HHS Public Access

Author manuscript

*Chromosome Res*. Author manuscript; available in PMC 2020 July 31.

Published in final edited form as:

*Chromosome Res*. 2020 March ; 28(1): 111–127. doi:10.1007/s10577-020-09628-z.

# Seq'ing identity and function in a repeat-derived noncoding RNA world

**Rachel J. O'Neill**[1,2,3]

[1]Institute for Systems Genomics, University of Connecticut, Storrs, CT 06269

[2]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269

[3]Department of Genetics and Genome Sciences, University of Connecticut Health Center, Farmington, CT 06030

## Abstract

Innovations in high throughout sequencing approaches are being marshaled to both reveal the composition of the abundant and heterogeneous noncoding RNAs that populate cell nuclei and lend insight to the mechanisms by which noncoding RNAs influence chromosome biology and gene expression. This review focuses on some of the recent technological developments that have enabled the isolation of nascent transcripts, chromatin-associated and DNA interacting RNAs. Coupled with emerging genome assembly and analytical approaches, the field is poised to achieve a comprehensive catalog of nuclear noncoding RNAs, including those derived from repetitive regions within eukaryotic genomes. Herein, particular attention is paid to the challenges and advances in the sequence analyses of repeat and transposable element derived noncoding RNAs and in ascribing specific function(s) to such RNAs.

### Keywords

noncoding RNA; chromatin associating RNAs; R-loop; triplex; repeat annotation

## Introduction

Since the discovery of *Xist* (Brown et al. 1992), a long noncoding RNA that directs inactivation of the mammalian X chromosome, our understanding of the role RNAs play in chromosome biology has expanded well beyond the fundamental "RNA codes for proteins" dogma. The vast majority of RNAs produced by RNA polymerase II are mRNAs, and as such are capped and polyadenylated for subsequent transport outside of the nucleus, yet a surprising amount of RNA remains in the nucleus, where the bulk of RNA turnover occurs. These nuclear residents are incredibly diverse and include trimmed and spliced portions of pre-mRNAs, RNA debris from RNA decay, repeat-derived RNAs, antisense RNAs and other forms of noncoding RNAs (ncRNA)(reviewed in (Nozawa and Gilbert 2019; Palazzo and Lee 2015)). In addition to simply being isolated from the translation pipeline, nuclear ncRNAs are in an environment where they can interact directly with DNA and/or chromatin and thus exert an influence over fundamental processes such as transcription and genome stability (Mattick 2001; Mattick 2005; Mattick 2009).

Early experiments indicated that ~10% of the mass of chromatin was RNA (Holmes et al. 1972), considered at that time to be part of the ribonucleoproteinaceous structures comprising a static "nuclear matrix" (Fey et al. 1986a; Fey et al. 1986b) supporting nuclear organization. Today, the idea of a static matrix has been abandoned (Pederson 2000) in favor of models invoking a dynamic nuclear organization of which RNA is an integral part. Since nuclear ncRNA content can vary across different cellular contexts, ncRNAs may serve as an architectural feature required for establishing specific chromatin states (Caudron-Herger and Rippe 2012; Mele and Rinn 2016), and thus foster a permutable form of control over genome organization (Michieletto and Gilbert 2019; Nozawa and Gilbert 2019). Additionally, sequence variation inherent to many ncRNAs, particularly repeat-derived ncRNAs, could provide a potent source of species-specific genome organization and evolutionary novelty (Hall and Lawrence 2016; Kapusta et al. 2013; Necsulea et al. 2014).

The capacity of RNAs to associate with chromatin, either through DNA or protein interactions, indicates they may act as molecular signals, regulators, guides and/or scaffolds (Chu et al. 2011; Guttman and Rinn 2012; Rinn et al. 2007). Moreover, they may contribute to the regulation of entire chromosomes, as *Xist* does, or specific chromosomal domains within a cell and thus may mediate specific cellular processes such as centromere function and chromosome inheritance (e.g. (Carone et al. 2009; Carone et al. 2013; Topp et al. 2004; Wong et al. 2007)) and thus foster chromosome evolution (Brown et al. 2012; Brown and O'Neill 2010; O'Neill and Carone 2009). Revealing the composition of RNAs that influence chromosome biology, defining how they interact with the genome and/or chromatin, and ascribing a cellular function, if any, to these interactions are among the grand challenges at the frontier of chromosome research.

These challenges are being met by innovations to high throughput sequencing (HTS) approaches (a.k.a. the growing menagerie of "…-seq"s) to the study of RNA. Coupled with revolutionary advances in long read sequencing, genome assembly and annotation methods, comprehensive cataloging of nuclear ncRNA is underway with a view towards understanding the cellular functions of these heterogeneous and fascinating biomolecules. This review focuses on some of the recent technological developments that have enabled isolation of both chromatin associated RNAs and DNA-associated RNAs. Moreover, computational approaches and initiatives to achieve chromosome level genome assemblies are discussed in light of the challenges in studying such RNAs.

## How do we define ncRNA?

Given that we have known about ncRNAs in the nucleus for over 50 years, why has it been so challenging to ascribe reasons for their existence? The first challenge, and arguably one that has yet to be fully overcome, is clarity on how one defines the component of nuclear RNAs that are *noncoding;* in other words, what exactly is a ncRNA? The phrase "noncoding RNA" at face value could refer to any RNA molecule that does not lead to a translated protein. However, this would include spliced introns, degradation products and RNA debris, as well as RNAs that are predictably transcribed and have a structured transcription unit, such as rRNAs and tRNAs. Current nomenclature distinguishes ncRNAs rather arbitrarily as either small RNAs of 200 nt and less, or RNAs 200 nt and longer, referred to as long or large

ncRNAs (lncRNAs) and long intergenic noncoding RNAs (lincRNA). Small RNAs are further classified into groups based on function, biogenesis, and/or other biochemical features (e.g. snoRNAs, tRNAs, miRNAs, piRNAs, etc) (Dupuis-Sandoval et al. 2015; Kim et al. 2009; Oberbauer and Schaefer 2018; Ozata et al. 2019; Pan 2018; Treiber et al. 2019).

Beyond the size designation of the larger ncRNAs fraction as >200 nt, relatively little else classifies or distinguishes lncRNAs and for many, the full transcription unit has not been adequately annotated in genome assemblies. Of the few lncRNAs that have been heavily studied, the underlying transcription units are often quite long, such as the 2.3 kb *H19*, the first lncRNA annotated in human (Brannan et al. 1990), the ~8 kb MALAT1 (Tripathi et al. 2010), the 17 kb *Xist* (Brown et al. 1992), and the 2.2 kb HOTAIR (Rinn et al. 2007). These RNAs, along with a few other well characterized transcripts, are known to participate in specific cellular functions, such as splicing, translation, RNA editing and transcription (see (Qian et al. 2019) for a review). The overall length of these lncRNAs has no doubt facilitated their annotation in assembled and well curated genomes (i.e. mouse and human), while smaller or more divergent lncRNAs have been more challenging to catalog and study.

## The road to defining ncRNA function

Recent comparative studies utilizing transcriptomic datasets and available genome assemblies have revealed a collection of lncRNAs with enough sequence conservation across species to enable at least partial annotations and functional inferences (Cabili et al. 2011; Guttman et al. 2009; Marques and Ponting 2009; Necsulea et al. 2014). However, the low sequence conservation among the vast majority of lncRNAs limits the ability to use sequence alone for annotation or to surmise functions. Further complicating the classification of lncRNAs is the observation that transposable element sequences (TEs) contribute to a significant portion of the lncRNA repertoire (Kapusta et al. 2013). In fact, TEs are ubiquitous in lncRNAs in vertebrates and account for a large fraction of total ncRNA sequences (Kapusta et al. 2013).

It is possible that the insertion of exonic portions of TEs into lncRNAs, termed Repeat Insertion Domains of LncRNAs (RIDLs) (Johnson and Guigo 2014), represent exaptations of TE sequences (Johnson 2019). For example, a short sequence motif found in several unrelated lncRNAs was identified in human cells that increases nuclear enrichment through binding to HNRNPK (Lubelsky and Ulitsky 2018). This motif, SIRLOIN (SINE-derived nuclear RNA localization), overlaps with antisense sequences of the Alu SINE repeat element, indicating the nuclear-retention of RNAs mediated by this motif may be part of a pathway to regulate transcripts that contain Alu insertions (Lubelsky and Ulitsky 2018). Some TE insertions, however, may have limited or no impact to the function of a lncRNA and thus are simply not selected against, as is the case with lineage-specific TE insertions found in the *Xist* lncRNA (Kapusta et al. 2013). Alternatively, the first portion of many lncRNAs, and often the entire lncRNA itself, is comprised of TE sequences, indicating that TE insertions in genomic sequences can provide the transcription start site, and subsequently produce a new lncRNA (Kapusta et al. 2013). Thus, divergence of genomic TE content across different lineages provides fodder for the recruitment of lineage-specific lncRNAs (Kapusta et al. 2013).

Further confounding the study of ncRNAs utilizing cross-species sequence comparisons is the fact that divergent, non-TE repeats are often expressed. Satellite repeats, for example, are a class of ncRNA that are found in most eukaryotes (reviewed in (Biscotti et al. 2015; Hartley and O'Neill 2019; Talbert and Henikoff 2018)). Satellite-derived ncRNAs are produced from genomic loci that vary in composition from simple repeats consisting of a small number nucleotides organized in tandem arrays to longer satellite arrays of repeated units that are each 10's to 1,000's of bases in length. In many cases, these ncRNA producing repeats are found in clusters in specific chromosome regions, such as large heterochromatin blocks on chromosome arms, centromeres and telomeres, linking transcription of highly repetitive ncRNAs to chromosome function.

Given their abundance and diversity, teasing apart functional from non-functional ncRNAs has been challenging and even controversial (e.g. (Graur et al. 2013; Palazzo and Lee 2015; Pennisi 2012). A series of commentaries highlight some of the problems with the use of the term "functional" when applied as a blanket descriptor to a ncRNA (Doolittle 2018; Laubichler et al. 2015; Palazzo and Lee 2015). The issues lie in the fact that "function" is interpreted differently in molecular (*what does the ncRNA do*) vs evolutionary (*why does the ncRNA exist*) contexts. Recently, a new lexicon to clarify "function" has been proposed, referred to as the "Pittsburg model of function". In this model, ncRNAs are further classified into five categories based on the depth and context of genetic information available to support functional classification (Table 1) (Keeling et al. 2019). Such a refined frame work for presenting data on ncRNAs is long overdue; the application of these categories offer clarity for the field as we navigate discoveries of the myriad chromatin-associated ncRNAs across different cell types and developmental stages, and particularly across different species (Doolittle 2018).

