16 RNA Rules

RNA IS A CORE COMPONENT OF CHROMATIN

DNA and proteins have been the focus of the study of chromosome structure, but RNA is also a major component, essential to the organization of chromatin and the ‘nuclear matrix’.[1–18] As long ago as 1989, Sheldon Penman and colleagues demonstrated that transcription is required to maintain nuclear structure, and that chromatin integrity is destroyed by treatment with RNase, noting that “ribonucleoprotein granules were dispersed throughout the euchromatic regions” and suggesting “that RNA is a structural component of the nuclear matrix, which in turn may organize the higher order structure of chromatin”[4] (Chapter 4).

Genome-wide mapping and sequencing studies subsequently showed that there are many chromatin-bound RNAs in animal cells and that the locations of long non-coding RNAs in chromatin are “focal, sequence-specific and numerous”,[19] with thousands of “tightly associated” non-coding RNAs tethered adjacent to active genes.[11,20,21] Well-studied non-coding RNAs such as 7SK, U1, B2 and Alu RNAs, Gas5 and SRA, and more recently a large coterie of enhancer-derived and other lncRNAs, have been shown to be involved in the regulation of transcription initiation, elongation, termination and splicing.[22–25] The stress response induces the transcription downstream of protein-coding genes of thousands of lncRNAs that remain chromatin bound.[26] Chromatin-associated RNAs, which include those transcribed from enhancers and repeats, have been shown to have roles in genome organization via enhancer-promoter interactions and the formation of transcription hubs, heterochromatin and nuclear bodies (or ‘granules’)[11–13,15,16,21,27–30] through their interaction with proteins containing intrinsically disordered regions and the formation of phase-separated domains, as set out below.

REGULATION OF CHROMOSOME STRUCTURE

The scaffolding of euchromatin involves highly abundant (‘CoT1’) repeat RNAs, predominantly from 5' truncated LINE elements,[31,32] the expression of which varies during development and is regulated by other RNAs.[33,34] Chromatin-associated RNA proximity ligation reveals an RNA-DNA contact map similar to that observed by DNA-DNA ligation in topologically associated domains.[13] LncRNAs have been shown to regulate TAD formation,[35–37] and a recent analysis identified more than 10,000 RNA–chromatin interactions mediated by protein-coding RNAs and non-coding RNAs.[38] The RNAi machinery has also been shown to regulate nuclear topology.[39,40]

Many binding sites for CTCF, a zinc-finger containing protein (see below) that appears to anchor boundary sequences in TADs[41] (Chapter 14), are derived from transposable elements[42] and transcriptionally active HERV-H retrotransposons demarcate TADs in human pluripotent stem cells.[43] Similar to that observed with ‘enhancer’ RNAs (see below), lncRNAs have been reported to regulate neighboring genes through interaction with the Mediator complex,[44,45] a master coordinator of transcription and cell lineage commitment that also organizes chromosome topology (Chapter 14).

LINE- and centromere-derived repeat RNAs are structural and functional components of centromeric chromatin.[46–49] Heterochromatin formation generally requires the expression of repetitive sequences[50] and the RNAi pathway,[51–55] and RNA binding is required for heterochromatic localization of HP1 and the Suv39h histone methyltransferase.[56–59] Chromatin compaction is also controlled by lncRNAs that target IAP retrotransposons.[60] Telomere formation and maintenance requires specialized non-coding RNAs,[61,62] as does pairing of

DOI: 10.1201/9781003109242-16

homologous chromosomes in meiosis[63,64] and many, if not most, chromatin-associated proteins bind RNA,[65] including those involved in other chromatin-regulated process such as DNA stability and damage repair.[66]

