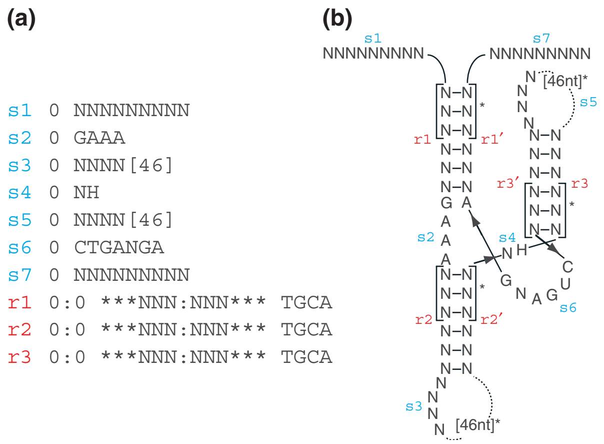
FIGURE 5 | Genomic systematic evolution of ligands by exponential enrichment (SELEX) for ribozymes. Human genomic DNA was partially digested and size selected to 150 base pairs and ligated to double-stranded hairpin primers. The loops were digested and then amplified by performing PCR to add the T7 promoter. A biotinylated primer was used to amplify and then used to extract one strand of single-stranded DNA (ssDNA). The ssDNA was ligated and then incubated with primers, Taq Pol, and dNTPs to produce nicked, circular, double-stranded DNA (dsDNA). Rolling circle transcription produced long transcripts with many potential cleavage sites (red arrows). Mg^{2+} induced cleavage produced singular, dimer, and multimer units. The dimers were isolated because they contain the full sequence, and then they were reverse transcribed (RT) and amplified [polymerase chain reaction (PCR)] for the next round of selection.[101]

human genome.[101] This ribozyme was discovered through a genome-wide search with a biochemical technique similar to genomic SELEX. A library of molecules of 150 base pairs in length derived from human genomic DNA was flanked by fixed primer sequences (Figure 5). The DNA library was circularized, transcribed in a rolling circle, and those that self-cleaved produced single unit and dimer RNAs that were easy to separate from the uncleaved RNAs. An alignment to the human genome showed that the selected ribozyme is located in a highly conserved region of a large intron of the *CPEB3* gene. Interestingly, the ribozyme has a high sequence and structural similarity to the HDV ribozyme. A comparison between the human and chimp ribozymes revealed that a single base change in the chimp ribozyme leads to a more stable native fold, resulting in faster cleavage rates.[102] The mutation in the human ribozyme allows for a stable alternative fold that turns off the ribozyme. In general, it is likely that regulatory factors, small molecules or proteins, stabilize the native folds of ribozymes within transcripts *in vivo*, allowing for modulation of cleavage activity and therefore, tight regulation of gene expression. This biochemical technique may be used with other genomic libraries to discover novel ribozymes within any new genome.

BIOINFORMATIC METHODS FOR DISCOVERING RIBOZYMES

The *de novo* discovery of many ribozymes has been powered by bioinformatic searches. Hammerhead ribozymes represent a unique challenge to their discovery because the sequence is only loosely conserved yet the structure is instrumental to their function. The problem is even further confounded by the fact that tertiary, non-Watson–Crick interactions are critical to their function. These base-backbone and other so-called *trans* interactions have not yet been reliably predicted algorithmically. However, there has been some success using clever combinations of packages to search and further filter the results.

To work around this limitation of traditional alignment methods, pattern matching methods such as PatScan[103] (applied in Ref 95), RNAMOT,[104] and RNABOB (Eddy, unpublished, applied in Refs 97, 105, 106) are used. The advantage over traditional alignment strategies is the ability to design a motif with indefinitely long insertions in regions where the lengths of stems and bulges are quite variable. In the case of RNABOB, one can also describe some types of tertiary contacts. RNABOB has recently been used for the discovery of hammerhead ribozymes in the human microbiome[106] (Figure 6). The permissiveness of the pattern used required additional analysis with the ViennaRNA package.[107] In this study, several novel hammerhead ribozymes were confirmed *in vitro* to be the fastest known natural hammerhead ribozymes; however, the sequence similarity to known hammerhead ribozymes was limited. While there has been some success in these, the challenge of computationally discovering loosely conserved motifs will remain difficult.



```
s1  0  NNNNNNNNN
s2  0  GAAA
s3  0  NNNN[46]
s4  0  NH
s5  0  NNNN[46]
s6  0  CTGANGA
s7  0  NNNNNNNNN
r1  0:0  ***NNN:NNN***  TGCA
r2  0:0  ***NNN:NNN***  TGCA
r3  0:0  ***NNN:NNN***  TGCA
```

FIGURE 6 | RNABOB search for hammerhead ribozymes. RNABOB (Eddy, Janelia Farms, unpublished) has been used to scan genomes for complicated, flexible structures which otherwise could not be predicted with alignment algorithms, thermodynamical algorithms such as the Zucker algorithm (ViennaRNA, mfold) or covariance models. Instead of inputting a sequence and finding a homolog with a similar structure, the user inputs a specific pattern with a flexible framework for insertions. Panel (a) shows the descriptors used to search hammerhead ribozymes in the human microbiome.[106] The elements are depicted in the canonical hammerhead structure cartoon in (b). 's's are single-stranded elements and they are identified by a sequence constraint on the right. 'N's mean any character, and ambiguity codes are allowed as well (e.g., 'H' indicates an 'A', 'C', or 'T'). Numbers in brackets (e.g., '[46]') indicate an insertion of up to that many characters with no constraint on the content of the sequence. Therefore, s3 and s5 may be 4–50 nucleotides long; 'r's are relational elements, which is a generalized form of a hairpin. It allows the user to specify the required base pairing in the fourth field in the parts of the pattern in the third field ('***NNN:NNN***') where only 'N's are specified. Here 'TGCA' indicates that T may pair with A and G with C, i.e., all pairs are allowed but GU. (To allow them 'TGYR' would be input instead.) The '*'s indicate an optional character, so in the case shown, all 'r' elements are allowed to be 3–6 pairs long. To specify the topology, the features are input in order. In this case, it would be 's1 r1 s2 r2 s3 r2′ s4 r3 s5 r3′ s6 r1′ s7'.

SELECTION AND DISCOVERY OF ARTIFICIAL RIBOZYMES

The discovery of catalytic RNA led to speculation of an ancient RNA world in which heredity and metabolism relied on RNA molecules alone. Therefore, many groups sought to discover catalytic RNAs that perform reactions that could have supported an RNA-only world. SELEX is the most widely used method for discovering ribozymes with new functions or to enhance the abilities of known ribozymes under various conditions. However, unlike a selection for binding a protein or small molecule, SELEX with ribozymes has particular challenges. Most challenging is the separation of active and inactive molecules for ribozymes that catalyze reactions *in trans*. In addition, for a randomized library of molecules that self-cleave, the substrate and enzyme are physically separated during the selection process. To address these issues, the ribozyme library can be designed in such a way that the cleavage site is part of the fixed primer for amplification. Therefore, the substrate gets cleaved, but the ribozyme remains intact for further rounds of selection and the substrate can be re-added during amplification. (For a review on designing a selection strategy for ribozymes see Ref 108.) Ribozymes have been selected for self-cleavage upon binding small molecules, selected for cleaving under certain conditions, and even selected for catalyzing new reactions.

Selection of Self-Cleaving Ribozymes

In an attempt to discover new self-cleaving ribozymes, Salehi Ashtiani and colleagues performed SELEX with a random pool of DNA, with fixed sequences on the 5′ and 3′ ends, and a fixed cleavage site within the 5′ fixed sequence.[109] After many rounds of selection, the authors found that the hammerhead ribozyme was the only ribozyme highly enriched. The authors conclude that the hammerhead is one of the simplest ribozymes. This experiment suggested that it maybe difficult to discover brand new active sites for ribozymes with SELEX. However, many groups in the late 1990s sought to improve the activities of the known small self-cleaving ribozymes using SELEX.[110–117]

Ribozyme activity *in vivo* is extremely appealing to those who wish to design artificial ribozymes that cleave targeted genes. Eckstein and colleagues specifically designed a hairpin ribozyme targeted toward cleavage of the *CTNNB1* gene, which is essential in cancer development.[118] The target mRNA sequence contributes to one side of an internal loop in the ribozyme and the nonessential base in the other side of the loop was randomized. In this way, they could tailor their hairpin sequence specifically to the mRNA target. Ribozymes can be engineered to correct genetic disorders by targeting a mutant gene transcript, which can be less dangerous than targeting the DNA.[119]

Ribozymes can also be selected for activity upon binding to a small molecule, similar to the *glmS* ribozyme.[56] In one study Meli and colleagues randomized the two loops of the hairpin ribozyme and first selected for loss of catalytic activity, then selected for recovery of activity upon addition of an adenine molecule.[120] An adenine-dependent ribozyme was recently used to control gene expression of the Tpl2/Cot oncogene.[121] Allosteric ribozymes can also be designed such that a small molecule-binding apatmer is fused to a randomized ribozyme. In this case, the ribozyme is selected based on the cleavage activity upon the structural rearrangement induced by the metabolite binding. This has been done to design theophylline, FMN[122] and ATP[123] sensitive ribozymes. Recently, Ausländer and colleagues took such a designer ribozyme one step further and introduced the theophylline sensitive ribozyme into the 5′ UTR of eukaryotic mRNA. They were able to optimize the control of gene expression using theophylline *in vivo*.[124] (For a protocol for selection of allosteric ribozymes see Ref 125.)

Selection of Ribozymes that Perform Other Chemical Reactions

Although many studies have been done on ribozymes that cleave RNA, ribozymes can also perform chemical reactions with other substrates. Ribozymes have been selected for performing reactions such as acylation,[126] *N*-glycosidic bond formation of nucleic acids,[127] and peptide bond formation.[128] One challenge in selecting a truly catalytic ribozyme, in which the enzyme remains unchanged throughout the reaction, is the discrimination between the active and inactive ribozymes. For self-cleavage reactions, this separation can easily be done by size. However, when the ribozyme performs a reaction *in trans*, there must be a physical link between the reactant and the product in order to identify the active molecules. Agresti and coworkers used *in vitro* compartmentalization to aid in the discovery of a ribozyme that catalyzes a Diels-Alder cycloaddition.[129] They fused a randomized library to an anthracene molecule, and in a water/oil emulsion were able to isolate each molecule individually (Figure 7). This allowed one oil bubble to contain a single gene that was transcribed into a single ribozyme. *In vitro* selection of the active molecules requires some type of capture tag for recovery. In this case, a biotinylated maleimide was added and, if the reaction took place, the gene would become biotinylated, allowing for separation from inactive molecule. Because the reaction takes place *in trans*, the isolation of the ribozyme with the gene that coded for it was a key to the success of this method.