## Entering a new era of transcriptome profiling

Early genomics approaches that were designed to assess transcriptional output across different samples often employed exon-based screens (e.g. microarrays), ignoring repeat-derived and intergenic ncRNAs, thus rendering only a partial view of transcriptome dynamics. HTS approaches support transcriptome-scale sequencing (RNA-seq) that include ncRNAs by capturing potentially all RNAs present in a given sample, representing newly transcribed RNAs, stable RNAs, and RNAs heading for imminent decay. While RNA-seq was the first global transcriptomic approach enabled by HTS, new techniques have been developed to score the density of RNA polymerase II binding across the genome or to measure nascent, active transcription and delineate transcription start sites (TSSs), eliminating the need to account for the variable half-life of different RNAs. Sequencing data outputs are subsequently mapped to a reference genome and intersected with gene annotations to tease apart mRNAs from cell-specific ncRNAs.

Immunoprecipitation of RNA polymerase II (Churchman and Weissman 2011; Larson et al. 2014; Nojima et al. 2015) and isolation of insoluble chromatin (Weber et al. 2014) have been used to identify nascent transcripts, revealing the involvement of nucleosome positioning in transcription elongation (Churchman and Weissman 2011). However, variation in antibody specificity or the efficiency of chromatin purification may affect experimental outcomes of

such approaches (Mahat et al. 2016). Adaptations to nuclear run-on experiments (see (Smale 2009)) that enable genome-wide capture of nascent transcripts bypass immunoprecipitation, instead using labels incorporated into nascent RNA to isolate purified transcripts. In GRO-seq (global run-on sequencing), bromouridine is used to label nascent RNAs (Core et al. 2008); the incorporation of multiple labelled nucleotides in the run-on reaction allows a mapping resolution of 10's of bases. In a modification of this technique, PRO-seq (precision run-on sequencing), biotin-labelled NTP's are added to the run-on reaction and nascent transcripts with an incorporated biotin-NTP are sequenced from the 3' end to afford single bp resolution of the site of RNA polymerase engagement with nascent RNA when mapped back to a reference genome (Figure 1A)(Kwak et al. 2013; Mahat et al. 2016). PRO-cap, an adaptation of PRO-seq, incorporates steps to repair the 5'end of the nascent transcript (i.e. capping) for adaptor ligation and subsequent sequencing from the 5' end, providing TSS identification (Figure 1A)(Kwak et al. 2013; Mahat et al. 2016). Further building upon the principle of PRO-seq is the recent development of ChRO-seq (chromatin run-on sequencing) (Chu et al. 2018), wherein the input material is not nuclei isolated from cells, but rather is fractionated, insoluble chromatin that includes engaged RNA polymerase II (Wuarin and Schibler 1994), increasing the diversity of samples that can be queried.

By uncovering nascent transcripts independent of innate transcript stability, a model of transcription initiation and elongation is emerging, revealing some of the fundamental signatures of RNA polymerase II activity. For example, promoters and enhancers share the genomic signal of divergent transcription profiles for nascent transcripts, but can be distinguished based on the transcription level and stability of the resulting transcripts (Core et al. 2008). From these observations, it appears that histone modifications that vary between promoters and enhancers are not necessarily dictated by the type of regulatory element at which they reside, but rather are associated with specific transcriptional signals. Revealing patterns of nascent transcription at the genome-scale is supporting more accurate annotations of regulatory regions and active transcription across different cell types/stages, independent of factors that can influence transcript abundance in the nucleus. Furthermore, ongoing efforts to capture a view of the changing transcriptional landscape among different tissues, conditions, developmental stages and across different species is starting to reveal the true, and indeed extremely diverse and dynamic repertoire of ncRNAs. Understanding the fate of these ncRNAs and delineating whether the ncRNA sequence itself, the act of its transcription, or both, impact genome dynamics requires a combination of innovative tools to capture ncRNAs, delineate their interacting partners and decipher their mode of function at the genome-scale.

## Looking beyond transcription for ncRNA partners

To begin to understand the ways in which ncRNAs may impact genomes at both local (gene transcription, local chromatin states) and regional (chromosomal regions and entire chromosomes) scales, one must consider *how* and *where* ncRNAs associate with chromatin beyond their site of nascent transcription (Guttman and Rinn 2012). ncRNAs can associate with chromatin in *cis* and/or *trans* through either direct RNA-DNA interactions or through an intermediary, such as chromatin-associated protein or protein complex. Different methods have been developed to tease apart ncRNAs based on these varied interactions. From these

studies, we have not only begun to unravel the ncRNA-chromatin interactome, but have gained an appreciation for the varied, and in some cases seemingly contradictory, roles ncRNAs play in processes such as gene regulation, chromosome function and genome organization.

## RNA: DNA partnerships - R-loop detection

Direct RNA-DNA interactions occur through complementary base pairing of DNA with RNA, resulting in the formation of a three stranded structure consisting of a DNA:RNA hybrid and the displaced complementary DNA strand (Drolet et al. 1995; Thomas et al. 1976). Tiny, three-stranded "bubbles"occur during RNA-priming of DNA replication and at the immediate site of RNA polymerase as transcription occurs; longer, stable forms of these three-stranded structures are called R-loops (RNA moiety loop) (Thomas et al. 1976). R-loops were originally considered an extension of the RNA:DNA hybrid found within the RNA polymerase II transcription bubble (Westover et al. 2004), but it appears more likely that they result from the fold-back of nascent RNA as it exits the polymerase, known as an RNA thread back model (Roy et al. 2008).

In normal cells, an equilibrium is maintained that balances the formation and resolution of R-loops to support genome integrity (e.g. (Chakraborty and Grosse 2011; El Hage et al. 2010; Zhou et al. 2014)). Although R-loop formation has been linked to genome instability and disease (reviewed in (Santos-Pereira and Aguilera 2015)), R-loop structures may also serve important roles in normal cells. For example, R-loops facilitate the programmed immunoglobulin class switch recombination in B cells (Roy et al. 2008; Yu et al. 2003). Bolstered by computational predictions that R-loops could be prevalent across the genome (Ginno et al. 2012), genome-scale methods have been developed to identify R-loops and potentially reveal novel regulatory functions.

A genome-wide assessment of R-loops that form under normal cellular conditions was afforded by the development of an antibody (S9.6) to RNA:DNA duplex structures specifically, independent of nucleic acid sequence (Boguslawski et al. 1986). Immunopreciptation with the S9.6 antibody coupled with deep sequencing, a technique known as DRIP-seq (DNA:RNA immunoprecipitation coupled to high-throughput sequencing), results in a genome-wide map in R-loop sites in specific tissues (Ginno et al. 2012). Variations of this technique, including S1-nuclease DRIP-seq (S1DRIP) (Wahba et al. 2016), bisulfide DNA:RNA immunopreciptation (bis-DRIP) (Dumelie and Jaffrey 2017), and RNA:DNA immunoprecipitation (RDIP) (Nadel et al. 2015) have built upon the original DRIP-seq method to collectively develop preliminary maps for R-loop formation in specific cells. However, these techniques have some limitations in that the harsh preparation of the chromatin for immunoprecipitation may disrupt all but the most stable R-loops (Yan et al. 2019) and the S9.6 antibody may also recognize dsRNA (Hartono et al. 2018), complicating data interpretation.

Alternative methods employ a form of RNAse H, which has an affinity towards RNA:DNA heteroduplexes that is catalytically incapable of cleaving RNA. These methods, DRIVE (DNA:RNA in vitro enrichment) (Ginno et al. 2012) and R-ChIP (R-loop chromatin enrichment) (Chen et al. 2017), no longer rely on S9.6, overcoming doubts about the

specificity of the antibody, but still suffer from challenges presented by the affinity purification steps. A method that no longer relies of affinity purification has been developed that is based on the cleavage under targets and release using nuclease (CUT&RUN) approach (Skene and Henikoff 2017) combined with RNAse H specificity for RNA:DNA heteroduplexes. This approach, MapR, revealed previously undetected transient R-loops at promoters and active enhancers (Yan et al. 2019).

Collectively, these types of approaches have revealed that R-loops are found in the terminators and enhancers of some genes, and thus can influence transcriptional control. For example, R-loops that form immediately following a transcription start site in a CpG island prevent DNA methylation of the underlying gene via DNA methyltransferase 3B1, thus facilitating transcription activation (Ginno et al. 2012). Moreover, the overlap between R-loops and GC-skew in the 5' end of genes is also correlated with the deposition of histone marks of active transcription, including H3K4me3, H4K20me1 and H3K79me2 (Ginno et al. 2013; Ginno et al. 2012), implicating these R-loops as intermediaries in chromatin dynamics. R-loops may also function in transcript termination processes, such as RNA polymerase II pausing (Skourti-Stathaki et al. 2011) and induction of antisense transcription. When antisense transcripts are formed, these ncRNAs trigger dsRNA formation and the deposition of H3K9me2 and HP1γ, marks of repressive heterochromatin (Skourti-Stathaki et al. 2014). The ability of R-loops to trigger the formation of heterochromatin, histone H3 S10 phosphorylation and chromatin condensation (Castellano-Pozo et al. 2013) may facilitate transcript silencing through establishment of repressive chromatin, but may also lead to replication fork stalls and DNA fragility/breakage (Castellano-Pozo et al. 2013; El Achkar et al. 2005; Groh et al. 2014).

R-loops, while largely considered in the context of *cis* ncRNA interactions, can be formed by *trans*-acting RNAs (Wahba et al. 2013), indicating that a single RNA species may affect many loci across the genome that share a similar sequence composition, such as repeated elements and satellite arrays. The single stranded DNA binding protein RPA (replication protein A) was recently identified at human centromeres. While RPA is known to participate in ATR (ataxia telangiectasia mutated and Rad3-related) kinase activation targeting DNA damage and stalled replication forks (Zou and Elledge 2003), normal centromeres do not appear to recruit RPA through damage response mechanisms (Minocherhomji et al. 2015). Instead, RPA is recruited by the single stranded DNA that is displaced in R-loops, indicating R-loops may be a general feature of centromeres (Kabeche et al. 2018). Indeed, staining with the S9.6 antibody indicates that R-loops are prevalent at human centromeres and their association with ATR activation implicates that the formation of R-loops may be required to for activation of Aurora B and accurate chromosome segregation (Kabeche et al. 2018). It is possible that nascent transcripts forming centromeric R-loops are acting *in cis,* facilitated by the repeat-derived transcripts produced in active centromeres (Carone et al. 2009; May et al. 2005; McNulty et al. 2017; Rosic et al. 2014; Ugarkovic 2005). Alternatively, centromeric R-loops may be mediated by a *trans*-activating ncRNA, perhaps recognizing the repeat motif present in CENP-B DNA binding sites shared across divergent and chromosome specific centromeric satellites (Masumoto et al. 1989). As the genomic landscape of highly repeated regions such as centromeres become more accessible (see below), RNA-DNA and RNA-Chromatin sequencing approaches combined with innovative computational approaches

offer promise in revealing the complex RNA interactions that mediate centromere function and chromosome stability.