RNA GUIDANCE OF CHROMATIN REMODELING

Chromatin structure is modulated during development by 'pioneer transcription factors' that alter cell fate in plants and animals by targeting nucleosomes and/or common DNA motifs.[67–71] The best known examples of reprogramming proteins are the 'Yamanaka' factors, Oct4 (*Pou5f1* gene), Sox2, Klf4 and c-Myc, which are (collectively) capable of converting differentiated cells to 'induced pluripotent stem cells' (iPSCs),[72–74] a process enhanced by inclusion of the RNA-binding protein, Lin28.[75] Oct4 is also involved, *inter alia*, in the differentiation of pluripotent cells to form the cranial neural crest.[76]

Another key pluripotency and reprogramming factor is Nanog, a homeobox-containing protein.[77,78] Homeoboxes are helix-loop-helix DNA-binding domains that exhibit a preference, but not specificity, for the common motif TAAT,[79,80] in the case of Nanog TAAT(G/T)(G/T).[81] Oct4 is also a homeobox-containing protein that recognizes the loose consensus sequence TTT(G/T)(G/C)(T/A)T(T/A), which occurs at thousands of sites around the genome.[82–84]

The expression of Oct4, Nanog and other pluripotency factors[a] is regulated by non-coding RNAs,[b] including pseudogene-derived lncRNAs,[93–98] one of which recruits the histone-lysine N-methyltransferase SUV39H1 to epigenetically silence Oct4 expression.[97,98] Reciprocally Oct4 and Nanog regulate the expression of lncRNAs that modulate pluripotency.[99] Oct4 and Nanog also have multiple pseudogenes,[100–103] some of which are differentially expressed in pluripotent and tumor cell lines.[102,104]

There are 16 classes of genes encoding homeobox proteins in animals, 11 in plants, with hundreds of orthologs in the human genome, most of which contain additional domains.[80] As noted already (Chapter 5), Hox proteins are 'master controllers' of gene expression patterns during animal and plant development, and regulate the expression of many genes at different developmental stages. While they recognize similar sequences *in vitro*, Hox proteins display wide functional diversity and identification of their *in vivo* genomic targets has proven elusive, as has the identification of the targets of Oct4 and Sox2.[80,83,105–109] Analysis of genome-wide DNase I hypersensitivity profiles and transcription factor (TF)-binding sites identified 120 and validated eight 'pioneer' TF families that dynamically open chromatin (including Sox2, Oct4 and Hoxa11), and identified 'settler' TFs (including c-Myc), and the nuclear hormone receptor RXR:RAR and NF-κB families, whose genomic binding is dependent on chromatin opening by pioneer TFs.[110]

The targets of Sox2, Oct4, Nanog and other Hox proteins change with developmental stage,[77,111,112] all of which suggests that other factors are involved in determining their locus specificity. In this context, it may not be an outlier observation that the *Drosophila* Hox protein Bicoid (which controls anterior-posterior patterning) binds RNA via its homeodomain,[113,114] nor that highly conserved lncRNAs are produced in vertebrate endoderm lineages from paralogous regions in *HOXA* and *HOXB* clusters.[115]

Sox2 is a member of a subclass of 'high mobility group' (HMG) proteins, the most abundant chromatin-associated proteins after histones. HMG proteins bend DNA structure, initiate chromatin opening and facilitate nucleosome remodeling.[116–118] There are three classes, one of which (HMG-A) is abundant in embryonic cells and binds AT-rich sequences, another (HMG-N) binds nucleosomes, and the third (HMG-B, which includes the Sox proteins) binds the DNA helix minor groove with no sequence specificity.[119,120] Sox2 influences development not only in pluripotent stem cells but also in the lung, ear and eye, and in neural lineages, but how it and other HMG-B proteins achieve their tissue-specific versatility is unclear.[107]

While Sox2 has low affinity for DNA,[116] it binds RNA with high affinity through its HMG domain,[121,122] as do other members of the HMG-B family,[123] "which requires a reassessment of how these proteins establish proper patterns of gene expression across the genome".[121] The HMG-B domain of the mammalian sex-determining protein Sry is homologous to the RNA-binding domain of a viral protein,[124] suggesting that their target selection *in vivo* is guided by trans-acting RNA signals. There are well-documented