In vitro compartmentalization was also recently used to select for RNA polymerase ribozymes.[130,131]

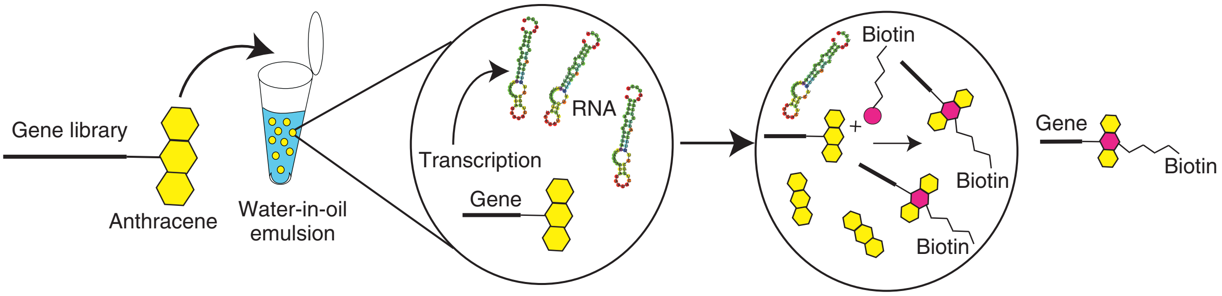


FIGURE 7 | Selection of Diels-Alder ribozymes by *in vitro* compartmentalization. A randomized dsDNA library of genes encoding potential ribozymes was fused to anthracene through a polyethylene glycol (PEG) linker. The genes were compartmentalized within droplets of a water-in-oil emulsion so that there was less than one molecule per compartment. The genes were transcribed and then Mg^{2+} and biotin-maleimide were added so that they diffused into the compartments. The active ribozymes carried out the Diels-Alder reaction, which fused the biotin to the gene of interest. The active genes are isolated by binding to streptavidin beads and amplified for further rounds of selection.[129]

The oil/water emulsion was used to first isolate the transcription of the ribozyme library in which the DNA and ribozyme (through a complementary hairpin) were both coupled to biotin/streptavidin beads. The emulsion was broken to couple the transcription primer to the beads and add the template, and then the complexes were again placed in an emulsion to isolate the ribozyme-templated transcription. This careful strategy resulted in a ribozyme-catalyzed transcription of up to 95 nucleotides. The discovery of a polymerase ribozyme is further evidence that an RNA-only world very well could have existed and even thrived. With current tools of selection and isolation of reactants and products, it is possible to recreate many of the essential chemical reactions for life using ribozymes.

CONCLUSION

The discovery of novel and unexpected principles often happens by serendipity and is then verified through well-structured experiments. In contrast, systematic searches detect mostly representatives from already known classes of molecules or slight variants of these. How can we identify more unknown and novel molecules in an unbiased way rather than look for ‘more of the same’? History has told us that we often anticipate principles. The best example is the fate of RNA aptamers. As soon as they were reported, instead of screening nature for examples, scientists designed sophisticated strategies to isolate synthetic aptamers. They were even used to regulate gene expression within cells[5,6] before they were identified as a part of riboswitches.[11,13] Since this landmark result, a large repertoire of natural aptamers has been discovered.

The situation is different for ribozymes. We still only know of a few classes, and the question remains open as to whether any more will be found. At the very least, there is evidence that self-cleavage is a widespread activity in many genomes: in the study of Seehafer et al.,[105] it was found that hammerhead ribozyme structures are prevalent across many clades, and they were found many more times than expected in random sequences. Their filtering system was confirmed on several predictions; further supporting the idea that self-cleavage of RNA is a pervasive mechanism.

In the near future, we will explore vast amounts of sequence space of many living and maybe some extinct organisms. We can construct genomic libraries of metagenomes from organisms that resist laboratory culture and characterization. With the combination of advanced computational analyses, unbiased experimental approaches and comparative genomics, we should succeed in expanding the repertoire of functional aptamers and ribozymes.

ACKNOWLEDGMENTS

Work in our laboratory was funded by the Austrian Science Fund FWF grants F4308 and I538 to R.S., the doctoral program in RNA Biology W1207 to B.Z. and the Austrian GenAU project D-110421-11 to R.S.

REFERENCES

1. Raney BJ, Cline MS, Rosenbloom KR, Dreszer TR, Learned K, Barber GP, Meyer LR, Sloan CA, Malladi VS, Roskin KM, et al. ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic Acids Res* 2011, 39:871–875.
2. Johnson JM, Edwards S, Shoemaker D, Schadt EE. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet* 2005, 21:93–102.
3. Doudna JA, Cech TR. The chemical repertoire of natural ribozymes. *Nature* 2002, 418:222–228.
4. Jäschke A. Artificial ribozymes and deoxyribozymes. *Curr Opin Struct Biol* 2001, 11:321–326.
5. Werstuck G, Green MR. Controlling gene expression in living cells through small molecule-RNA interactions. *Science* 1998, 282:296–298.
6. Suess B, Hanson S, Berens C, Fink B, Schroeder R, Hillen W. Conditional gene expression by controlling translation with tetracycline-binding aptamers. *Nucleic Acids Res* 2003, 31:1853–1858.
7. Bass BL, Cech TR. Specific interaction between the self-splicing RNA of Tetrahymena and its guanosine substrate: implications for biological catalysis by RNA. *Nature* 1984, 308:820–826.
8. Yarus M. A specific amino acid binding site composed of RNA. *Science* 1988, 240:1751–1758.
9. von Ahsen U, Davies J, Schroeder R: antibiotic inhibition of group I ribozyme function. *Nature* 1991, 353:368–370.
10. Hermann T, Patel DJ. Adaptive recognition by nucleic acid aptamers. *Science* 2000, 287:820–825.
11. Mironov AS, Gusarov I, Rafikov R, Lopez LE, Shatalin K, Kreneva RA, Perumov DA, Nudler E.

Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell* 2002, 111:747–756.

12. Winkler WC, Cohen-Chalamish S, Breaker RR. An mRNA structure that controls gene expression by binding FMN. *Proc Natl Acad Sci U S A* 2002, 99: 15908–15913.
13. Nahvi A, Sudarsan N, Ebert MS, Zou X, Brown KL, Breaker RR. Genetic control by a metabolite binding mRNA. *Chem Biol* 2002, 9:1043–1049.
14. Mendonsa SD, Bowser MT. In vitro selection of high-affinity DNA ligands for human IgE using capillary electrophoresis. *Anal Chem* 2004, 76:5387–5392.
15. Lorenz C, Gesell T, Zimmermann B, Schoeberl U, Bilusic I, Rajkowitsch L, Waldsich C, von Haeseler A, Schroeder R. Genomic SELEX for Hfq-binding RNAs identifies genomic aptamers predominantly in antisense transcripts. *Nucleic Acids Res* 2010, 38: 3794–3808.
16. Watrin M, Von Pelchrzim F, Dausse E, Schroeder R, Toulmé JJ. In vitro selection of RNA aptamers derived from a genomic human library against the TAR RNA element of HIV-1. *Biochemistry* 2009, 48:6278–6284.
17. Hofmann HP, Limmer S, Hornung V, Sprinzl M. Ni2+-binding RNA motifs with an asymmetric purine-rich internal loop and a G-A base pair. *RNA* 1997, 3: 1289–1300.
18. Sazani PL, Larralde R, Szostak JW. A small aptamer with strong and specific recognition of the triphosphate of ATP. *J Am Chem Soc* 2004, 126:8370–8371.
19. Niazi JH, Lee SJ, Gu MB. Single-stranded DNA aptamers specific for antibiotics tetracyclines. *Bioorg Med Chem* 2008, 16:7245–7253.
20. Gopinath SC, Misono TS, Kawasaki K, Mizuno T, Imai M, Odagiri T, Kumar PK. An RNA aptamer that distinguishes between closely related human influenza viruses and inhibits haemagglutinin-mediated membrane fusion. *J Gen Virol* 2006, 87:479–487.
21. Lee YJ, Seong-Wook L. *In vitro* Selection of Cancer-Specific RNA Aptamers. *J Microbiol Biotechnol* 2006, 16:1149–1153.
22. Fisher TS, Joshi P, Prasad VR. Mutations that confer resistance to template-analog inhibitors of human immunodeficiency virus (HIV) type 1 reverse transcriptase lead to severe defects in HIV replication. *J Virol* 2002, 76:4068–4072.
23. Jenison RD, Gill SC, Pardi A, Polisky B. High-resolution molecular discrimination by RNA. *Science* 1994, 263:1425–1429.
24. Ellington AD, Szostak JW. In vitro selection of RNA molecules that bind specific ligands. *Nature* 1990, 346:818–822.
25. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 1990, 249: 505–510.
26. Hirao I, Harada Y, Nojima T, Osawa Y, Masaki H, Yokoyama S. In vitro selection of RNA aptamers that bind to colicin E3 and structurally resemble the decoding site of 16S ribosomal RNA. *Biochemistry* 2004, 43: 3214–3221.
27. Vater A, Jarosch F, Buchner K, Klussmann S. Short bioactive Spiegelmers to migraine-associated calcitonin gene-related peptide rapidly identified by a novel approach: tailored-SELEX. *Nucleic Acids Res* 2003, 31:130.
28. Tang J, Xie J, Shao N, Yan Y. The DNA aptamers that specifically recognize ricin toxin are selected by two in vitro selection methods. *Electrophoresis* 2006, 27:1303–1311.
29. Lozupone C, Changayil S, Majerfeld I, Yarus M. Selection of the simplest RNA that binds isoleucine. *RNA* 2003, 9:1315–1322.
30. Gopinath SC, Kawasaki K, Kumar PK. Selection of RNA-aptamer against human influenza B virus. *Nucleic Acids Symp Ser (Oxf)* 2005:85–86.
31. Jensen KB, Atkinson BL, Willis MC, Koch TH, Gold L. Using in vitro selection to direct the covalent attachment of human immunodeficiency virus type 1 Rev protein to high-affinity RNA ligands. *Proc Natl Acad Sci U S A* 1995, 92:12220–12224.
32. Yao W, Adelman K, Bruenn JA. In vitro selection of packaging sites in a double-stranded RNA virus. *J Virol* 1997, 71:2157–2162.
33. Berezovski MV, Musheev MU, Drabovich AP, Jitkova JV, Krylov SN. Non-SELEX: selection of aptamers without intermediate amplification of candidate oligonucleotides. *Nat Protoc* 2006, 1:1359–1369.
34. Mayer G, Ahmed MS, Dolf A, Endl E, Knolle PA, Famulok M. Fluorescence-activated cell sorting for aptamer SELEX with cell mixtures. *Nat Protoc* 2010, 5:1993–2004.
35. Tsuji S, Hirabayashi N, Kato S, Akitomi J, Egashira H, Tanaka T, Waga I, Ohtsu T. Effective isolation of RNA aptamer through suppression of PCR bias. *Biochem Biophys Res Commun* 2009, 386:223–226.
36. Zimmermann B, Gesell T, Chen D, Lorenz C, Schroeder R. Monitoring genomic sequences during SELEX using high-throughput sequencing: neutral SELEX. *PLoS One* 2010, 5:e9169.
37. Tereshko V, Skripkin E, Patel DJ. Encapsulating streptomycin within a small 40-mer RNA. *Chem Biol* 2003, 10:175–187.
38. Mosing RK, Mendonsa SD, Bowser MT. Capillary electrophoresis-SELEX selection of aptamers with affinity for HIV-1 reverse transcriptase. *Anal Chem* 2005, 77:6107–6112.
39. Collett JR, Cho EJ, Ellington AD. Production and processing of aptamer microarrays. *Methods* 2005, 37:4–15.
40. Chushak Y, Stone MO. In silico selection of RNA aptamers. *Nucleic Acids Res* 2009, 37:e87.

41. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008, 26:1135–1145.

42. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007, 447:799–816.

43. Grundy FJ, Henkin TM. The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in gram-positive bacteria. *Mol Microbiol* 1998, 30:737–749.

44. Yao Z, Weinberg Z, Ruzzo WL. CMfinder–a covariance model based RNA motif finding algorithm. *Bioinformatics* 2006, 22:445–452.