**RNA: DNA partnerships - Triplex detection—**Without disrupting the hydrogen bonds of the DNA helix, RNA is still capable of direct nucleic acid interaction via the formation of a DNA:RNA triple helix (an RNA:DNA triplex, or simply "triplex" (Felsenfeld and Rich 1957) (not to be confused with the three strandedness of R-loops). A triplex forms when RNA binds to the major groove of a purine-rich stretch of duplex DNA through Hoogsteen or reverse Hoogsteen hydrogen bonding (reviewed in (Bacolla et al. 2015; Li et al. 2016)). Triplex formation has been shown to affect chromatin state through the recruitment of epigenetic modifiers, particularly when the interacting RNA in the triplex structure is a lncRNA. For example, local tethering of PRC to *Foxf1*, and subsequent trimethylation of histone 3 lysine 27 residues (H3K27me3), is mediated by a triplex containing the *Fendrr* lncRNA (Grote and Herrmann 2013). The ability of lncRNA-triplex structures to act as scaffold structures to recruit chromatin remodeling complexes (Bacolla et al. 2015) offers another means by which lncRNAs can impact gene regulation and chromosome biology. If tandem arrays of repeats (simple, satellite, TE, etc), such as those found in centromeres, pericentromeres, telomeres and heterochromatin blocks, produce triplex structures, scaffolding and chromatin factor recruitment could impact regional chromosome function and/or sub-cellular localization. For example, rDNA promoter methylation and regional silencing of rDNA transcription is initiated by the recruitment of DNMT3B, and subsequent interactions with the nucleolar remodeling complex NorC, following triplex formation with an antisense RNA (Bierhoff et al. 2010; Schmitz et al. 2010).

Computational methods have led to the prediction of the possible sites in the human genome that could form RNA:DNA triplex structures (Buske et al. 2012; Goni et al. 2004; Jalali et al. 2017; Wu et al. 2007) indicating that at least one putative triplex site exists for each gene, promoter and intergenic region. To avoid the isolation of RNA-DNA interactions formed through a protein intermediary, *in vivo* approaches to isolate RNA:DNA triplex structures should not rely on cross-linked samples. Rather, a recently described pair of methods (Senturk Cetin et al. 2019) removes free RNA from RNA that is bound to DNA through Hoogsteen pairing using a urea/NP40 extraction to isolate chromatin that is then treated with proteinase K to remove RNA bound to DNA via a protein intermediary. DNA:RNA triplex structures are further enriched using two complementary methods, paramagnetic bead selection and RNA immunoprecipitation via an anti-DNA antibody. Isolated RNA is then subjected to strand-specific RNA-seq and the sequencing data mapped back to the genome.

These genome-scale methods revealed that a surprising number of protein coding genes produced RNAs that associated with DNA in triplex structures (Senturk Cetin et al. 2019). These RNAs may represent noncoding isoforms of protein coding transcripts or other ncRNAs embedded within the transcript, such as miRNAs or antisense RNAs (Ayupe et al. 2015), or could be intragenic enhancer RNAs (Andersson et al. 2014; Cinghu et al. 2017). For these protein coding genes, the triplex may be fundamental to the gene's function or transcriptional output (Senturk Cetin et al. 2019). In addition to these intragenic ncRNAs, an abundance of TEs and repeated elements were identified in these screens as triplex bound

RNAs (Senturk Cetin et al. 2019), revealing the possibility that repeat-derived RNAs could interact with multiple genomic locations sharing sequence identity.

Given the observation that repeats within specific TEs can act as super-enhancers (Goni et al. 2004; Soibam 2017) or control nuclear localization of RNAs, such as SIRLOIN elements (Lubelsky and Ulitsky 2018), triplexes formed with repeated sequences could provide a potent means for repeat-bearing TEs to interact with DNA *in trans*. In support of this idea is the recent discovery that a defined, short motif is shared between *Xist* RNA and LINE1s in mouse and human that is predicted to mediate redundant lncRNA-triplex structures between *Xist* RNAs and X-linked LINEs during X-inactivation (Matsuno et al. 2019). Intriguingly, while a redundant UC/TC (r-UC/TC) motif was found in the two eutherian species, a redundant AG (r-AG) motifs was found to be shared between the putative marsupial X inactivation mediating lncRNA, *RNA-on-the-silent X (Rsx),* and LINEs within opossum. The lineage-specific convergence in redundant motif sequences shared between lncRNAs involved in X chromosome inactivation and X-linked LINEs may indicate that lncRNA-LINE triplexes are essential for inactivation of the X in females (Matsuno et al. 2019).

**Beyond RNA:DNA interactions—**The identity of a specific RNA's interacting partners can be revealed by screening the entire genome for those partners (also referred to as a ONE vs MANY approach). Three techniques employing the ONE vs MANY approach, ChIRP (chromatin isolation by RNA purification)(Chu et al. 2011), RAP (RNA antisense purification)(Engreitz et al. 2013), and CHART (capture hybridization analysis of RNA targets)(Simon et al. 2011), isolate all interacting partners for a specific RNA using biotinylated, complementary oligonucleotides for the RNA in cells that have been treated with cross linking reagents to allow isolation of nucleic acid - protein interactions. Where these applications vary are in the cross liking reagent and chromatin treatments and in the design of the oligonucleotide probes for the target RNA (Simon 2016). Long probes are used in RAP and probes spanning the entire RNA (i.e. tiling) are used in both RAP and ChIRP, alleviating the need to predict accessible parts of an RNA molecule when in its folded form. CHART, on the other hand, utilizes RNAse H mapping to identify accessible regions of the RNA target for oligonucleotide probe design. Complexes isolated from these techniques can be further purified to identify RNA-protein partners via mass spectrophotometry (e.g. (West et al. 2014)), or the genomic locations of RNA interactions using deep sequencing (Chu et al. 2011; Engreitz et al. 2013; Simon et al. 2011). While useful in guiding the study ncRNAs of unknown function, these hybridization-based approaches also come with some caveats as artifacts such as hybridization to off target DNA or RNAs, directly or indirectly, can undermine precision of the data analysis (Simon 2016).

Alternative approaches for revealing RNA-chromatin interactions have been developed that do not rely on a known RNA and thus scan for all RNAs that may interact with chromatin. CheRNA-seq (chromatin-enriched RNA-seq), an approach to isolate chromatin-proximal RNAs, uses nuclear fractionation followed by RNA deep sequencing (Werner and Ruthenburg 2015) to separate soluble mRNAs and ncRNAs from RNAs that may function at the chromatin interface. Using a urea and Nonidet-P40 solution to separate released mRNAs from ternary complexes of RNA polymerase II and its DNA template (Bhatt et al. 2012;

Wuarin and Schibler 1994), cheRNAs are isolated and sequenced at relatively high depth to ensure capture of low-abundance RNAs (Werner and Ruthenburg 2015).

Based on several of the same principles as the ONE vs MANY approaches, these MANY vs MANY approaches also begin with cross linking RNA-protein complexes. Relying on proximity ligation, these methods employ a bivalent and biotinylated linker molecule that consists of single stranded RNA at one end and double stranded DNA at the other. Proximity ligation, wherein protein complexes that bring RNA and DNA together (i.e. on chromatin), is enabled by a bivalent linker containing a biotinylated bridge sequence, ligating the RNA portion to nascent RNA and the double stranded DNA portion to proximal DNA. The MANY vs MANY techniques that rely on this type of proximity ligation, RNA-DNA heteroduplex capture include (Figure 1B): MARGI (mapping RNA-genome interactions) (Sridhar et al. 2017), GRID-seq (global RNA interactions with DNA by deep sequencing) (Li et al. 2017) and ChAR-seq (chromatin associate RNA sequencing) (Bell et al. 2018). One technical component that distinguishes MARGI from ChAR and GRID is that the proximity ligation in the former is performed on extracted chromatin complexes (Sridhar et al. 2017) while in the latter two, proximity ligation is performed *in situ* on intact nuclei (Figure 1B) (Bell et al. 2018; Li et al. 2017). Further distinguishing GRID and ChAR approaches is the post ligation processing. GRID-seq includes a restriction enzyme digestion following reverse transcriptase conversion of the RNA-DNA duplex to a cDNA-DNA duplex. The targeted digestion 19–23bp from the bridge sequence (this is done using the enzyme *MmeI* whose recognition sequence is within the bridge but cuts 18–20bp away) allows size selection prior to sequencing to enrich for fragments containing RNA-DNA ligations (Figure 1B, left) (Li et al. 2017). ChAR-seq, on the other hand, isolates 100–125bp each of the DNA and cDNA sequences (Bell et al. 2018). The five fold greater length of sequence obtained in ChAR-seq supports more accurate mapping to the reference, which can influence the interpretation of global RNA-seq data (Figure 1B, right)(Li et al. 2017), particularly when repeats are considered.

From these collective approaches, a model of how transcription, transcripts and chromatin remodeling are coordinated is emerging that indicates there is no single rule that defines lncRNA-chromatin interactions. For example, these studies confirmed the previous work demonstrating some lncRNAs interact *in cis* near their site of transcription while others work across larger regions or even across different chromosomes. Surprisingly, promoters/TSSs were found to have an association with *trans*-interacting RNAs (Li et al. 2017; Sridhar et al. 2017) while enhancers were found to associate with transcripts of their regulating gene (Li et al. 2017). Regions with *trans*-interacting RNA attachment were also correlated with open chromatin histone marks, H3K27ac and H3K4me3 (Sridhar et al. 2017), but this correlation was not consistent across all RNAs. snoRNAs interactions, for example, are enriched for marks of heterochromatin rather than active transcription (Bell et al. 2018).

The application of ChAR-seq to *Drosophila* cells indicated that transcription-associated RNAs are enriched at TAD boundaries, linking RNA-chromatin interactions to 3D genome architecture (Bell et al. 2018). In fact, a recently described technique RADICL-seq (RNA and DNA interacting complexes ligated and sequenced) was applied to mouse cells, revealing an enrichment of RNA-chromatin interactions at TAD boundaries specifically

associated with TEs (Bonetti et al. 2019), indicating such interactions may be a conserved mechanism for the control of genome organization.

## The final frontier: incorporation of repeats and TEs in ncRNA data analyses

The descriptors for genome-scale studies often include "all" rather than "many", as used in this review. However, the use of "all" is misleading as it implies that the entirety of the genome serves as a reference for mapping NGS datasets. Rather, it is understood that these data analyses are contemporaneous with *available* genome sequence. Herein lies one of the major challenges for the field: *how do we obtain a comprehensive understanding of RNA-chromatin relationships, particularly when ncRNAs containing, or derived from, repeats are considered, when we have yet to fully annotate the complete sequence content of the genome?* Reference genomes for most model species are not chromosome-level to the extent that all scaffolds are provided with both chromosomal assignment and linear arrangement (Lewin et al. 2019). An estimated 10% of the human reference genome (hg38), considered to be one of the best eukaryotic genome assemblies to date, remains on orphan scaffolds enriched for repeat-dense regions of the genome, such as rDNA loci, centromeres, interstitial repeat clusters, telomeres and pericentric regions (Altemose et al. 2014; Miga 2015; Rosenbloom et al. 2015).

The short read lengths inherent to modern high-depth sequencing technologies, coupled with the difficulty in assigning highly similar repeats to a specific location in a reference genome, are major limitations to closing gaps in genome assemblies for most complex eukaryotic genomes. Techniques such as Hi-C (Lieberman-Aiden et al. 2009) greatly improve the ability to assign contigs to chromosomes (Burton et al. 2013; Kaplan and Dekker 2013; Marie-Nelly et al. 2014), but are not capable of building full, chromosome-scale scaffolds on their own (Lewin et al. 2019). Despite these seemingly insurmountable challenges, researchers have developed an ever-growing set of tools to both catalog and analyze repeats across the genome. For example, RepeatMasker is used to classify repeats based on a compiled database (such as Repbase (Jurka 2000)) using gapped aligners, affording the ability to classify highly variable sequences (Smit et al. 2015). While traditionally considered for repeat annotations in genome assemblies, this tool can be applied to HTS reads, regardless of their source (RNA or DNA from various applications, as described in this review) (Figure 2A).

In like fashion, if a particular repeat class is known, any sequences within HTS datasets with identity to this class can be isolated from a pool of sequences and a k-mer approach can be used to define the phylogenetic relationships among repeats (Smalec et al. 2019) or derive graphical models of repeat content (Miga et al. 2014; Rosenbloom et al. 2015). For example, the linear order and frequency of individual repeats within large tandem arrays, exemplified by alpha satellites in human centromeres, was inferred from linked pairs of sequencing reads from whole genome shotgun data (Figure 2B) (Miga et al. 2014; Rosenbloom et al. 2015). In addition, the frequency and classification of transposable element insertions into repeat arrays can be assessed using this graphical model approach (Figure 2C).

In the absence of a complete, telomere-to-telomere genome assembly, other approaches can be applied to study the contribution of repeats to the RNA-chromatin relationship. Current mapping tools, such as BWA and Bowtie2, (Langmead and Salzberg 2012; Li 2013), are typically implemented to report unique mappers only; in other words, sequencing reads that map to more than 1 location in the queried genome are ignored. In doing so, the contribution of repeats are often overlooked or minimized. To complement standard mapping strategies, HTS datasets can be explored for repeat content via genome independent methods. For example, sequencing reads can be annotated for repeat content using Repeat Masking pipelines to reveal the types of repeats and their frequency within a given HTS dataset. K-mer based approaches can also be used to classify reads into specific repeat groups (Lefebvre et al. 2003; Marcais and Kingsford 2011). Approaches that derive *de novo* assemblies from HTS data have also been developed, such as RepARK (Koch et al. 2014) (Figure 3), REPdenovo (Chu et al. 2016), and ChIPtigs from ChIP-seq data (He et al. 2015). These methods rely on k-mers rather than alignments to build contigs, but in doing so less-frequent and rare k-mers may be lost in the final assembled contigs. While none of these methods offer a full replacement for a reference genome, they illuminate regions that are either missing, or highly variable, when compared to a single reference genome, such as those enriched in TEs, satellites and/or tandem arrays.

The arrival of long-read sequencing technologies in the genome sequencing market has provided a boost to the initiatives to derive genome assemblies that include repeats, particularly those with relevance to chromosome segregation. For example, the genome sequence of the koala, based on ~58X PacBio long-read sequencing and polishing with 30X Illumina short read sequencing, afforded assembly of scaffolds that contained centromeres (Johnson et al. 2018). These scaffolds were functionally annotated with ChIP-seq data for a pool of centromere-binding proteins, revealing that transposable elements are a major contributor to centromere identity in this species (Johnson et al. 2018). In Drosophila, centromere scaffolds were assembled with the aid of long-read data from PacBio and chromosome assignment was achieved using oligo-paints derived from these assemblies (Chang et al. 2019). The annotations of repeats containing centromeric histones using a combination of ChIP-seq and ChIP-tig analyses showed that islands of transposable elements within satellite arrays define chromosome-specific centromere identity in Drosophila (Chang et al. 2019).

## Where do we go from here?

The combination of long-read sequencing data (i.e. Oxford Nanopore, PacBio) and applications such as Hi-C, accompanied by increasingly accessible high-coverage short read sequencing are supporting efforts to complete telomere-to-telomere (T2T) assemblies for a reference human genome. Successes in this approach have been realized for the X chromosome (Miga et al. 2019), and are being expanded to the entire human genome (Miga et al. 2019). New computational tools (e.g.(Bongartz 2019; Russo et al. 2019; Shafin et al. 2019)) and assembly improvements for model species are facilitating additional analyses with existing "-seq" datasets from diverse applications. Moreover, genome-scale applications developed for short-read NGS technologies are being modified to incorporate

long-read sequencing to enable more accurate mapping with the inclusion of junctions between repeats and unique sequences and the assembly of tandem arrays of repeats.

Such advances will enable a full appreciation of the dynamic and diverse RNA-chromatin relationships that exist in eukaryotic genomes. However, a major challenge will be to "carryover" exiting datasets developed to study RNA-chromatin interactions to new assemblies and repeat annotation pipelines as they emerge. Furthermore, the diversity of genomes across individuals within a population should be incorporated into studies exploring the role of ncRNAs in instability and disease. The lack of T2T-scale genomes that support comprehensive comparative approaches must be overcome (Doolittle 2018; Lewin et al. 2018) to fully appreciate conserved RNA-chromatin functions as well as divergent functions that enable evolutionary novelty (Kapusta et al. 2013). This is an exciting time where we are witnessing a re-emergence of the synergy of RNA biology and chromosome biology through innovations in genomics.

## Acknowledgements

RJO is supported by the National Science Foundation (1613806 and 1643825) and the National Institutes of Health (R01GM123312 and R21CA240199). Editorial comments were kindly provided by M. O'Neill; title was suggested by Michelle Neitzey.

## Abbreviations

| | |
|---|---|
| **ncRNA** | noncoding RNA |
| **lncRNA** | long noncoding RNA |
| **lincRNA** | long intergenic noncoding RNAs |
| **caRNAs** | chromatin associated RNAs |
| **HTS** | high throughput sequencing |
| **snoRNA** | small nucleolar RNAs |
| **tRNAs** | transfer RNAs |
| **miRNAs** | micro RNAs |
| **piRNAs** | piwi interacting RNAs |
| **TEs** | transposable elements |
| **RIDLs** | Repeat Insertion Domains of LncRNAs |
| **SIRLOIN** | SINE-derived nuclear organization |
| **GRO-seq** | global run-on sequencing |
| **PRO-seq** | precision run-on sequencing |
| **PRO-cap** | precision run-on sequencing of capped RNA |

| | |
|---|---|
| **TSS** | transcription start site |
| **ChRO-seq** | chromatin run-on sequencing |
| **R-loop** | RNA moiety loop |
| **DRIP-seq** | DNA:RNA immunoprecipitation |
| **S1DRIP** | S1-nuclease DRIP-seq |
| **bis-DRIP** | bisulfide DNA:RNA immunopreciptation |
| **RDIP** | RNA:DNA immunoprecipitation |
| **DRIVE** | DNA:RNA in vitro enrichment |
| **R-ChIP** | R-loop chromatin enrichment |
| **CUT&RUN** | cleavage under targets and release using nuclease |
| **RPA** | replication protein A |
| **ATR** | ataxia telangiectasia mutated and Rad3-related |
| **ChIRP** | chromatin isolation by RNA purification |
| **RAP** | RNA antisense purification |
| **CHART** | capture hybridization analysis of RNA targets |
| **MARGI** | mapping RNA genome interactions |
| **GRID-seq** | global RNA interactions with DNA by deep sequencing |
| **ChAR-seq** | chromatin associate RNA sequencing |
| **cheRNA-seq** | chromatin enriched RNA-seq |
| **RADICL-seq** | RNA and DNA interacting complexes ligated and sequenced |

## Bibliography

Altemose N, Miga KH, Maggioni M, Willard HF (2014) Genomic characterization of large heterochromatic gaps in the human genome assembly PLoS Comput Biol 10:e1003628 doi:10.1371/journal.pcbi.1003628 [PubMed: 24831296]

Andersson R et al. (2014) An atlas of active enhancers across human cell types and tissues Nature 507:455–461 doi:10.1038/nature12787 [PubMed: 24670763]

Ayupe AC, Tahira AC, Camargo L, Beckedorff FC, Verjovski-Almeida S, Reis EM (2015) Global analysis of biogenesis, stability and sub-cellular localization of lncRNAs mapping to intragenic regions of the human genome RNA Biol 12:877–892 doi:10.1080/15476286.2015.1062960 [PubMed: 26151857]

Bacolla A, Wang G, Vasquez KM (2015) New Perspectives on DNA and RNA Triplexes As Effectors of Biological Activity PLoS Genet 11:e1005696 doi:10.1371/journal.pgen.1005696 [PubMed: 26700634]

Bell JC et al. (2018) Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts Elife 7 doi:10.7554/eLife.27024

Bhatt DM et al. (2012) Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions Cell 150:279–290 doi:10.1016/j.cell.2012.05.043 [PubMed: 22817891]