45. Weinberg Z, Barrick JE, Yao Z, Roth A, Kim JN, Gore J, Wang JX, Lee ER, Block KF, Sudarsan N, et al. Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Res* 2007, 35:4809–4819.

46. Weinberg Z, Wang JX, Bogue J, Yang J, Corbino K, Moy RH, Breaker RR. Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol* 2010, 11:R31.

47. Yao Z, Barrick J, Weinberg Z, Neph S, Breaker R, Tompa M, Ruzzo WL. A computational pipeline for high- throughput discovery of cis-regulatory noncoding RNA in prokaryotes. *PLoS Comput Biol* 2007, 3:e126.

48. Barrick JE, Breaker RR. The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome Biol* 2007, 8:R239.

49. Sudarsan N, Hammond MC, Block KF, Welz R, Barrick JE, Roth A, Breaker RR. Tandem riboswitch architectures exhibit complex gene control functions. *Science* 2006, 314:300–304.

50. Mandal M, Lee M, Barrick JE, Weinberg Z, Emilsson GM, Ruzzo WL, Breaker RR. A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science* 2004, 306:275–279.

51. Sudarsan N, Barrick JE, Breaker RR. Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA* 2003, 9:644–647.

52. Bocobza S, Adato A, Mandel T, Shapira M, Nudler E, Aharoni A. Riboswitch-dependent gene regulation and its evolution in the plant kingdom. *Genes Dev* 2007, 21:2874–2879.

53. Wachter A, Tunc-Ozdemir M, Grove BC, Green PJ, Shintani DK, Breaker RR. Riboswitch control of gene expression in plants by splicing and alternative 3' end processing of mRNAs. *Plant Cell* 2007, 19: 3437–3450.

54. Kubodera T, Watanabe M, Yoshiuchi K, Yamashita N, Nishimura A, Nakai S, Gomi K, Hanamoto H. Thiamine-regulated gene expression of Aspergillus oryzae thiA requires splicing of the intron containing a riboswitch-like domain in the 5'-UTR. *FEBS Lett* 2003, 555:516–520.

55. Cheah MT, Wachter A, Sudarsan N, Breaker RR. Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature* 2007, 447:497–500.

56. Winkler WC, Nahvi A, Roth A, Collins JA, Breaker RR. Control of gene expression by a natural metabolite-responsive ribozyme. *Nature* 2004, 428: 281–286.

57. Joyce GF. The antiquity of RNA-based evolution. *Nature* 2002, 418:214–221.

58. Mandal M, Boese B, Barrick JE, Winkler WC, Breaker RR. Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria. *Cell* 2003, 113:577–586.

59. Bohn C, Rigoulay C, Chabelskaya S, Sharma CM, Marchais A, Skorski P, Borezée-Durant E, Barbet R, Jacquet E, Jacq A, et al. Experimental discovery of small RNAs in Staphylococcus aureus reveals a riboregulator of central metabolism. *Nucleic Acids Res* 2010, 38:6620–6636.

60. Toledo-Arana A, Dussurget O, Nikitas G, Sesto N, Guet-Revillet H, Balestrino D, Loh E, Gripenland J, Tiensuu T, Vaitkevicius K, et al. The Listeria transcriptional landscape from saprophytism to virulence. *Nature* 2009, 459:950–956.

61. Loh E, Dussurget O, Gripenland J, Vaitkevicius K, Tiensuu T, Mandin P, Repoila F, Buchrieser C, Cossart P, Johansson J. A trans-acting riboswitch controls expression of the virulence regulator PrfA in Listeria monocytogenes. *Cell* 2009, 139:770–779.

62. Bengert P, Dandekar T. Riboswitch finder–a tool for identification of riboswitch RNAs. *Nucleic Acids Res* 2004, 32:W154–W159.

63. Chang TH, Huang HD, Wu LC, Yeh CT, Liu BJ, Horng JT. Computational identification of riboswitches based on RNA conserved functional sequences and conformations. *RNA* 2009, 15: 1426–1430.

64. Eddy SR, Durbin R. RNA sequence analysis using covariance models. *Nucleic Acids Res* 1994, 22: 2079–2088.

65. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, Bateman A. Rfam: updates to the RNA families database. *Nucleic Acids Res* 2009, 37: D136–D140.

66. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. *Bioinformatics* 2009, 25: 1335–1337.

67. Abreu-Goodger C, Ontiveros-Palacios N, Ciria R, Merino E. Conserved regulatory motifs in bacteria: riboswitches and beyond. *Trends Genet* 2004, 20: 475–479.

68. Abreu-Goodger C, Merino E. RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements. *Nucleic Acids Res* 2005, 33: W690–692.

69. Lyngsø RB, Pedersen CN. RNA pseudoknot prediction in energy-based models. *J Comput Biol* 2000, 7: 409–427.

70. Meyer MM, Hammond MC, Salinas Y, Roth A, Sudarsan N, Breaker RR. Challenges of ligand identification for riboswitch candidates. *RNA Biol* 2011, 8: 5–10.

71. Zimmermann B, Bilusic I, Lorenz C, Schroeder R. Genomic SELEX: a discovery tool for genomic aptamers. *Methods* 2010, 52:125–132.

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

73. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010, 464:59–65.

74. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, et al. Enterotypes of the human gut microbiome. *Nature* 2011, 473:174–180.

75. Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M, Pääbo S. Analysis of one million base pairs of Neanderthal DNA. *Nature* 2006, 444: 330–336.

76. Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Chen F, Platt D, Pääbo S, Pritchard JK, Rubin EM. Sequencing and analysis of Neanderthal genomic DNA. *Science* 2006, 314:1113–1118.

77. Zaug AJ, Cech TR. In vitro splicing of the ribosomal RNA precursor in nuclei of Tetrahymena. *Cell* 1980, 19:331–338.

78. Cech TR, Zaug AJ, Grabowski PJ. In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell* 1981, 27:487–496.

79. Peebles CL, Perlman PS, Mecklenburg KL, Petrillo ML, Tabor JH, Jarrell KA, Cheng HL. A self-splicing RNA excises an intron lariat. *Cell* 1986, 44:213–223.

80. van der Veen R, Arnberg AC, van der Horst G, Bonen L, Tabak HF, Grivell LA. Excised group II introns in yeast mitochondria are lariats and can be formed by self-splicing in vitro. *Cell* 1986, 44: 225–234.

81. Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* 1983, 35:849–857.

82. Michel F, Umesono K, Ozeki H. Comparative and functional anatomy of group II catalytic introns–a review. *Gene* 1989, 82:5–30.

83. Symons RH. Small catalytic RNAs. *Annu Rev Biochem* 1992, 61:641–671.

84. Feldstein PA, Buzayan JM, Bruening G. Two sequences participating in the autolytic processing of satellite tobacco ringspot virus complementary RNA. *Gene* 1989, 82:53–61.

85. Haseloff J, Gerlach WL. Sequences required for self-catalysed cleavage of the satellite RNA of tobacco ringspot virus. *Gene* 1989, 82:43–52.

86. Prody GA, Bakos JT, Buzayan JM, Schneider IR, Bruening G. Autolytic processing of dimeric plant virus satellite RNA. *Science* 1986, 231:1577–1580.

87. Sharmeen L, Kuo MY, Dinter-Gottlieb G, Taylor J. Antigenomic RNA of human hepatitis delta virus can undergo self-cleavage. *J Virol* 1988, 62:2674–2679.

88. Saville BJ, Collins RA. A site-specific self-cleavage reaction performed by a novel RNA in Neurospora mitochondria. *Cell* 1990, 61:685–696.

89. de la Peña M, García-Robles I. Ubiquitous presence of the hammerhead ribozyme motif along the tree of life. *RNA* 2010, 16:1943–1950.

90. Ferré-D'Amaré AR, Scott WG. Small self-cleaving ribozymes. *Cold Spring Harb Perspect Biol* 2010, 2:a003574.

91. Martick M, Scott WG. Tertiary contacts distant from the active site prime a ribozyme for catalysis. *Cell* 2006, 126:309–320.

92. Khvorova A, Lescoute A, Westhof E, Jayasena SD. Sequence elements outside the hammerhead ribozyme catalytic core enable intracellular activity. *Nat Struct Biol* 2003, 10:708–712.

93. De la Peña M, Gago S, Flores R. Peripheral regions of natural hammerhead ribozymes greatly increase their self-cleavage activity. *EMBO J* 2003, 22:5561–5570.

94. Boots JL, Canny MD, Azimi E, Pardi A. Metal ion specificities for folding and cleavage activity in the Schistosoma hammerhead ribozyme. *RNA* 2008, 14: 2212–2222.

95. Przybilski R, Gräf S, Lescoute A, Nellen W, Westhof E, Steger G, Hammann C. Functional hammerhead ribozymes naturally encoded in the genome of Arabidopsis thaliana. *Plant Cell* 2005, 17:1877–1885.

96. de la Peña M, García-Robles I. Intronic hammerhead ribozymes are ultraconserved in the human genome. *EMBO Rep* 2010, 11:711–716.

97. Martick M, Horan LH, Noller HF, Scott WG. A discontinuous hammerhead ribozyme embedded in a mammalian messenger RNA. *Nature* 2008, 454: 899–902.

98. Yen L, Svendsen J, Lee JS, Gray JT, Magnier M, Baba T, D'Amato RJ, Mulligan RC. Exogenous control of mammalian gene expression through

modulation of RNA self-cleavage. *Nature* 2004, 431:471–476.

99. Breaker RR. Natural and engineered nucleic acids as tools to explore biology. *Nature* 2004, 432:838–845.

100. Lee ER, Baker JL, Weinberg Z, Sudarsan N, Breaker RR. An allosteric self-splicing ribozyme triggered by a bacterial second messenger. *Science* 2010, 329: 845–848.

101. Salehi-Ashtiani K, Lupták A, Litovchick A, Szostak JW. A genomewide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene. *Science* 2006, 313:1788–1792.

102. Chadalavada DM, Gratton EA, Bevilacqua PC. The human HDV-like CPEB3 ribozyme is intrinsically fast-reacting. *Biochemistry* 2010, 49:5321–5330.

103. Dsouza M, Larsen N, Overbeek R. Searching for patterns in genomic data. *Trends Genet* 1997, 13: 497–498.

104. Ferbeyre G, Bourdeau V, Pageau M, Miramontes P, Cedergren R. Distribution of hammerhead and hammerhead-like RNA motifs through the GenBank. *Genome Res* 2000, 10:1011–1019.

105. Seehafer C, Kalweit A, Steger G, Gräf S, Hammann C. From alpaca to zebrafish: hammerhead ribozymes wherever you look. *RNA* 2011, 17:21–26.

106. Jimenez RM, Delwart E, Lupták A. Structure-based search reveals hammerhead ribozymes in the human microbiome. *J Biol Chem* 2011, 286:7737–7743.

107. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. *Monatsh Chem* 1994, 125:167–188.

108. Eckstein F, Kore AR, Nakamaye KL. In vitro selection of hammerhead ribozyme sequence variants. *Chembiochem* 2001, 2:629–635.

109. Salehi-Ashtiani K, Szostak JW. In vitro evolution suggests multiple origins for the hammerhead ribozyme. *Nature* 2001, 414:82–84.

110. Wright MC, Joyce GF. Continuous in vitro evolution of catalytic function. *Science* 1997, 276:614–617.

111. Thomson JB, Sigurdsson ST, Zeuch A, Eckstein F. In vitro selection of hammerhead ribozymes containing a bulged nucleotide in stem II. *Nucleic Acids Res* 1996, 24:4401–4406.