Bierhoff H, Schmitz K, Maass F, Ye J, Grummt I (2010) Noncoding transcripts in sense and antisense orientation regulate the epigenetic state of ribosomal RNA genes Cold Spring Harb Symp Quant Biol 75:357–364 doi:10.1101/sqb.2010.75.060 [PubMed: 21502405]

Biscotti MA, Canapa A, Forconi M, Olmo E, Barucca M (2015) Transcription of tandemly repetitive DNA: functional roles Chromosome Res 23:463–477 doi:10.1007/s10577-015-9494-4 [PubMed: 26403245]

Boguslawski SJ, Smith DE, Michalak MA, Mickelson KE, Yehle CO, Patterson WL, Carrico RJ (1986) Characterization of monoclonal antibody to DNA.RNA and its application to immunodetection of hybrids J Immunol Methods 89:123–130 doi:10.1016/0022-1759(86)90040-2 [PubMed: 2422282]

Bonetti A et al. (2019) RADICL-seq identifies general and cell type-specific principles of genome-wide RNA-chromatin interactions bioRxiv:681924 doi:10.1101/681924

Bongartz P (2019) Resolving repeat families with long reads BMC Bioinformatics 20:232 doi:10.1186/s12859-019-2807-4 [PubMed: 31072311]

Brannan CI, Dees EC, Ingram RS, Tilghman SM (1990) The product of the H19 gene may function as an RNA Mol Cell Biol 10:28–36 doi:10.1128/mcb.10.1.28 [PubMed: 1688465]

Brown CJ, Hendrich BD, Rupert JL, Lafreniere 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 doi:10.1016/0092-8674(92)90520-m [PubMed: 1423611]

Brown JD, Mitchell SE, O'Neill RJ (2012) Making a long story short: noncoding RNAs and chromosome change Heredity (Edinb) 108:42–49 doi:10.1038/hdy.2011.104 [PubMed: 22072070]

Brown JD, O'Neill RJ (2010) Chromosomes, conflict, and epigenetics: chromosomal speciation revisited Annu Rev Genomics Hum Genet 11:291–316 doi:10.1146/annurev-genom-082509-141554 [PubMed: 20438362]

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions Nat Biotechnol 31:1119–1125 doi:10.1038/nbt.2727 [PubMed: 24185095]

Buske FA, Bauer DC, Mattick JS, Bailey TL (2012) Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data Genome Res 22:1372–1381 doi:10.1101/gr.130237.111 [PubMed: 22550012]

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses Genes Dev 25:1915–1927 doi:10.1101/gad.17446611 [PubMed: 21890647]

Carone DM et al. (2009) A new class of retroviral and satellite encoded small RNAs emanates from mammalian centromeres Chromosoma 118:113–125 doi:10.1007/s00412-008-0181-5 [PubMed: 18839199]

Carone DM, Zhang C, Hall LE, Obergfell C, Carone BR, O'Neill MJ, O'Neill RJ (2013) Hypermorphic expression of centromeric retroelement-encoded small RNAs impairs CENP-A loading Chromosome Res 21:49–62 doi:10.1007/s10577-013-9337-0 [PubMed: 23392618]

Castellano-Pozo M et al. (2013) R loops are linked to histone H3 S10 phosphorylation and chromatin condensation Mol Cell 52:583–590 doi:10.1016/j.molcel.2013.10.006 [PubMed: 24211264]

Caudron-Herger M, Rippe K (2012) Nuclear architecture by RNA Curr Opin Genet Dev 22:179–187 doi:10.1016/j.gde.2011.12.005 [PubMed: 22281031]

Chakraborty P, Grosse F (2011) Human DHX9 helicase preferentially unwinds RNA-containing displacement loops (R-loops) and G-quadruplexes DNA Repair (Amst) 10:654–665 doi:10.1016/j.dnarep.2011.04.013 [PubMed: 21561811]

Chang CH et al. (2019) Islands of retroelements are major components of Drosophila centromeres PLoS Biol 17:e3000241 doi:10.1371/journal.pbio.3000241 [PubMed: 31086362]

Chen L et al. (2017) R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters Mol Cell 68:745–757 e745 doi:10.1016/j.molcel.2017.10.008 [PubMed: 29104020]

Chu C, Nielsen R, Wu Y (2016) REPdenovo: Inferring De Novo Repeat Motifs from Short Sequence Reads PLoS One 11:e0150719 doi:10.1371/journal.pone.0150719 [PubMed: 26977803]

Chu C, Qu K, Zhong FL, Artandi SE, Chang HY (2011) Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions Mol Cell 44:667–678 doi:10.1016/j.molcel.2011.08.027 [PubMed: 21963238]

Chu T et al. (2018) Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme Nat Genet 50:1553–1564 doi:10.1038/s41588-018-0244-3 [PubMed: 30349114]

Churchman LS, Weissman JS (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution Nature 469:368–373 doi:10.1038/nature09652 [PubMed: 21248844]

Cinghu S et al. (2017) Intragenic Enhancers Attenuate Host Gene Expression Mol Cell 68:104–117 e106 doi:10.1016/j.molcel.2017.09.010 [PubMed: 28985501]

Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters Science 322:1845–1848 doi:10.1126/science.1162228 [PubMed: 19056941]

Doolittle WF (2018) We simply cannot go on being so vague about 'function' Genome Biol 19:223 doi:10.1186/s13059-018-1600-4 [PubMed: 30563541]

Drolet M, Phoenix P, Menzel R, Masse E, Liu LF, Crouch RJ (1995) Overexpression of RNase H partially complements the growth defect of an Escherichia coli delta topA mutant: R-loop formation is a major problem in the absence of DNA topoisomerase I Proc Natl Acad Sci U S A 92:3526–3530 doi:10.1073/pnas.92.8.3526 [PubMed: 7536935]

Dumelie JG, Jaffrey SR (2017) Defining the location of promoter-associated R-loops at near-nucleotide resolution using bisDRIP-seq Elife 6 doi:10.7554/eLife.28306

Dupuis-Sandoval F, Poirier M, Scott MS (2015) The emerging landscape of small nucleolar RNAs in cell biology Wiley Interdiscip Rev RNA 6:381–397 doi:10.1002/wrna.1284 [PubMed: 25879954]

El Achkar E, Gerbault-Seureau M, Muleris M, Dutrillaux B, Debatisse M (2005) Premature condensation induces breaks at the interface of early and late replicating chromosome bands bearing common fragile sites Proc Natl Acad Sci U S A 102:18069–18074 doi:10.1073/pnas.0506497102 [PubMed: 16330769]

El Hage A, French SL, Beyer AL, Tollervey D (2010) Loss of Topoisomerase I leads to R-loop-mediated transcriptional blocks during ribosomal RNA synthesis Genes Dev 24:1546–1558 doi:10.1101/gad.573310 [PubMed: 20634320]

Engreitz JM et al. (2013) The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome Science 341:1237973 doi:10.1126/science.1237973 [PubMed: 23828888]

Felsenfeld G, Rich A (1957) Studies on the formation of two- and three-stranded polyribonucleotides Biochim Biophys Acta 26:457–468 doi:10.1016/0006-3002(57)90091-4 [PubMed: 13499402]

Fey EG, Krochmalnic G, Penman S (1986a) The nonchromatin substructures of the nucleus: the ribonucleoprotein (RNP)-containing and RNP-depleted matrices analyzed by sequential fractionation and resinless section electron microscopy J Cell Biol 102:1654–1665 doi:10.1083/jcb.102.5.1654 [PubMed: 3700470]

Fey EG, Ornelles DA, Penman S (1986b) Association of RNA with the cytoskeleton and the nuclear matrix J Cell Sci Suppl 5:99–119 doi:10.1242/jcs.1986.supplement_5.6 [PubMed: 3477558]

Ginno PA, Lim YW, Lott PL, Korf I, Chedin F (2013) GC skew at the 5' and 3' ends of human genes links R-loop formation to epigenetic regulation and transcription termination Genome Res 23:1590–1600 doi:10.1101/gr.158436.113 [PubMed: 23868195]

Ginno PA, Lott PL, Christensen HC, Korf I, Chedin F (2012) R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters Mol Cell 45:814–825 doi:10.1016/j.molcel.2012.01.017 [PubMed: 22387027]

Goni JR, de la Cruz X, Orozco M (2004) Triplex-forming oligonucleotide target sequences in the human genome Nucleic Acids Res 32:354–360 doi:10.1093/nar/gkh188 [PubMed: 14726484]

Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E (2013) On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE Genome Biol Evol 5:578–590 doi:10.1093/gbe/evt028 [PubMed: 23431001]

Groh M, Lufino MM, Wade-Martins R, Gromak N (2014) R-loops associated with triplet repeat expansions promote gene silencing in Friedreich ataxia and fragile X syndrome PLoS Genet 10:e1004318 doi:10.1371/journal.pgen.1004318 [PubMed: 24787137]

Grote P, Herrmann BG (2013) The long non-coding RNA Fendrr links epigenetic control mechanisms to gene regulatory networks in mammalian embryogenesis RNA Biol 10:1579–1585 doi:10.4161/rna.26165 [PubMed: 24036695]

Guttman M et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals Nature 458:223–227 doi:10.1038/nature07672 [PubMed: 19182780]

Guttman M, Rinn JL (2012) Modular regulatory principles of large non-coding RNAs Nature 482:339–346 doi:10.1038/nature10887 [PubMed: 22337053]

Hall LL, Lawrence JB (2016) RNA as a fundamental component of interphase chromosomes: could repeats prove key? Curr Opin Genet Dev 37:137–147 doi:10.1016/j.gde.2016.04.005 [PubMed: 27218204]

Hartley G, O'Neill RJ (2019) Centromere Repeats: Hidden Gems of the Genome Genes (Basel) 10 doi:10.3390/genes10030223

Hartono SR, Malapert A, Legros P, Bernard P, Chedin F, Vanoosthuyse V (2018) The Affinity of the S9.6 Antibody for Double-Stranded RNAs Impacts the Accurate Mapping of R-Loops in Fission Yeast J Mol Biol 430:272–284 doi:10.1016/j.jmb.2017.12.016 [PubMed: 29289567]

He X, Cicek AE, Wang Y, Schulz MH, Le HS, Bar-Joseph Z (2015) De novo ChIP-seq analysis Genome Biol 16:205 doi:10.1186/s13059-015-0756-4 [PubMed: 26400819]

Holmes DS, Mayfield JE, Sander G, Bonner J (1972) Chromosomal RNA: its properties Science 177:72–74 doi:10.1126/science.177.4043.72 [PubMed: 5041779]

Jalali S, Singh A, Maiti S, Scaria V (2017) Genome-wide computational analysis of potential long noncoding RNA mediated DNA:DNA:RNA triplexes in the human genome J Transl Med 15:186 doi:10.1186/s12967-017-1282-9 [PubMed: 28865451]

Johnson R, Guigo R (2014) The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs RNA 20:959–976 doi:10.1261/rna.044560.114 [PubMed: 24850885]

Johnson RN et al. (2018) Adaptation and conservation insights from the koala genome Nat Genet 50:1102–1111 doi:10.1038/s41588-018-0153-5 [PubMed: 29967444]

Johnson WE (2019) Origins and evolutionary consequences of ancient endogenous retroviruses Nat Rev Microbiol 17:355–370 doi:10.1038/s41579-019-0189-2 [PubMed: 30962577]

Jurka J (2000) Repbase update: a database and an electronic journal of repetitive elements Trends Genet 16:418–420 doi:10.1016/s0168-9525(00)02093-x [PubMed: 10973072]

Kabeche L, Nguyen HD, Buisson R, Zou L (2018) A mitosis-specific and R loop-driven ATR pathway promotes faithful chromosome segregation Science 359:108–114 doi:10.1126/science.aan6490 [PubMed: 29170278]

Kaplan N, Dekker J (2013) High-throughput genome scaffolding from in vivo DNA interaction frequency Nat Biotechnol 31:1143–1147 doi:10.1038/nbt.2768 [PubMed: 24270850]

Kapusta A et al. (2013) Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs PLoS Genet 9:e1003470 doi:10.1371/journal.pgen.1003470 [PubMed: 23637635]

Keeling DM, Garza P, Nartey CM, Carvunis AR (2019) The meanings of 'function' in biology and the problematic case of de novo gene emergence Elife 8 doi:10.7554/eLife.47014

Kim VN, Han J, Siomi MC (2009) Biogenesis of small RNAs in animals Nat Rev Mol Cell Biol 10:126–139 doi:10.1038/nrm2632 [PubMed: 19165215]

Koch P, Platzer M, Downie BR (2014) RepARK--de novo creation of repeat libraries from whole-genome NGS reads Nucleic Acids Res 42:e80 doi:10.1093/nar/gku210 [PubMed: 24634442]

Kwak H, Fuda NJ, Core LJ, Lis JT (2013) Precise maps of RNA polymerase reveal how promoters direct initiation and pausing Science 339:950–953 doi:10.1126/science.1229386 [PubMed: 23430654]

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2 Nat Methods 9:357–359 doi:10.1038/nmeth.1923 [PubMed: 22388286]

Larson MH et al. (2014) A pause sequence enriched at translation start sites drives transcription dynamics in vivo Science 344:1042–1047 doi:10.1126/science.1251871 [PubMed: 24789973]

Laubichler MD, Stadler PF, Prohaska SJ, Nowick K (2015) The relativity of biological function Theory Biosci 134:143–147 doi:10.1007/s12064-015-0215-5 [PubMed: 26449352]

Lefebvre A, Lecroq T, Dauchel H, Alexandre J (2003) FORRepeats: detects repeats on entire chromosomes and between genomes Bioinformatics 19:319–326 doi:10.1093/bioinformatics/btf843 [PubMed: 12584116]

Lewin HA, Graves JAM, Ryder OA, Graphodatsky AS, O'Brien SJ (2019) Precision nomenclature for the new genomics Gigascience 8 doi:10.1093/gigascience/giz086

Lewin HA et al. (2018) Earth BioGenome Project: Sequencing life for the future of life Proc Natl Acad Sci U S A 115:4325–4333 doi:10.1073/pnas.1720115115 [PubMed: 29686065]

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM arXiv:13033997v2

Li X, Zhou B, Chen L, Gou LT, Li H, Fu XD (2017) GRID-seq reveals the global RNA-chromatin interactome Nat Biotechnol 35:940–950 doi:10.1038/nbt.3968 [PubMed: 28922346]

Li Y, Syed J, Sugiyama H (2016) RNA-DNA Triplex Formation by Long Noncoding RNAs Cell Chem Biol 23:1325–1333 doi:10.1016/j.chembiol.2016.09.011 [PubMed: 27773629]

Lieberman-Aiden E et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome Science 326:289–293 doi:10.1126/science.1181369 [PubMed: 19815776]

Lubelsky Y, Ulitsky I (2018) Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells Nature 555:107–111 doi:10.1038/nature25757 [PubMed: 29466324]

Mahat DB et al. (2016) Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq) Nat Protoc 11:1455–1476 doi:10.1038/nprot.2016.086 [PubMed: 27442863]

Marcais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers Bioinformatics 27:764–770 doi:10.1093/bioinformatics/btr011 [PubMed: 21217122]

Marie-Nelly H et al. (2014) High-quality genome (re)assembly using chromosomal contact data Nat Commun 5:5695 doi:10.1038/ncomms6695 [PubMed: 25517223]

Marques AC, Ponting CP (2009) Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness Genome Biol 10:R124 doi:10.1186/gb-2009-10-11-r124 [PubMed: 19895688]

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 109:1963–1973 doi:10.1083/jcb.109.5.1963 [PubMed: 2808515]

Matsuno Y, Yamashita T, Wagatsuma M, Yamakage H (2019) Convergence in LINE-1 nucleotide variations can benefit redundantly forming triplexes with lncRNA in mammalian X-chromosome inactivation Mob DNA 10:33 doi:10.1186/s13100-019-0173-4 [PubMed: 31384315]

Mattick JS (2001) Non-coding RNAs: the architects of eukaryotic complexity EMBO Rep 2:986–991 doi:10.1093/embo-reports/kve230 [PubMed: 11713189]

Mattick JS (2005) The functional genomics of noncoding RNA Science 309:1527–1528 doi:10.1126/science.1117806 [PubMed: 16141063]

Mattick JS (2009) The genetic signatures of noncoding RNAs PLoS Genet 5:e1000459 doi:10.1371/journal.pgen.1000459 [PubMed: 19390609]

May BP, Lippman ZB, Fang Y, Spector DL, Martienssen RA (2005) Differential regulation of strand-specific transcripts from Arabidopsis centromeric satellite repeats PLoS Genet 1:e79 doi:10.1371/journal.pgen.0010079 [PubMed: 16389298]

McNulty SM, Sullivan LL, Sullivan BA (2017) Human Centromeres Produce Chromosome-Specific and Array-Specific Alpha Satellite Transcripts that Are Complexed with CENP-A and CENP-C Dev Cell 42:226–240 e226 doi:10.1016/j.devcel.2017.07.001

Mele M, Rinn JL (2016) "Cat's Cradling" the 3D Genome by the Act of LncRNA Transcription Mol Cell 62:657–664 doi:10.1016/j.molcel.2016.05.011 [PubMed: 27259198]

Michieletto D, Gilbert N (2019) Role of nuclear RNA in regulating chromatin structure and transcription Curr Opin Cell Biol 58:120–125 doi:10.1016/j.ceb.2019.03.007 [PubMed: 31009871]

Miga KH (2015) Completing the human genome: the progress and challenge of satellite DNA assembly Chromosome Res 23:421–426 doi:10.1007/s10577-015-9488-2 [PubMed: 26363799]

Miga KH et al. (2019) Telomere-to-telomere assembly of a complete human X chromosome bioRxiv:735928 doi:10.1101/735928

Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ (2014) Centromere reference models for human chromosomes X and Y satellite arrays Genome Res 24:697–707 doi:10.1101/gr.159624.113 [PubMed: 24501022]

Minocherhomji S et al. (2015) Replication stress activates DNA repair synthesis in mitosis Nature 528:286–290 doi:10.1038/nature16139 [PubMed: 26633632]

Nadel J et al. (2015) RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships Epigenetics Chromatin 8:46 doi:10.1186/s13072-015-0040-6 [PubMed: 26579211]

Necsulea A et al. (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods Nature 505:635–640 doi:10.1038/nature12943 [PubMed: 24463510]

Nojima T et al. (2015) Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing Cell 161:526–540 doi:10.1016/j.cell.2015.03.027 [PubMed: 25910207]

Nozawa RS, Gilbert N (2019) RNA: Nuclear Glue for Folding the Genome Trends Cell Biol 29:201–211 doi:10.1016/j.tcb.2018.12.003

O'Neill RJ, Carone DM (2009) The role of ncRNA in centromeres: a lesson from marsupials Prog Mol Subcell Biol 48:77–101 doi:10.1007/978-3-642-00182-6_4

Oberbauer V, Schaefer MR (2018) tRNA-Derived Small RNAs: Biogenesis, Modification, Function and Potential Impact on Human Disease Development Genes (Basel) 9 doi:10.3390/genes9120607

Ozata DM, Gainetdinov I, Zoch A, O'Carroll D, Zamore PD (2019) PIWI-interacting RNAs: small RNAs with big functions Nat Rev Genet 20:89–108 doi:10.1038/s41576-018-0073-3 [PubMed: 30446728]

Palazzo AF, Lee ES (2015) Non-coding RNA: what is functional and what is junk? Front Genet 6:2 doi:10.3389/fgene.2015.00002 [PubMed: 25674102]

Pan T (2018) Modifications and functional genomics of human transfer RNA Cell Res 28:395–404 doi:10.1038/s41422-018-0013-y [PubMed: 29463900]

Pederson T (2000) Half a century of "the nuclear matrix" Mol Biol Cell 11:799–805 doi:10.1091/mbc.11.3.799 [PubMed: 10712500]

Pennisi E (2012) Genomics. ENCODE project writes eulogy for junk DNA Science 337:1159, 1161 doi:10.1126/science.337.6099.1159 [PubMed: 22955811]

Qian X, Zhao J, Yeung PY, Zhang QC, Kwok CK (2019) Revealing lncRNA Structures and Interactions by Sequencing-Based Approaches Trends Biochem Sci 44:33–52 doi:10.1016/j.tibs.2018.09.012 [PubMed: 30459069]

Rinn JL et al. (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs Cell 129:1311–1323 doi:10.1016/j.cell.2007.05.022 [PubMed: 17604720]

Rosenbloom KR et al. (2015) The UCSC Genome Browser database: 2015 update Nucleic Acids Res 43:D670–681 doi:10.1093/nar/gku1177

Rosic S, Kohler F, Erhardt S (2014) Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division J Cell Biol 207:335–349 doi:10.1083/jcb.201404097 [PubMed: 25365994]