112. Sargueil B, Burke JM. In vitro selection of hairpin ribozymes. *Methods Mol Biol* 1997, 74:289–300.

113. Jayasena VK, Gold L. In vitro selection of self-cleaving RNAs with a low pH optimum. *Proc Natl Acad Sci U S A* 1997, 94:10612–10617.

114. Vaish NK, Heaton PA, Eckstein F. Isolation of hammerhead ribozymes with altered core sequences by in vitro selection. *Biochemistry* 1997, 36:6495–6501.

115. Conaty J, Hendry P, Lockett T. Selected classes of minimised hammerhead ribozyme have very high cleavage rates at low Mg2+ concentration. *Nucleic Acids Res* 1999, 27:2400–2407.

116. Tuschl T, Sharp PA, Bartel DP. Selection in vitro of novel ribozymes from a partially randomized U2 and U6 snRNA library. *EMBO J* 1998, 17:2637–2650.

117. Zillmann M, Limauro SE, Goodchild J. In vitro optimization of truncated stem-loop II variants of the hammerhead ribozyme for cleavage in low concentrations of magnesium under non-turnover conditions. *RNA* 1997, 3:734–747.

118. Drude I, Strahl A, Galla D, Müller O, Müller S. Design of hairpin ribozyme variants with improved activity for poorly processed substrates. *FEBS J* 2011, 278: 622–633.

119. Müller S. Engineered ribozymes as molecular tools for site-specific alteration of RNA sequence. *Chembiochem* 2003, 4:991–997.

120. Meli M, Vergne J, Maurel MC. In vitro selection of adenine-dependent hairpin ribozymes. *J Biol Chem* 2003, 278:9835–9842.

121. Li YL, Vergne J, Torchet C, Maurel MC. In vitro selection of adenine-dependent ribozyme against Tpl2/Cot oncogene. *FEBS J* 2009, 276:303–314.

122. Soukup GA, Breaker RR. Design of allosteric hammerhead ribozymes activated by ligand-induced structure stabilization. *Structure* 1999, 7:783–791.

123. Tang J, Breaker RR. Rational design of allosteric ribozymes. *Chem Biol* 1997, 4:453–459.

124. Ausländer S, Ketzer P, Hartig JS. A ligand-dependent hammerhead ribozyme switch for controlling mammalian gene expression. *Mol Biosyst* 2010, 6: 807–814.

125. Piganeau N. In vitro selection of allosteric ribozymes. *Methods Mol Biol* 2009, 535:45–57.

126. Wilson C, Szostak JW. In vitro evolution of a self-alkylating ribozyme. *Nature* 1995, 374:777–782.

127. Unrau PJ, Bartel DP. RNA-catalysed nucleotide synthesis. *Nature* 1998, 395:260–263.

128. Zhang B, Cech TR. Peptide bond formation by in vitro selected ribozymes. *Nature* 1997, 390:96–100.

129. Agresti JJ, Kelly BT, Jäschke A, Griffiths AD. Selection of ribozymes that catalyse multiple-turnover Diels-Alder cycloadditions by using in vitro compartmentalization. *Proc Natl Acad Sci U S A* 2005, 102:16170–16175.

130. Zaher HS, Unrau PJ. Selection of an improved RNA polymerase ribozyme with superior extension and fidelity. *RNA* 2007, 13:1017–1026.

131. Wochner A, Attwater J, Coulson A, Holliger P. Ribozyme-catalyzed transcription of an active ribozyme. *Science* 2011, 332:209–212.

Boots, J., MATYLLA-KULINSKA, K., Zywicki, M., Zimmermann, B., and Schroeder, R. (2012) **Genomic SELEX** in “Handbook of RNA Biochemistry” 2nd Edition Edt. Hartmann R.K., Bindereif, A., Westhof, E. Wiley-VCH

I contributed to this publication in discussing the content and the outline of the chapter as well as in writing the following sections: Introduction, Library construction, Choice of bait, Biochemical analysis of the genomic aptamers.

53
Genomic SELEX

Jennifer L. Boots, Katarzyna Matylla-Kulinska, Marek Zywicki, Bob Zimmermann, and Renée Schroeder

53.1
Introduction

SELEX (Systematic Evolution of Ligands by EXponential enrichment) was developed to screen large libraries for sequences that bind with high affinity to ligands of choice [1, 2]. This is achieved by binding the library to a target, separating the bound from unbound sequences, and finally amplifying the selected sequences for further rounds of selection [2, 3]. The libraries were initially obtained by randomly synthesizing nucleic acids, leading to pools of highly complex sequence space. In the recent years, with the onset of tools to sequence entire pools in parallel and with so many sequenced genomes, the SELEX procedure has been adapted to screen genomes for functional DNA or RNA motifs that bind to interesting targets. When the libraries are derived from genomic DNA instead of random sequences, the procedure is referred to as *genomic SELEX*.

Genomic SELEX is an important tool for the discovery of genome-encoded aptamers and regulatory sequences that interact with proteins or other ligands. For example, genomic SELEX has been particularly useful in finding DNA targets for transcription factors [4–6], RNA targets for splicing factors [7], and novel RNA–RNA loop–loop interactions [8]. In genomic SELEX, the initial library is not random but is composed of varying lengths of genomic DNA that represent the entire genome [9]. The DNA or RNA molecules selected from genomic SELEX experiments are referred to as *genomic aptamers* [10]. The main advantages of genomic SELEX over the classical one are that it uses a significantly reduced allowable sequence space and increases the likelihood that a biologically relevant target is selected. Moreover, since the initial library originates from genomic DNA, it screens for aptamers regardless of RNA expression levels. Thus it is possible to select RNAs that are expressed at very low levels, or at a specific cell cycle point or developmental stage. However, a limitation of genomic SELEX is that the selected genomic aptamers are derived from RNAs that may or may not be expressed. Classical methods, such as massive sequencing of RNAs bound to a target, guarantee to result in expressed RNAs.

Handbook of RNA Biochemistry, 2nd Edition. Edited by R.K. Hartmann, A. Bindereif, A. Schön, and E. Westhof.

In this chapter, we give a comprehensive introduction into the genomic SELEX method. We discuss in detail how to construct a genomic RNA library starting with any available genomic DNA, then how to select for RNAs that bind to a target of interest (the bait), and finally how to evaluate the sequences obtained from the selection. Depending on the bait, a very large number of aptamers might be expected. In that case, high-throughput sequencing is essential. If the source genome is large and contains highly repetitive elements, finding the original genomic location of the selected sequences and statistical evaluation of the data requires special care. The major goal of the computational analysis is to create a basis for selecting candidates for further biochemical analysis. Since genomic aptamers discovered with genomic SELEX only represent the binding domain within an encoded RNA, and not necessarily the full transcript *in vivo*, further characterization of the transcript may be necessary to understand the biological relevance of the RNA–ligand interaction. We discuss methods to characterize these transcripts and ways to evaluate the potential biological function of the interaction.

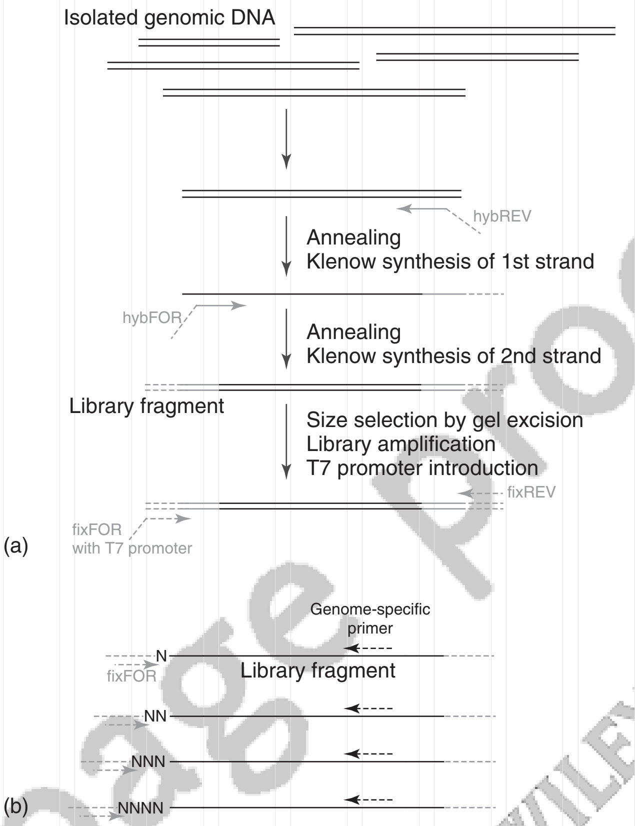
53.2 Description of the Methods

53.2.1 Library Construction

The initial library for the genomic SELEX procedure is created from the genomic DNA pool of an organism of interest, which is randomly primed and transcribed into RNA [9, 11]. As a result, the library entirely covers the genome of interest, so every potential genomic aptamer is represented in the starting pool. The advantage of constructing a library from genomic DNA is to screen for genomic aptamers irrespective of their expression profile in general and as it relates to phases of the cell cycle or developmental stages.

After isolating or purchasing high-quality genomic DNA, the first strand is synthesized with the Klenow fragment of DNA polymerase. The hybREV primer (both hyb primers consist of a specific sequence followed by randomized nucleotides at the 3′end) is annealed to the genomic DNA at 25 °C and then extended. Before second strand synthesis with the hybFOR, the excess hybREV should be thoroughly removed (for example, with a microconcentrator) in order to reduce the formation of fragments flanked with the same sequence on both sides. After synthesis of both strands, size selection follows. The lengths of fragments should correspond to the size of potential aptamers being targeted. Klenow reaction products are usually resolved on denaturing polyacrylamide gel, and fragments of desired size are excised. After DNA elution from a gel piece, a T7 promoter sequence is introduced by a subsequent polymerase chain reaction (PCR) amplification with primers fixFOR/fixREV (see Figure 53.1a).

We have previously used the sequence 5′-CCAAGTAATACGACTCACTATAGGGGAATTCGGAGCGGGCAGC-3′ (T7 promoter sequence underlined) for the



Q1 **Figure 53.1** (a) Scheme for a genomic SELEX● library construction. High-quality genomic DNA is used to construct the genomic SELEX library. First, the hybREV primer is annealed to genomic DNA at 25 °C and extended by the Klenow fragment of DNA polymerase. Then, before the second strand synthesis, the excess hybREV should be carefully removed. After second strand synthesis with hybFOR, fragments of desired size are selected in the denaturing gel electrophoresis and eluted from the polyacrylamide gel. Next, a T7 promoter sequence is introduced by a subsequent PCR amplification with primers fixFOR/fixREV. Gray parts of hybFOR/hybREV represent nine random nucleotides. The dashed gray sections represent specific fixFOR/fixREV sequence. The T7 promoter sequence in the fixFOR is depicted by the dotted gray line. (b) Library quality control by analysis of the distribution of end points. If the genomic library represents the entire genome, PCR amplification with the combination of one genome-specific primer and one library-end-point-specific (fixFOR/fixREV) primer should result in amplicons that differ by a single nucleotide.

fixFOR-T7 primer [11]. The downstream sequence has the advantage that it is not present in current assemblies of human, yeast, and *Escherichia. coli* genomes. It also functions well under the recommended PCR annealing temperature (55 °C) of the reverse primer, fixREV, 5′-CGGGATCCTCGGGGCTGGGATG-3′, which is also nongenomic. Restriction sites are also included at the 3′ ends of fixFOR (*Bbv*I) and fixREV (*Fok*I). This can be useful in swapping fixed primers between selection rounds to guarantee that the selected motifs do not include the fixed primer sequences. Since the T7 promoter is not needed in the Klenow phase, hybFOR consists of only the sequence 5′-AGGGGAATTCGGAGCGGGCAGC-3′ followed by nine random nucleotides. The sequence of hybREV is the same as fixREV with nine random nucleotides at the 3′ end.