Roy D, Yu K, Lieber MR (2008) Mechanism of R-loop formation at immunoglobulin class switch sequences Mol Cell Biol 28:50–60 doi:10.1128/MCB.01251-07 [PubMed: 17954560]

Russo M, De Lucca B, Flati T, Gioiosa S, Chillemi G, Capranico G (2019) DROPA: DRIP-seq optimized peak annotator BMC Bioinformatics 20:414 doi:10.1186/s12859-019-3009-9 [PubMed: 31387525]

Santos-Pereira JM, Aguilera A (2015) R loops: new modulators of genome dynamics and function Nat Rev Genet 16:583–597 doi:10.1038/nrg3961 [PubMed: 26370899]

Schmitz KM, Mayer C, Postepska A, Grummt I (2010) Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes Genes Dev 24:2264–2269 doi:10.1101/gad.590910 [PubMed: 20952535]

Senturk Cetin N, Kuo CC, Ribarska T, Li R, Costa IG, Grummt I (2019) Isolation and genome-wide characterization of cellular DNA:RNA triplex structures Nucleic Acids Res 47:2306–2321 doi:10.1093/nar/gky1305 [PubMed: 30605520]

Shafin K et al. (2019) Efficient de novo assembly of eleven human genomes using PromethION sequencing and a novel nanopore toolkit bioRxiv:715722 doi:10.1101/715722

Simon MD (2016) Insight into lncRNA biology using hybridization capture analyses Biochim Biophys Acta 1859:121–127 doi:10.1016/j.bbagrm.2015.09.004

Simon MD et al. (2011) The genomic binding sites of a noncoding RNA Proc Natl Acad Sci U S A 108:20497–20502 doi:10.1073/pnas.1113536108 [PubMed: 22143764]

Skene PJ, Henikoff S (2017) An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites Elife 6 doi:10.7554/eLife.21856

Skourti-Stathaki K, Kamieniarz-Gdula K, Proudfoot NJ (2014) R-loops induce repressive chromatin marks over mammalian gene terminators Nature 516:436–439 doi:10.1038/nature13787 [PubMed: 25296254]

Skourti-Stathaki K, Proudfoot NJ, Gromak N (2011) Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination Mol Cell 42:794–805 doi:10.1016/j.molcel.2011.04.026 [PubMed: 21700224]

Smale ST (2009) Nuclear run-on assay Cold Spring Harb Protoc 2009:pdb prot5329 doi:10.1101/pdb.prot5329

Smalec BM, Heider TN, Flynn BL, O'Neill RJ (2019) A centromere satellite concomitant with extensive karyotypic diversity across the Peromyscus genus defies predictions of molecular drive Chromosome Res 27:237–252 doi:10.1007/s10577-019-09605-1 [PubMed: 30771198]

Smit AF, Hubley R, Green P (2015) RepeatMasker Open-4.0 <http://www.repeatmasker.org>.

Soibam B (2017) Super-lncRNAs: identification of lncRNAs that target super-enhancers via RNA:DNA:DNA triplex formation RNA 23:1729–1742 doi:10.1261/rna.061317.117 [PubMed: 28839111]

Sridhar B et al. (2017) Systematic Mapping of RNA-Chromatin Interactions In Vivo Curr Biol 27:610–612 doi:10.1016/j.cub.2017.01.068

Talbert PB, Henikoff S (2018) Transcribing Centromeres: Noncoding RNAs and Kinetochore Assembly Trends Genet 34:587–599 doi:10.1016/j.tig.2018.05.001 [PubMed: 29871772]

Thomas M, White RL, Davis RW (1976) Hybridization of RNA to double-stranded DNA: formation of R-loops Proc Natl Acad Sci U S A 73:2294–2298 doi:10.1073/pnas.73.7.2294 [PubMed: 781674]

Topp CN, Zhong CX, Dawe RK (2004) Centromere-encoded RNAs are integral components of the maize kinetochore Proc Natl Acad Sci U S A 101:15986–15991 doi:10.1073/pnas.0407154101 [PubMed: 15514020]

Treiber T, Treiber N, Meister G (2019) Regulation of microRNA biogenesis and its crosstalk with other cellular pathways Nat Rev Mol Cell Biol 20:5–20 doi:10.1038/s41580-018-0059-1 [PubMed: 30228348]

Tripathi V et al. (2010) The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation Mol Cell 39:925–938 doi:10.1016/j.molcel.2010.08.011 [PubMed: 20797886]

Ugarkovic D (2005) Functional elements residing within satellite DNAs EMBO Rep 6:1035–1039 doi:10.1038/sj.embor.7400558 [PubMed: 16264428]

Wahba L, Costantino L, Tan FJ, Zimmer A, Koshland D (2016) S1-DRIP-seq identifies high expression and polyA tracts as major contributors to R-loop formation Genes Dev 30:1327–1338 doi:10.1101/gad.280834.116 [PubMed: 27298336]

Weber CM, Ramachandran S, Henikoff S (2014) Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase Mol Cell 53:819–830 doi:10.1016/j.molcel.2014.02.014 [PubMed: 24606920]

Werner MS, Ruthenburg AJ (2015) Nuclear Fractionation Reveals Thousands of Chromatin-Tethered Noncoding RNAs Adjacent to Active Genes Cell Rep 12:1089–1098 doi:10.1016/j.celrep.2015.07.033 [PubMed: 26257179]

West JA et al. (2014) The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites Mol Cell 55:791–802 doi:10.1016/j.molcel.2014.07.012 [PubMed: 25155612]

Westover KD, Bushnell DA, Kornberg RD (2004) Structural basis of transcription: nucleotide selection by rotation in the RNA polymerase II active center Cell 119:481–489 doi:10.1016/j.cell.2004.10.016 [PubMed: 15537538]

Wong LH et al. (2007) Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere Genome Res 17:1146–1160 doi:10.1101/gr.6022807 [PubMed: 17623812]

Wu Q, Gaddis SS, MacLeod MC, Walborg EF, Thames HD, DiGiovanni J, Vasquez KM (2007) High-affinity triplex-forming oligonucleotide target sequences in mammalian genomes Mol Carcinog 46:15–23 doi:10.1002/mc.20261 [PubMed: 17013831]

Wuarin J, Schibler U (1994) Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing Mol Cell Biol 14:7219–7225 doi:10.1128/mcb.14.11.7219 [PubMed: 7523861]

Yan Q, Shields EJ, Bonasio R, Sarma K (2019) Mapping Native R-Loops Genome-wide Using a Targeted Nuclease Approach Cell Rep 29:1369–1380 e1365 doi:10.1016/j.celrep.2019.09.052 [PubMed: 31665646]

Yu K, Chedin F, Hsieh CL, Wilson TE, Lieber MR (2003) R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells Nat Immunol 4:442–451 doi:10.1038/ni919 [PubMed: 12679812]

Zhou R, Zhang J, Bochman ML, Zakian VA, Ha T (2014) Periodic DNA patrolling underlies diverse functions of Pif1 on R-loops and G-rich DNA Elife 3:e02190 doi:10.7554/eLife.02190 [PubMed: 24843019]

Zou L, Elledge SJ (2003) Sensing DNA damage through ATRIP recognition of RPA-ssDNA complexes Science 300:1542–1548 doi:10.1126/science.1083430 [PubMed: 12791985]

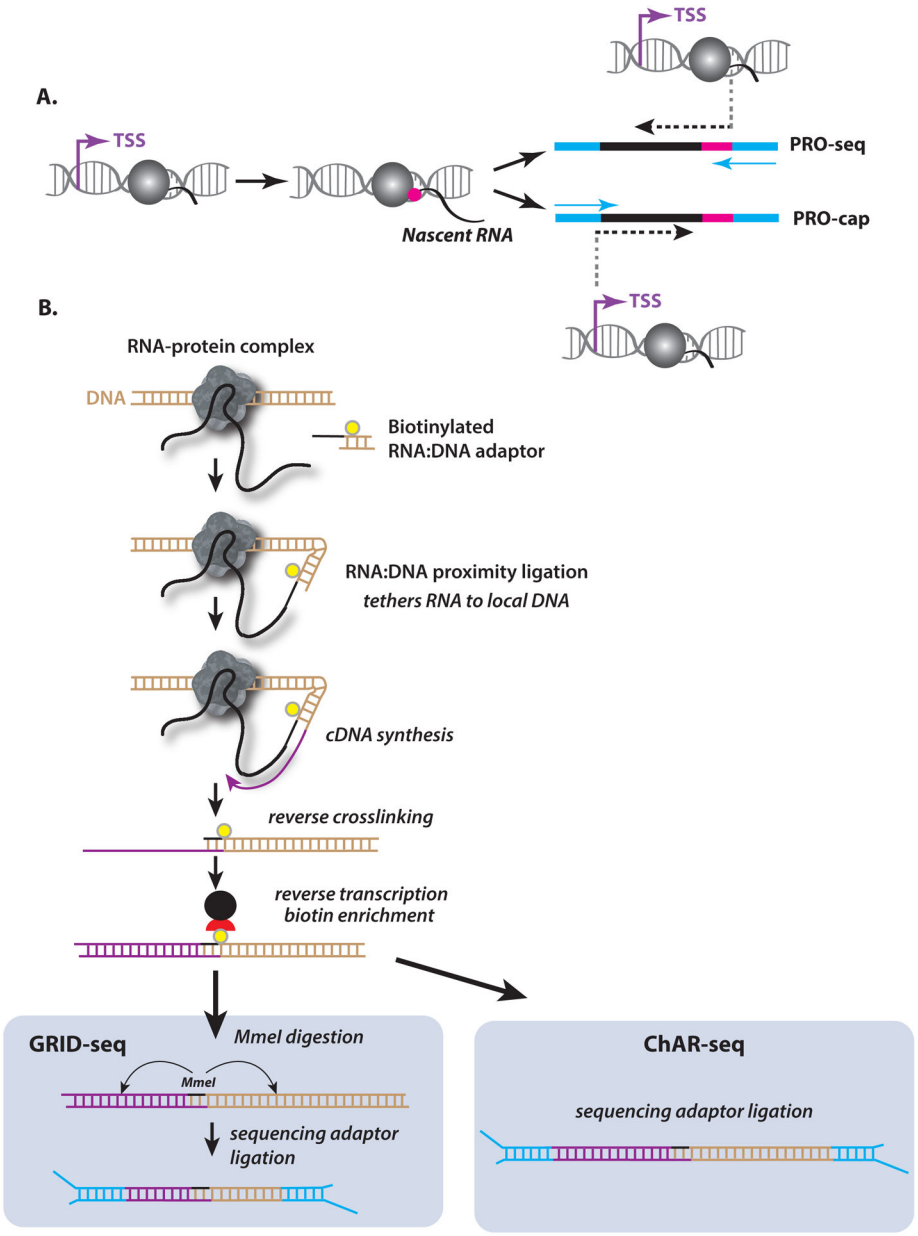


**Figure 1.**

**A.** Using a genome-wide nuclear run-on reaction incorporating a biotin-labelled ribonucleotide (pink) followed by sequencing adaptor (blue) ligation, PRO-seq (top) is used to capture sites of active RNA polymerase engagement and PRO-cap is used to identify transcription start sites (TSS). **B.** Both GRID-seq and ChAR-seq start by cross linking RNA-protein-DNA complexes and proximity ligation of an RNA-DNA hybrid adaptor that is biotinylated (yellow). cDNA synthesis (purple) proceeds from the adaptor, resulting in sequences containing cDNA (purple), the biotinylated adaptor (black and yellow), and presumed interacting DNA sequence (tan). After reversal of crosslinks, proximity-ligation products are enriched using streptavidin-coated magnetic beads. GRID-seq (left) proceeds with *MmeI* digestion based on the *MmeI* recognition sequence within the adaptor. Following

cleavage, which occurs ~20bp from the hybrid adaptor, sequencing adaptors are ligated (blue) for subsequent HTS. ChAR-seq (right) does not rely on *MmeI* digestion, allowing for capturing more sequence information following sequencing adaptor ligation (blue) and HTS.