Before the first SELEX round, it is advisable to test the quality of the genomic library. To ensure that the genome coverage is reasonable, PCRs with several arbitrarily chosen primer pairs are performed to give amplicons corresponding to the length of fragments in the library. For easier evaluation, it is recommended to test amplicons from single-copy regions. As an additional control, a •PCR with a gene-specific primer in combination with fixREV or fixFOR can be performed to ensure that the obtained products have sizes that vary in the desired range (Figure 53.1b). Q2

53.2.2 Choice of Bait

The choice of bait is an essential step. In principle, aptamers against any ligand can be obtained as long as the ligand is soluble under conditions in which the RNA is stable. SELEX has been performed against a very long list of small molecules, ranging from primary metabolites, coenzymes, antibiotics, to synthetic drugs [3]. When searching for genomic aptamers against small molecules, it is advisable to design a procedure that does not require a linker for immobilization. This is because it became apparent from the X-ray structures of ligand–aptamer complexes of riboswitches [12] that the ligand is entirely embedded within the aptamer, leaving no space for a linker. The most common baits for genomic SELEX are proteins that are chosen because they are involved in the regulation of RNA expression, folding, or activity. Many proteins contain predicted RNA-recognition motifs where the target RNA is not known. RNA-binding proteins involved in transcription, processing, stability, and degradation are good candidates for genomic SELEX. However, proteins that have been shown to bind RNA nonspecifically or only transiently would not be good candidates since they will most likely not enrich any specific aptamers [13]. Finally, it is important for genomic SELEX that the protein of interest be highly pure and stable *in vitro*.

53.2.3 SELEX Procedure

A summary of the SELEX procedure is shown in Figure 53.2. The specific details of the procedure are discussed in this section. This procedure was adapted from Ref. [11].

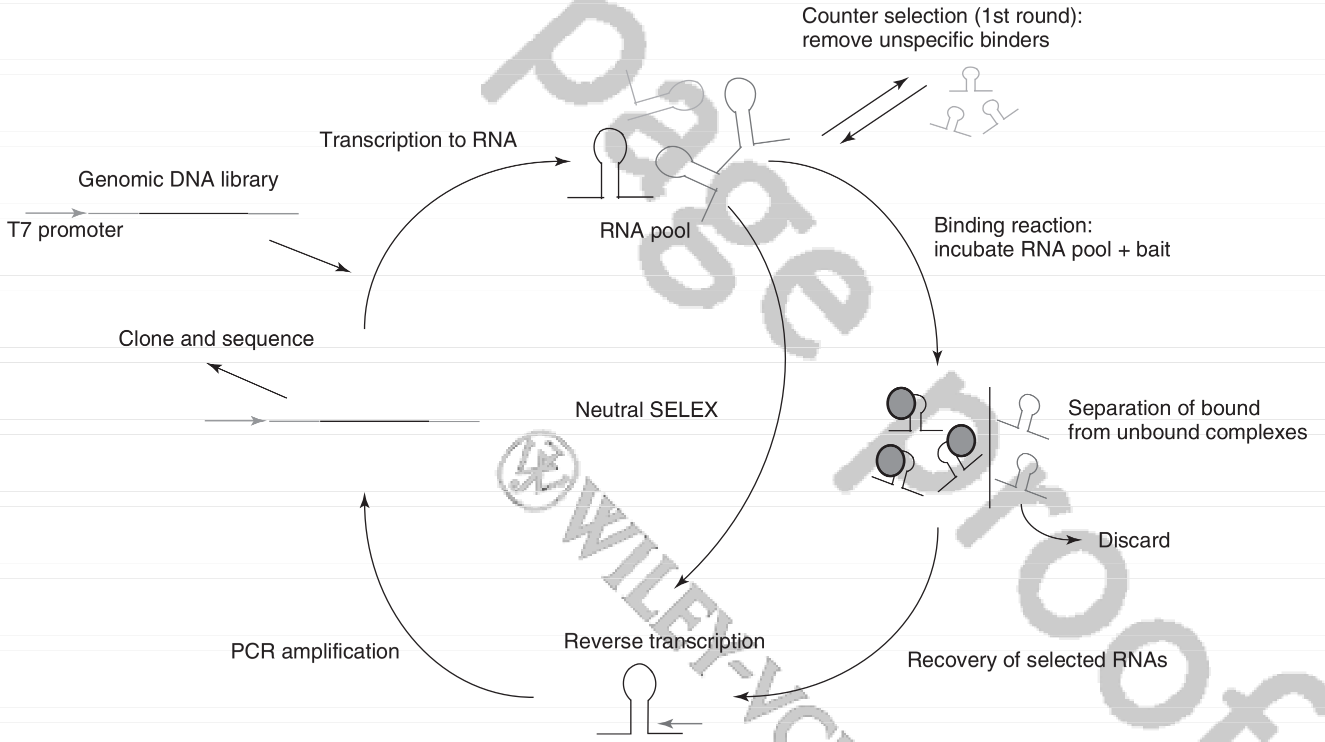


Figure 53.2 The SELEX procedure overview. The genomic DNA library is transcribed into RNA using T7 polymerase. In the first round of selection, a counter selection is used to remove unspecific binding RNAs that bind to the apparatus for separation. The cleared RNA library is incubated with the target of interest (the bait). The RNAs that bind to the target are then separated from the unbound using a variety of methods. Next, bound RNAs are recovered and amplified using reverse transcription and PCR. Finally, the RNA pool is either subjected to further rounds of selection or cloned and sequenced.

53.2.3.1 Transcription of Genomic Library into RNA Library

Genomic SELEX can be performed directly with the DNA library if DNA aptamers are the targets of interest. However, in this chapter, we focus on how to perform genomic SELEX with an RNA library. The first step in the SELEX procedure is to transcribe the DNA library into RNA. To do this, incubate 10 μg of the genomic library DNA with 20 μl of 25 mM rNTP mix, 1 μl of RNase inhibitor (Promega), T7 polymerase, 10 μl 10X buffer, and trace amounts of α-^{32}P GTP in a 100 μl reaction volume for 4 h at 37 °C. The amount of template, $MgCl_2$, rNTPs, T7 polymerase, and dithiothreitol can be varied to optimize the transcription reaction. The radioactivity is used to follow the RNA pool throughout the selection procedure and to estimate the enrichment after each round of selection.

After transcription, DNase I is added to degrade the template DNA library, and then the library is incubated at 37 °C for 30 min. It is important to eliminate the template DNA library when performing RNA genomic SELEX as certain DNA sequences may become enriched if the bait can also bind double-stranded DNA. To stop the reaction, heat-inactivate the DNase I at 65 °C for 10 min in the presence of EDTA (or according to manufacturer's guidelines). It is recommended to check the quality and size of the RNA fragments on an agarose gel or low-percentage (4%) polyacrylamide gel. Finally, dilute the RNA library to 500 μl in a binding buffer suitable for the RNA–bait interaction.

53.2.3.2 Counter Selection

To avoid enriching nonspecific RNA aptamers in the library, which can bind to the apparatus used for separation of bound from unbound complexes, a counter selection must be performed. For example, the diluted pool can be precleared by incubation with the membrane or the column matrix. The precleared library will flow through the membrane and can be recovered and purified by ethanol precipitation. For column separation, the RNA library is incubated with the beads, and then the RNAs that do not bind are recovered by either centrifugation or gravity flow in the column apparatus and ethanol precipitated. In addition, if the bait protein has a tag for column purification, the library can be incubated with beads containing the tag to eliminate RNAs that bind specifically to the tag. Counter selection may be performed with an inactive form of the bait protein to assure that the selected RNA sequences are specific for the active bait.

53.2.3.3 Positive Selection

The positive selection of RNA–bait complexes involves first an incubation step, to allow for complex formation, and then the separation of bound from unbound complexes. The concentration of the RNA library should be measured with UV spectroscopy and a scintillation counter to determine the counts per mole of RNA. This will be important for the calculation of the concentration of recovered RNA from each round of selection. Next, decide on the ratio of the RNA to ligand concentration. If the bait is a protein, a good starting point is a 10 : 1 molar ratio of RNA library : protein. It is critical that the RNA be in molar excess of the protein to establish an environment of competition for binding different species of RNA molecules in the library. For a protein with a known activity, it is advisable to use buffer conditions and time of incubation in which the protein is known to be active. However, if the protein function is unknown or if the bait is a small molecule, it is recommended to begin with near physiological buffer conditions. A good starting point for binding is room temperature (23 °C), where RNA secondary structure is stable and there is minimal denaturation of the protein [11].

Initial rounds of selection are normally carried out under moderate conditions, and then, during later rounds, the conditions are more stringent to increase the specificity of the selected complexes. For example, increasing the salt concentration, changing the RNA–bait ratio, or adding nonspecific competitors in later rounds of selection may increase the stringency of binding. Before binding, denature the RNA library for 1 min at 95 °C and then let it slowly cool to room temperature for ∼10 min to ensure refolding of the RNA. Next, incubate the RNA with the bait in the desired binding buffer for the amount of time required for the interaction. UV cross-linking may also be used to stabilize the RNA–protein complex [14].

After incubation of the RNA library with the bait, the bound and unbound complexes must be separated. This can be done using a variety of techniques. For RNA–protein complexes, the most convenient methods are membrane filtration [1, 13–16] and affinity chromatography [2, 17, 18]. For example, if the bait protein was purified with a fused tag, the RNA–protein complexes can be incubated with the appropriate affinity column. In addition, if there is an antibody that recognizes the target protein, the RNA–protein complexes can be immunoprecipitated. However, it must be kept in mind that the RNA may block binding to the antibody, so a control must be done to determine if this is a feasible method for separation.

If the protein is small, size exclusion methods, such as gel electrophoresis or size selection chromatography, are useful for separation [19]. Other methods include fluorescence-activated cell sorting [20] and surface plasmon resonance [21]. It is helpful to perform multiple selections in parallel using different binding conditions, by varying ligand concentrations, and using mutants of the bait protein to increase the specificity of the selection [10]. In order to confirm the selection results, it is also advisable to perform parallel selections using different immobilization methods, as well as to perform technical replicates.

Although genomic SELEX is usually performed *in vitro*, it is also possible to confirm a direct interaction between the RNA and protein by performing one or two cycles *in vivo*. In this case, the enriched library and the bait need to be fused to reporter molecules for a three-hybrid system readout [13]. Alternatively, the protein–RNA complex can be cross-linked and immunoprecipitated with an antibody.

53.2.3.4 Recovery and Amplification of Selected Sequences

After the positive selection step, the RNA sequences need to be recovered and amplified for further rounds of selection. After each round of selection, the membrane (or column) is counted on a scintillation counter to determine the approximate amount of selected RNA. This is useful for later estimation of the enrichment of the RNA pools during each round of selection.

The recovery of the selected RNAs depends on the method of separation. For membrane filtration, the RNA sequences can be recovered by incubating the membrane with 7 M urea, 20 mM sodium citrate, pH 5.0, 1 mM EDTA, and phenol (pH 5.2) and then shaking at 1400 rpm at room temperature for 10 min. This is followed by ethanol precipitation. For recovery of selected RNAs from a column, the protein–RNA complexes can be removed from the column by competition with a small molecule that binds the tag, the same as would be done for the purification of the tagged protein. Alternatively, the selected RNAs can be eluted from the column by digesting the protein with proteinase K. In both cases, it is advisable to then phenol, chloroform, isoamyl alcohol (PCI)-extract and ethanol precipitate the RNA pool. For gel electrophoresis, the RNA is recovered through crushing and soaking in elution buffer (10 mM Tris-HCl, pH 8, 0.3 M NaOAc pH 5.4, 2 mM EDTA, 0.1% SDS) and ethanol precipitation.

After recovery of genomic aptamers, the pool is reverse transcribed into DNA. We recommend using an enzyme that is active at elevated temperatures, such as 50–60 °C, to allow reverse transcription (RT) of highly structured RNAs. Alternatively, an RNA helicase can be used in combination with the reverse transcriptase. The RT step is followed by PCR amplification with a polymerase with high fidelity (to minimize sequence artifacts during amplification) and the fixed primers (see Section 11.2.1). It is recommended to perform 7–10 cycles in the PCR to avoid dimerization of incomplete products, which can be extended when primer–template ratio decreases, resulting in chimeric products (see Table• 53.1). Q3
After PCR, the DNA pool is phenol extracted and ethanol precipitated. The selected DNA pool is then transcribed, as previously discussed (see Section 11.2.3.1), and further rounds of selection can be performed. The number of cycles required to enrich the library depends on its initial complexity and on the desired affinity of the RNAs to the bait. About 7–12 rounds are typical for a genomic SELEX experiment. Usually, depending on the RNA–protein ratio used for complex formation, 30–60% of the input RNA will bind to the bait. Then the DNA fragments can be cloned and sequenced immediately (see Section 11.2.3.6), otherwise further rounds of selection are carried out.

53.2.3.5 Neutral SELEX

The amplification steps (PCR, *in vitro* transcription, RT) of genomic SELEX may introduce some bias in which sequences are ultimately selected. We previously developed a parallel control to SELEX in order to evaluate the effect of these steps on the initial library as SELEX proceeds [22]. To do this, the selection steps are omitted from the SELEX cycle (see Figure 53.2), and the results of each round can be sequenced. When we performed this, each round of the so-called neutral SELEX was sequenced. The average sequence had a less stable structure as the cycles progressed. We hypothesized that this is caused by the difficulty of the reverse transcriptase in denaturing highly structured RNAs. However, sometimes the selective pressures of binding can still lead to the enrichment of a highly structured RNA. For example, the SELEX-derived streptomycin aptamer was crystallized and shown to have a stable structure [23]. Other characteristic biases that were

Table 53.1

Step	Problem	Possible cause	Solution
Library construction	Uneven coverage of the genome	Genome is not amenable to random priming	As long as every region is represented, selection should be possible. Alternatively, whole genome amplification (WGA) kits can be used
	Fragments become shorter with subsequent rounds of genomic SELEX	Shorter products are amplified more efficiently	If the effect is strong, try increasing the elongation time in PCR. Also, relaxed stringency of selection can result in neutral selection effects; therefore, varying the stringency could help
	No PCR products during quality control	The library does not represent the entire genome	Generate more material during Klenow reaction, perhaps by varying the primer amounts Ensure that the isolated genomic DNA is highly pure WGA kits can be used (above)
		Fragments are too long for amplification	Pick shorter fragments, preferably within the size range of the library
		PCR conditions are not optimal	Vary $MgSO_4$, dNTPs, annealing temperature, or elongation times
	PCR products do not vary in expected size during quality control	Incorrect size selected in construction	Repeat size selection step
		Mispriming of designed primers	Check the annealing conditions and sequence of the primers being used
Amplification	Sequences become longer	Chimeric products formed because of low primer to template ratio	Decrease the number of PCR cycles Increase initial primer concentration Vary annealing temperature
Transcription	Not enough material obtained	Suboptimal reaction conditions	Vary $MgCl_2$, rNTPs, and template ratios
Reverse transcription	Structured RNAs not recovered	Reverse transcriptase does not denature RNA	Include an initial denaturation step Use a high-temperature reverse transcriptase Use a helicase during the reverse transcription step
Positive selection	RNA was not completely removed from membrane/column	Protein was not fully denatured on membrane	Repeat filter-elution steps
		Protein was not efficiently competed off column by small molecule	Adjust levels of small molecule; repeat filter elution
	No RNAs were enriched/selected	Protein is nonspecific	—
		Binding conditions are not optimal	Start with a decreased stringency by increasing the molarity of both RNA and protein in the solution

analyzed from the neutral SELEX sequences such as length, nucleotide content, and divergence from the initial library were only mildly affected, but this can vary depending on the features of the initial library. Thus we advise performing a parallel neutral SELEX control to any genomic SELEX in order to analyze the background signal.

53.2.3.6 Cloning and Sequencing

At any point during the initial rounds of selection, it is possible to clone and sequence the selected RNA pool to determine if any sequence is being enriched, and furthermore, if there are artifacts of DNA contamination or PCR chimers (see Table 53.1). The pool can be cloned into any commercially available T/A cloning vector according to the manufacturer's instructions.

For baits that may have a large number of RNA targets, it is essential to use high-throughput sequencing. Current technologies are● advancing at a staggering Q4 rate, so while discussing any one in particular, the other becomes obsolete. In any case, since the length of the sequences is constrained, no current technology would require sequence fragmentation. Without fragmentation, the fix primers on either end of the aptamers can be used to gain information about which genomic strand the aptamer lies. Additionally, both ends of the aptamer, or the whole aptamer, should be sequenced. Since the lengths are varying, this is essential to elucidate the enrichment patterns, binding motifs, and any potential structural elements encoded in the aptamer.

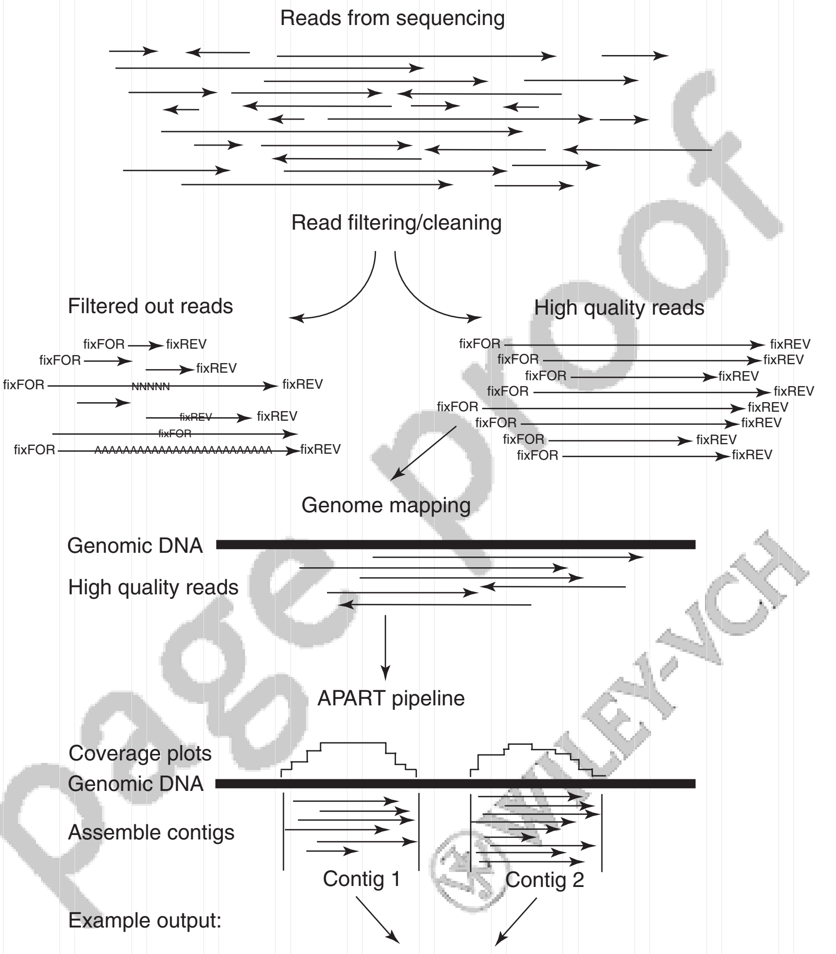
53.2.4 Troubleshooting

In Table 53.1, we describe common problems that could be faced during the genomic SELEX procedure. We suggest possible causes of these problems and recommend solutions.

53.3 Evaluation of Obtained Sequences

53.3.1 Computational Analysis of SELEX-Derived Sequences

The analysis of sequencing data obtained from the genomic SELEX experiment is usually focused on identification of genomic aptamers that have been enriched during the selection process. The sequenced data are referred to as *reads*. The typical procedure is to first perform an assembly of the reads into ''contigs'' based on sequence similarity to the reference genome, and then to identify high coverage peaks as putative binding motifs (Figure 53.3). All the activities can be performed step by step using a variety of available software or in a single run by using an automated pipeline named APART (*A*utomated *P*ipeline for *A*nalysis of



Contig name	Reads	Length	Exon/intron	Orientation	Position	Feature description
loc.III + 1 5294	3898	368	Intergenic	Antisense	chrIII:2-369	Telomeric region…
loc.XIII + 409766	1123	288	Exon	Sense	chrXIII:409767	Nuclear transcription…

Figure 53.3 Computational analysis flowchart. The reads obtained from sequencing must first be filtered and cleaned in order to proceed with only the highest quality reads. Reads that are too short, contain none or only one of the fixed primers, have fixed primers in the middle of the sequence, are made up of more than 50% homopolymer, or contain unknown nucleotides (N) from sequencing must be filtered out. The high-quality reads are then mapped to the genome using a variety of programs. The APART pipeline can then be used to group reads that map to the same location in the genome into contigs. Finally, the APART pipeline gives a table containing information such as number of reads, contig length, and location in the genome and a description of what is known about sequences from this region of the genome.

*R*NA *T*ranscripts) [24]. We recommend using the APART pipeline for most of the steps because it has been well optimized for handling nonunique reads and includes the identification of highly abundant regions within the assembled contigs. The initial steps, including read preparation and genomic alignment, are highly dependent on the quality and content of the library. Thus it is recommended that these steps be performed before the automated APART workflow in order to adjust and optimize the parameters set.

53.3.1.1 **Read Filtering and Cleaning**

In the first step, low-quality sequence reads have to be removed from the pool. The threshold depends on the sequencing technology and should be weighted according to the quality score distribution within the library so the vast majority of the reads pass through to the subsequent steps. When no quality criteria are already given for selecting the reads, removing the sequence of the bottom 5% quality scores would be a good starting point. Optional filters remove reads with low information content (containing more than 50% of homopolymer) or reads containing unresolved bases (''N''). The next task is to locate the adaptor sequences (including the fixed primers, see Section 11.2.1) surrounding the genomic aptamer. This can be achieved using any pattern-matching program. Our recommendation is patmatch [25]. Its major benefit is the possibility to separately control insertions, deletions, and substitutions. For the first pass, we recommend using a value of 3 for every type of change. However, depending on library quality, it is worth to test mismatch allowance values between 2 and 4. Usually, the 3′ ends of the reads are of lower quality than the 5′, thus an increase of allowed mismatches by 1 or 2 for the 3′ adaptor is usually a good solution. For downstream analysis, only the reads that contain both adapters should be used. When using the APART pipeline, all the above tasks can be performed by calling up a single automated script.