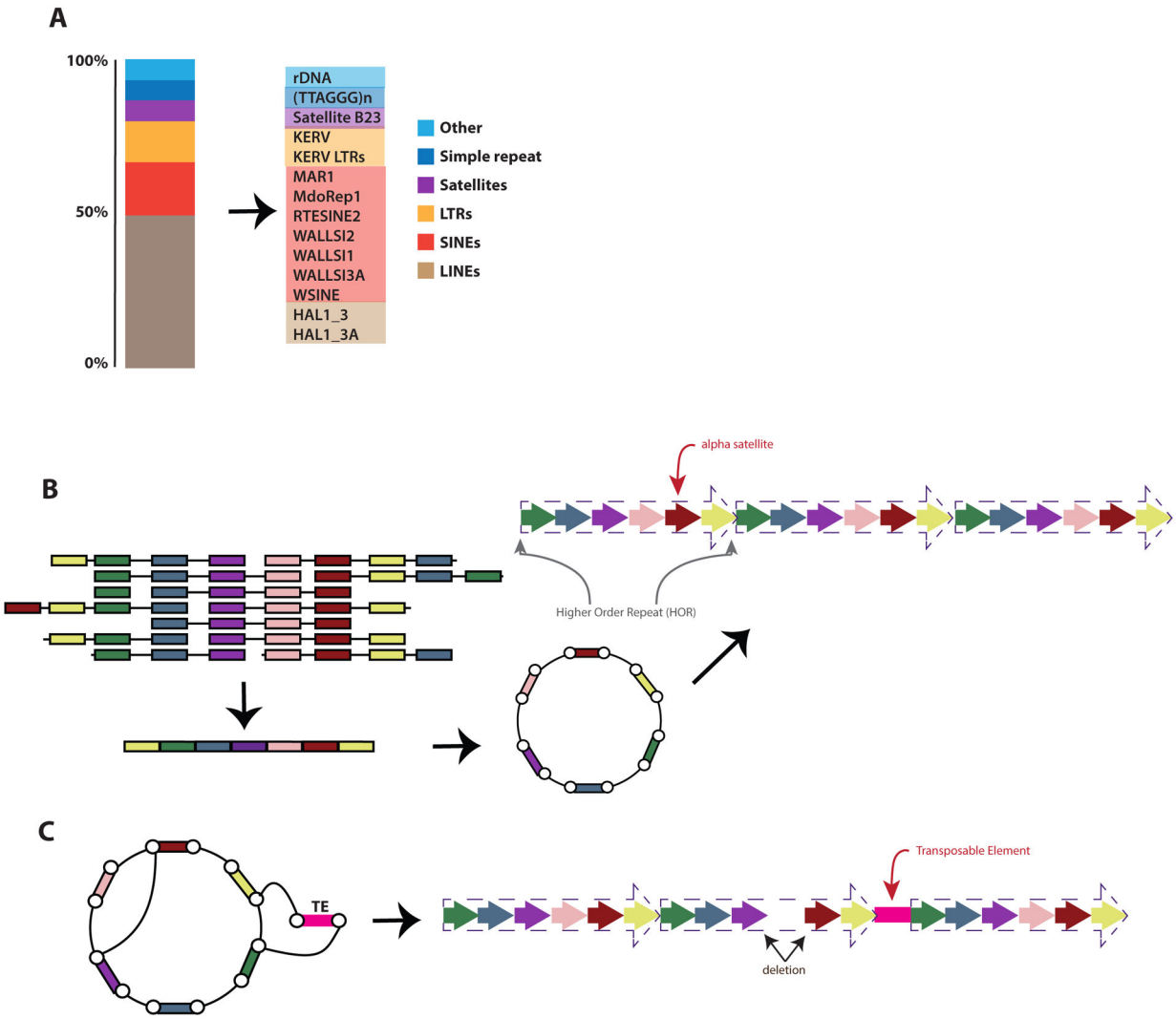


**Figure 2.**

Examples of methods used to study repeats in the absence of a genome assembly. **A.** Repeat Masker applied to raw sequencing data provides details on overall frequency of repeats by class (left) and specific type (right). **B.** The linear order of highly repeated sequences, such as human alpha satellites found in centromeres, can be inferred from whole genome shotgun data (paired end sequencing). The resulting graphical model illustrates the frequency and order of satellite sequences (colored blocks). From the circular model, a linear arrangement of centromere satellites (colored arrows) can be inferred, including higher order repeat arrays (dotted arrows). **C.** Variations within centromere arrays, such as deletions and insertions, can be captured with the graphical model approach.

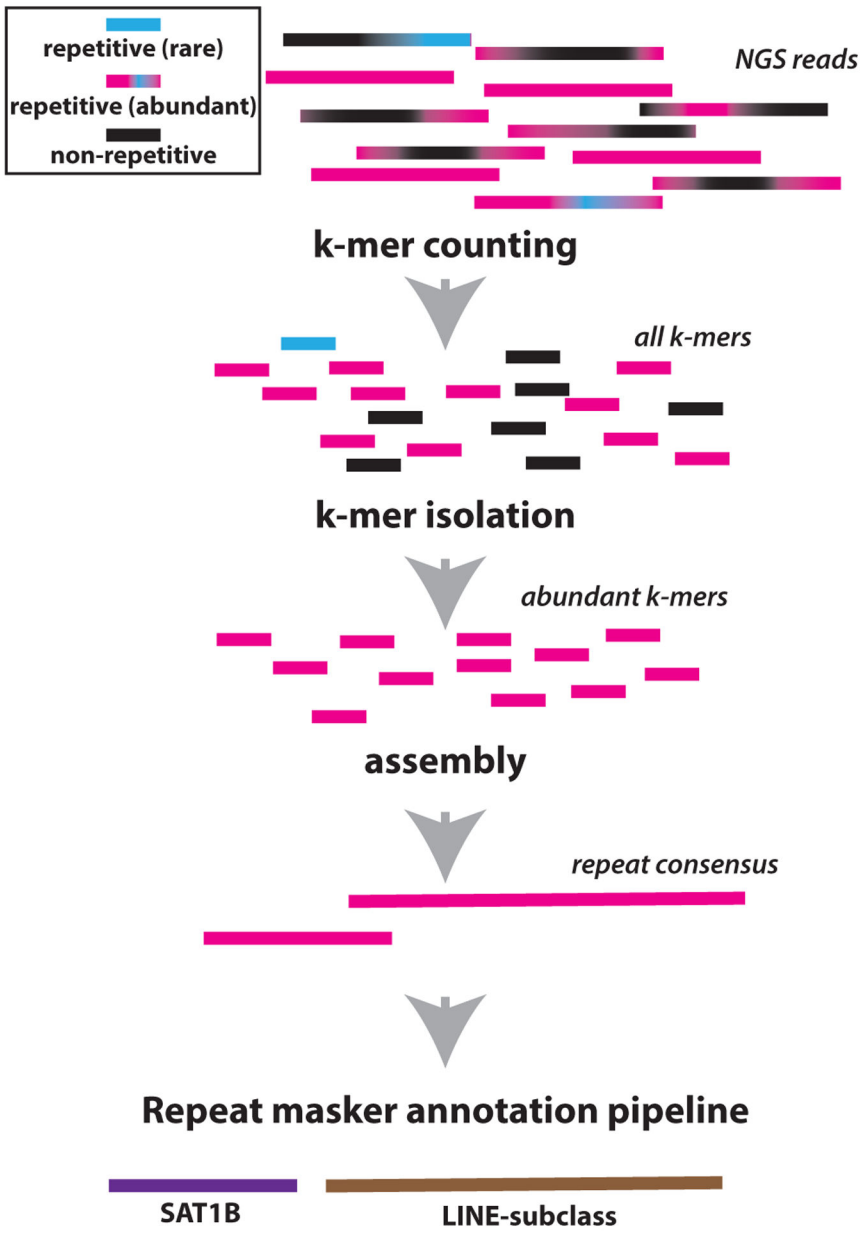


**Figure 3.**
RepARK uses k-mers to build a *de novo* assembly of repeats that can be further annotated for specific repeat type.

**Table 1.**

The Pittsburgh Model of Function as it relates to describing the function of any given ncRNA. The functional classification beings with the defined occurrence of a ncRNA (*expression)* and sequentially increases in the level of the classification based on the type of functional information garnered from studying the ncRNA in its biological system.

| Classification/Meaning | Definition |
|---|---|
| Vague | Insufficient evidence to infer one or more meanings of function within this model, nor to derive a new meaning |
| Expression | The presence or amount of ncRNA transcript |
| Capacities | Intrinsic physical properties of ncRNA; the necessity of the object's behavior given an environment (eg., structural constraints) |
| Interactions | Physical contacts, direct or indirect, between the ncRNA and the other components of a system |
| Physiological Implications | The ncRNA's involvement in biological processes as enabled by a set of its capacities, interactions and expression patterns, independent of cross-generational considerations. |
| Evolutionary Implications | The ncRNA's influence on population dynamics over successive generations, as enabled by its physiological implications and their interplay with environmental pressures. |