53.3.1.2 **Genome Mapping**

In order to map the reads to the reference genome, we recommend using the bowtie aligner [26] due to its speed, ability of reporting all matches for nonunique reads, and the extensive possibilities for control over the output. When using bowtie, we recommend the value of 1 up to 2 for libraries sequenced with high accuracy methods or 2 up to 3 for libraries obtained with technologies of lower base call accuracy.

For setting the alignments that bowtie should be allowed to report, we recommend the combined use of the ''-a,'' ''--best,'' and ''–strata'' options. This forces bowtie to report all matches for the particular read (-a option) sorted from the best to the worst (--best option) and limits the list to those within the best ''stratum'' (--strata option). ''Stratum'' refers to a level of alignment score (e.g., perfect match, one mismatch). The reasoning is that if a read is mapped to a repetitive or multicopy sequence, it will be mapped to all places where it could have originated, based on the sequencing data, and not be mapped to places that it is less likely to have originated. Reporting all best alignments maximizes the coverage of any given feature, allowing for higher copy features to show enrichment patterns.

Other options are mostly related to the speed performance of the aligner and should be adjusted according to the needs of the user. For compatibility with the APART pipeline, it is required to print the output in SAM● format (-S option) Q5 and report the reads matching more than the maximum allowed number of times with -M option. The APART pipeline can automatically perform a genomic alignment, using the bowtie aligner with parameters set for the highest quality output. However, this will take longer than running bowtie manually.

53.3.1.3 Assembly and Annotation

The next step is to group reads into contigs, or regions of the genome where overlapping reads are found. The APART pipeline can automatically assemble the reads into contigs and generate genomic browser-compatible tracks in bed and wig formats. It will also utilize reads that map to multiple regions of the genome and group contigs together if they contain identical sets of reads. APART will provide a comprehensive functional annotation of the contigs based on a genome annotation and sequence similarity, including identification of all known noncoding RNAs and repeat units.

Running APART is straightforward. However, the default parameter set can be optimized for RNA-seq projects. When using it for genomic SELEX analysis, we recommend a couple of deviations from the default. The minimum number of reads per contig should be set to 1 in order to include all the reads in the statistics. Additionally, the contig clustering method should be set to use read name sets, instead of contig sequence. However, for libraries of low quality or derived from organisms with high genetic variability, which contain substantial number of mismatches in genomic alignment, sequence-based clustering may perform better.

53.3.1.4 Enrichment Analysis

The major aim of the SELEX procedure is to enrich the initial RNA pool with molecules that bind the bait. Thus, the investigation of the global enrichment of the output library is the primary analysis that indicates the success of the experiment. The first look should be focused on verification of the read distribution among the contigs. A successful experiment will have a highly stratified distribution. That is, the top few contigs should contain a large percentage of the total reads recovered from the experiment. For example, in our Hfq genomic SELEX experiment, over one-third of the sequenced reads belonged to the top 15 contigs [13]. However, there was a total of 1522 contigs, indicating a high stratification of enrichment. In case of the opposite situation, the efficiency of the selection procedure should be reconsidered (especially the ionic conditions for binding and stringency of the washing steps).

Another recommended analysis focuses on enrichment of the specific functional genomic regions. In some cases, it can result in conclusive statements about possible functions of the molecule used as bait. The basic analysis includes two steps. First, we recommend performing a calculation of the enrichment of genomic features (introns, exons, protein-coding genes, ncRNA genes, etc.) by comparing the read number statistics generated by APART with the size of the respective feature in the genome. This step can also be used as a quality control for SELEX procedure if one knows what kind of sequences are expected to bind the bait (e.g., if one uses the intronic splicing enhancer as a bait, introns should be enriched). Second, the identification of enriched Gene Ontology (GO) categories within annotated contigs by using GeneTrail [27] or other GO enrichment analysis tool can be helpful in estimating the cellular functions of the molecule used as bait.

53.3.1.5 Benefits of Sequencing the Initial Library

For computational analysis of the genomic-SELEX-derived sequences, it is also recommended to sequence the initial library. There are two major benefits to sequencing the library. First, the presence of the selected read sequences in the initial library confirms its genomic origin. Thus, reads from a selected pool, which are confirmed to be in the initial library, can be utilized even if they are not matching the sequenced part of the reference genome. Second, after the assembly, the initial pool can be used as an exact background distribution for the enrichment analysis. It is possible that the initial library will vary from random genomic distribution because of differential accessibility of certain genomic regions for random priming, and therefore, it is useful to determine the possible artificial enrichment of certain sequences in the initial library.

53.3.1.6 Identification of the Binding Motif

The RNA motif responsible for binding to a target is usually determined by both the sequence and structure of the RNA. Unfortunately, the software available at present for *de novo* identification of complex RNA motifs is based on the assumption that all supplied sequences contain a unique motif. Since this is not always the case for genomic-SELEX-derived contigs, we suggest performing the motif search in the following several steps:

1) Cluster all contigs obtained from APART pipeline with sequence similarity threshold set to 70% using cd-hit [28] or any other clustering program of choice.
2) For each identified cluster, calculate the joint number of reads as the sum of reads for member contigs.
3) Depending on clustering results, identify a couple of clusters with the highest joint read number and perform a sequence motif search for contig sequences within the clusters. We recommend the use of Glam2 [29] program, which allows for the identification of gapped motifs. Check if there are similarities between the sequence motifs identified for individual clusters.
4) In parallel, perform the secondary-structure-based clustering of the contigs using either RNA Forester [30] or RNACluster [31] software. Compare clusters obtained with sequence-based clustering results.
5) Identify or refine consensus secondary structures for clusters of interest with Alifold [32] and RNA Consensus Shapes [33]. It is worthwhile to compare results from both tools, since they are based on different approaches.

53.3.2 Biochemical Analysis of the Genomic Aptamers

In this section, we discuss biochemical approaches suggested for the characterization of RNA molecules containing genomic aptamers mapped to the genome.

53.3.2.1 Validation of the RNA–Protein Interaction

The most evident control for the validation of selected genomic aptamers is to confirm the interaction between the enriched RNAs and the bait. Electrophoretic mobility shift assay (EMSA) [7, 34, 35] and filter binding assays [36, 37] are straightforward methods to check the interaction *in vitro* and to assess the binding strength. Alternatively, affinity between the transcript and the bait can be analyzed by surface plasmon resonance analysis [38] or fluorescence anisotropy [38]. It is also important to be aware that the entire RNA molecule may fold differently than the short genomic aptamer. In the context of the entire transcript, the selected domain may be involved in intramolecular interactions or be sterically inaccessible and therefore unable to bind the protein partner. Hence is advisable to repeat the binding assays once the full-length transcript is determined (see Section 11.3.2.3).

It is important to test whether the RNA interacts with the bait in the cellular environment. *In vivo* binding analysis is greatly facilitated when a specific antibody for the bait is available, so coimmunoprecipitation methods can be used (for example, see [39]). We recommend coimmunoprecipitation coupled with *in vivo* cross-linking, called *CLIP* [40], where the interaction is captured within the cell before cell lysis and therefore the amount of nonspecific contaminating RNA is reduced. The precipitated RNA pool is then analyzed by Northern blot, RNase protection assay (RPA), or strand-specific RT-PCR for the presence of the genomic aptamer of interest.

Genomic SELEX provides an RNA-binding domain, but not necessarily the minimal binding site. Among the well-established methods to determine the exact contact sites, there are boundary determination analysis [35] and RNA footprinting [38]. In addition, if the bait protein has a metal-ion-binding pocket and the RNA binds in proximity to it, an iron-directed cleavage assay may be chosen [41].

53.3.2.2 Expression Analysis of Genomic Aptamers

Given that the selected genomic aptamers are merely randomly transcribed binding domains, it is crucial to confirm their expression in a target cell at a specific time point. Northern blot analysis is a convenient tool for detection of abundant genomic aptamers. Moreover, it provides information about the size of the entire RNA molecule comprising the selected genomic aptamer. However, the method lacks sensitivity and requires large amount of RNA material. Therefore, for analysis of genomic aptamers that are not abundant, RPA and strand-specific RT-PCR are the methods of choice. Nevertheless, the accuracy of RT-PCR results has been recently questioned because experimental artifacts are suspected. The main source of error is primer-independent cDNA synthesis caused primarily by RNA self-priming

[42, 43] or priming by other short RNAs or residual DNA after DNaseI treatment during RNA preparation. For that reason, it is essential to provide appropriate negative controls. It is recommended to perform the RT step in the absence of primer and then compare PCR products with those carried out with specific primer. It has also been reported that use of actinomycin D in the RT step blocks spurious synthesis of the cDNA [42].

53.3.2.3 Reconstruction of the Whole-Transcript-Comprising Genomic Aptamer

For studying the biological significance of genomic aptamers, it is important to know the size of the native transcript comprising the selected binding domain. As mentioned in Section 11.3.2.2, a Northern blot serves as a good method for both verification of the cellular expression and size determination of the native transcript. However, it lacks sensitivity. To assess the length of the entire RNA molecule, we suggest performing 3′- and 5′ rapid amplification of cDNA ends [44–46] or RNA self-circularization followed by RT-PCR [47].

Alternatively, primer walking, an RT-PCR-based method, may be used (Figure 53.4). The genomic aptamer is first reverse transcribed from the desired RNA pool with a sequence-specific reverse primer and then amplified by PCR. In consecutive amplification steps, the reverse primer (used in the RT reaction) remains fixed, whereas the forward primer is placed several dozen nucleotides upstream. In subsequent PCR amplifications, the forward primer walks along the transcript by being shifted toward the 5′ end of the cDNA, as long as the amplicon is detectable. On the other hand, in the 3′ end mapping, the forward primer (used in the RT reaction) stays constant and the reverse primers are continuously placed downstream toward the 3′ terminus. When both extreme ends are reached, the amplification with the outermost primers is performed to prove that the detected transcript spans the full length. Once the characterization of the genomic aptamer–protein interaction and the analysis of the RNA transcript containing the genomic aptamer are completed, it is possible to speculate about the function of the protein–RNA complex.

53.3.2.4 Determining the Function of the RNA–Protein Interaction

Finally, it is important to demonstrate the biological relevance of the protein–RNA interaction *in vivo*. Since RNA binding to a protein is not synonymous with function, the relevance of the protein–RNA interaction must be determined. It is difficult to generalize a strategy for all genomic aptamers because it is inherently connected to the nature of the bait. In cases where the function of a protein is unknown, the identity of the target RNA may give insight into the function [36]. Moreover, the RNA–protein interaction may be disrupted *in vivo* by mutating or knocking down the protein, and then, by analyzing the effect it has on the RNA, one may gain insight into the function [7]. If the protein of interest is important for the RNA localization in the cell, a knockdown of the protein may show a mislocalization of the RNA. The third strategy is specific for genomic SELEX against enzymes. The most obvious functional assay for these is to see if the RNA inhibits or accelerates the enzymatic activity of the protein [37]. These are only a few examples of the

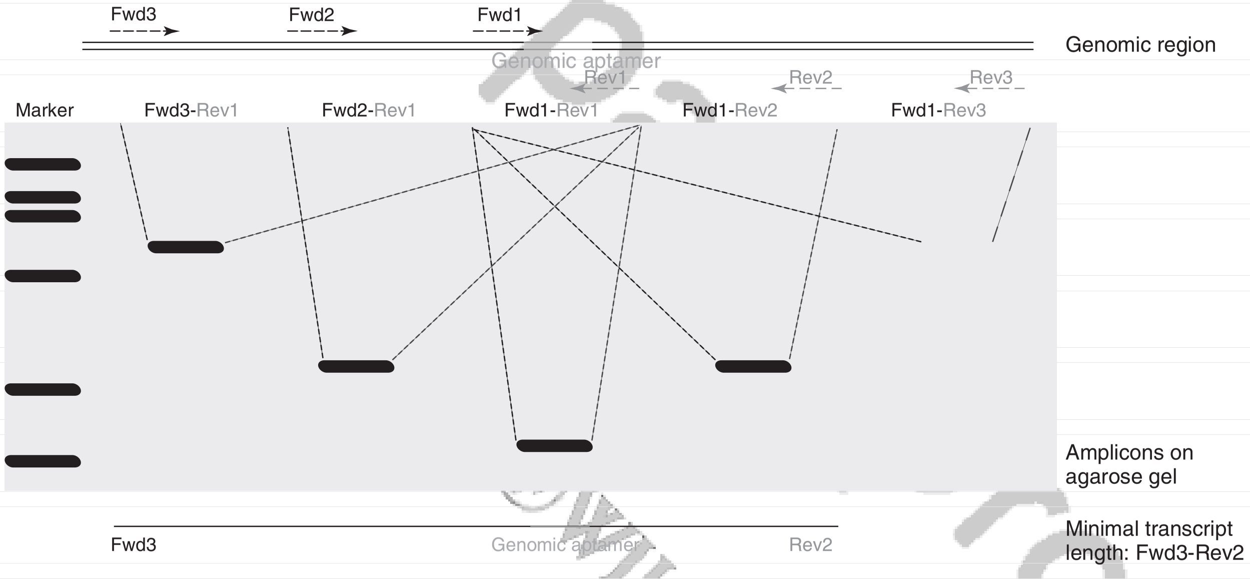


Figure 53.4 Reconstruction of the full-length transcript using a primer-walking method. Primer walking is a useful method to determine the full size of a transcript comprising the genomic aptamer selected in the genomic SELEX. The desired RNA pool is first reverse transcribed with a specific Rev1 primer, and then the genomic aptamer is amplified with PCR using Rev1 and Fwd1 primers. In consecutive amplification reactions, the 5′ end of the putative transcript is mapped: the reverse primer Rev1 remains fixed but the forward primer is placed further upstream (Fwd2 and Fwd3) as long as the amplicon is detectable on agarose gel. In the 3′ end mapping, the Fwd1 primer is used in the reverse transcription reaction and stays constant in the subsequent amplification reactions, whereas the reverse primers (Rev2 and Rev3) are shifted toward the 3′end of the transcript. Lack of Fwd1-Rev3 amplicon on agarose gel suggests that the 3′end of the transcript lies between sequences Rev2 and Rev3. When both extreme ends are reached, amplification with the most outer primers (Fwd3 and Rev2) is performed to prove that the detected transcript spans the full length.

many strategies that can be utilized to understand the importance of a particular protein–RNA interaction.

53.4
Conclusions and Outlook

Whole transcriptome analyses are delivering an unexpected high number of diverse transcripts, leading to the idea that probably every region of a genome is transcribed into RNA at some point of the organism's life cycle. Furthermore, the identity of a cell can be defined by its transcriptome. With this in mind, we need to find approaches to detect and functionally characterize those transcripts that are expressed rarely and at a low level. Because genomic SELEX is performed with libraries derived from the total DNA of an organism, every single part of the genome should be represented in the initial pool. We envision that in the near future, all genomic aptamers encoded within a genome, which interact with cellular proteins, RNAs, and metabolites, will have to be identified in order to describe the RNA regulon. To reach this goal, all available approaches will be necessary, and genomic SELEX will be a valuable approach to detect the low-abundance regulatory aptamers that otherwise might escape our attention.

Acknowledgments

Q6 Work in our laboratory● is funded by the Austrian Genome Research Program GEN-AU ''ncRNAs'' and the Austrian Science Fund FWF grant no.° F4308 to R.S. M.Z. was funded by a grant from the Austrian Ministry of Science and Research (GenAU project consortium ''non-coding RNAs'' D-110420-012-012 to N. Polacek)

References

1. Tuerk, C. and Gold, L. (1990) *Science*, **249**, 505–510.
2. Ellington, A.D. and Szostak, J.W. (1990) *Nature*, **346**, 818–822.
3. Gold, L., Polisky, B., Uhlenbeck, O., and Yarus, M. (1995) *Annu. Rev. Biochem.*, **64**, 763–797.
4. Umezawa, Y., Shimada, T., Kori, A., Yamada, K., and Ishihama, A. (2008) *J. Bacteriol.*, **190**, 5890–5897.
5. Ogasawara, H., Hasegawa, A., Kanda, E., Miki, T., Yamamoto, K., and Ishihama, A. (2007) *J. Bacteriol.*, **189**, 4791–4799.
6. Shimada, T., Fujita, N., Maeda, M., and Ishihama, A. (2005) *Genes Cells*, **10**, 907–918.
7. Kim, S., Shi, H., Lee, D.K., and Lis, J.T. (2003) *Nucleic Acids Res.*, **31**, 1955–1961.
8. Watrin, M., Von Pelchrzim, F., Dausse, E., Schroeder, R., and Toulmé, J.J. (2009) *Biochemistry*, **48**, 6278–6284.
9. Singer, B.S., Shtatland, T., Brown, D., and Gold, L. (1997) *Nucleic Acids Res.*, **25**, 781–786.
10. Zimmermann, B., Bilusic, I., Lorenz, C., and Schroeder, R. (2010) *Methods*, **52**, 125–132.

11. Lorenz, C., Von Pelchrzim, F., and Schroeder, R. (2006) *Nat. Protoc.*, **1**, 2204–2212.
12. Montange, R.K. and Batey, R.T. (2008) *Annu. Rev. Biophys.*, **37**, 117–133.
13. Lorenz, C., Gesell, T., Zimmermann, B., Schoeberl, U., Bilusic, I., Rajkowitsch, L., Waldsich, C., Von Haeseler, A., and Schroeder, R. (2010) *Nucleic Acids Res.*, **38**, 3794–3808.
14. Jensen, K.B., Atkinson, B.L., Willis, M.C., Koch, T.H., and Gold, L. (1995) *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 12220–12224.
15. Gopinath, S.C., Misono, T.S., Kawasaki, K., Mizuno, T., Imai, M., Odagiri, T., and Kumar, P.K. (2006) *J. Gen. Virol.*, **87**, 479–487.
16. Hirao, I., Harada, Y., Nojima, T., Osawa, Y., Masaki, H., and Yokoyama, S. (2004) *Biochemistry*, **43**, 3214–3221.
17. Tang, J., Xie, J., Shao, N., and Yan, Y. (2006) *Electrophoresis*, **27**, 1303–1311.
18. Lozupone, C., Changayil, S., Majerfeld, I., and Yarus, M. (2003) *RNA*, **9**, 1315–1322.
19. Yao, W., Adelman, K., and Bruenn, J.A. (1997) *J. Virol.*, **71**, 2157–2162.
20. Mayer, G., Ahmed, M.S., Dolf, A., Endl, E., Knolle, P.A., and Famulok, M. (2010) *Nat. Protoc.*, **5**, 1993–2004.
21. Tsuji, S., Hirabayashi, N., Kato, S., Akitomi, J., Egashira, H., Tanaka, T., Waga, I., and Ohtsu, T. (2009) *Biochem. Biophys. Res. Commun.*, **386**, 223–226.
22. Zimmermann, B., Gesell, T., Chen, D., Lorenz, C., and Schroeder, R. (2010) *PLoS ONE*, **5**, e9169.
23. Tereshko, V., Skripkin, E., and Patel, D.J. (2003) *Chem. Biol.*, **10**, 175–187.
24. Zywicki, M., Bakowska-Zywicki, K., and Polacek, N. (2011) (manuscript● submitted). Q7
25. Yan, T., Yoo, D., Berardini, T.Z., Mueller, L.A., Weems, D.C., Weng, S., Cherry, J.M., and Rhee, S.Y. (2005) *Nucleic Acids Res.*, **33**, W262–W266.
26. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009) *Genome Biol.*, **10**, R25.
27. Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., Elnakady, Y.A., Müller, R., Meese, E., and Lenhof, H.P. (2007) *Nucleic Acids Res.*, **35**, W186–W192.
28. Li, W. and Godzik, A. (2006) *Bioinformatics*, **22**, 1658–1659.
29. Frith, M.C., Saunders, N.F., Kobe, B., and Bailey, T.L. (2008) *PLoS Comput. Biol.*, **4**, e1000071.
30. Höchsmann, M., Voss, B., and Giegerich, R. (2004) *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **1**, 53–62.
31. Liu, Q., Olman, V., Liu, H., Ye, X., Qiu, S., and Xu, Y. (2008) *J. Comput. Chem.*, **29**, 1517–1526.
32. Hofacker, I.L., Fekete, M., and Stadler, P.F. (2002) *J. Mol. Biol.*, **319**, 1059–1066.
33. Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J., and Giegerich, R. (2006) *Bioinformatics*, **22**, 500–503.
34. Zolotukhin, A.S., Michalowski, D., Smulevitch, S., and Felber, B.K. (2001) *J. Virol.*, **75**, 5567–5575.
35. Kwang-Sun, K., Hyejin, R., Hyung, L.J., Meehyun, K., Taeyeon, K., Yool, K., Kook, H., Seol-Hoon, L., and Younghoon, L. (2006)● *Bull. Korean Chem. Soc*, 699. Q8
36. Shtatland, T., Gill, S.C., Javornik, B.E., Johansson, H.E., Singer, B.S., Uhlenbeck, O.C., Zichi, D.A., and Gold, L. (2000) *Nucleic Acids Res.*, **28**, E93.
37. Windbichler, N., Von Pelchrzim, F., Mayer, O., Csaszar, E., and Schroeder, R. (2008) *RNA Biol.*, **5**, 30–40.
38. Kim, H.J., Kwon, M., and Yu, J. (2007) *Bioorg. Med. Chem.*, **15**, 7688–7695.
39. Peritz, T., Zeng, F., Kannanayakal, T.J., Kilk, K., Eiríksdóttir, E., Langel, U., and Eberwine, J. (2006) *Nat. Protoc.*, **1**, 577–580.
40. Ule, J., Jensen, K., Mele, A., and Darnell, R.B. (2005) *Methods*, **37**, 376–386.
41. Wassarman, K.M. and Saecker, R.M. (2006) *Science*, **314**, 1601–1603.
42. Perocchi, F., Xu, Z., Clauder-Münster, S., and Steinmetz, L.M. (2007) *Nucleic Acids Res.*, **35**, e128.
43. Haddad, F., Qin, A.X., Giger, J.M., Guo, H., and Baldwin, K.M. (2007) *BMC Biotechnol.*, **7**, 21.

44. Scotto-Lavino, E., Du, G., and Frohman, M.A. (2006) *Nat. Protoc.*, **1**, 2742–2745.

45. Scotto-Lavino, E., Du, G., and Frohman, M.A. (2006) *Nat. Protoc.*, **1**, 2555–2562.

46. Scotto-Lavino, E., Du, G., and Frohman, M.A. (2006) *Nat. Protoc.*, **1**, 3056–3061.

47. Meyer, B.J. and Southern, P.J. (1993) *J. Virol.*, **67**, 2621–2627.