[a] These factors have distinct roles in cell lineage specification[77] and the regulation of their expression is intertwined.[74,83,85–89] Nanog exerts its action in part via TET1/2 methylcytosine hydroxylases.[90]

[b] Noncoding RNAs also regulate the expression of the nuclear hormone receptor ESR[191] and the CEBPA (CCAAT enhancer-binding protein alpha).[92]

examples of lncRNAs that interact with Sox2 to regulate pluripotency, neurogenesis, neuronal differentiation and brain development,[122,125–128] and a lncRNA has been shown to interact with a chromatin-remodeling complex to induce nucleosome repositioning.[129]

Sox2, Nanog and Oct4 are often found at super enhancers[130] and state-specific differences in enhancer activity correspond with reconfiguration of Sox2, Nanog and Oct4 binding and target gene expression.[111] The lncRNA Evf2 selectively represses genes across megabase distances by coupling recruitment and sequestration of Sox2 into phase-separated domains (see below), affecting enhancer targeting and activity, with genome-wide effects.[122] In human embryonic stem cells Oct4 and Nanog associate with transcripts of the human endogenous retrovirus subfamily H (HERV-H) transposable elements, which are required to maintain stem cell identity and whose terminal repeats function as enhancers.[131,132]

The classic master switch transcription factor, MyoD, which can reprogram fibroblasts into muscle cells and is central to muscle differentiation *in vivo*,[133] is regulated by lncRNAs,[134–136] as are other aspects of muscle gene expression.[137–139] The pioneer transcription factor CBP also binds RNAs, including those transcribed from enhancers, to stimulate histone acetylation and transcription.[140]

Nucleosome repositioning and remodeling is accomplished by the ATP-dependent imitation switch (ISWI), chromodomain helicase DNA-binding (CHD), SWI/SNF (switch/'sucrose non-fermentable') (SWI/SNF) and INO80 complexes.[141,142] These complexes are directed to specific sites in chromatin or antagonized by lncRNAs, including Xist and enhancer RNAs, in processes as diverse as rRNA synthesis, myogenic differentiation and proliferation, endothelial proliferation, migration and angiogenic function, atherosclerosis, cardiomyopathy, liver regeneration and stem cell renewal, immunity and inflammation, and various cancers,[129,136,143–160] leading one group to conclude that "every cell type expresses precise lncRNA signatures to control lineage-specific regulatory programs".[160]

However, the patchy data on the binding of RNAs by the various proteins that control chromatin remodeling during development reflects limited investigations because of the expectation that all 'transcription factors' bind to DNA, rather than be directed by RNA-DNA and other RNA-mediated interactions.

GUIDANCE OF TRANSCRIPTION FACTORS

Loose DNA sequence specificities are a feature of eukaryotic transcription factors generally. While eukaryotic genomes are orders of magnitude larger than those of prokaryotes, their more conventional TFs have shorter DNA recognition sites (6–10bp versus 15–25bp in *E. coli*[161]), often expressed as a 'consensus' sequence, but better represented by multiple sequences, with many TFs recognizing different primary and secondary motifs.[162–166]

Moreover, high-throughput chromatin immunoprecipitation experiments with antibodies against specific TFs show different patterns of binding in different cell types, so additional factors must be involved. Such factors can be either (or both) trans-acting signals or chromatin accessibility, the latter supported by the observation that TF-binding sites are nucleosome depleted and DNase-sensitive, indicating that epigenomic decisions precede TF factor binding.[167–170]

The largest class of TFs in animals and plants contain 'zinc-finger' (ZF) domains,[c] specifically the C2H2 class, of which there are over 700 encoded in the human genome, and which recognize more sequence motifs than all other transcription factors combined.[171] Human C2H2-ZF proteins contain an average ~10 C2H2 domains (ranging from 1 to 30), classified into three groups: 'triple', 'multiple-adjacent', and 'separated-paired' C2H2 finger proteins, enabling some to bind multiple ligands.

It is thought that most ZFs bind to DNA, although most of the binding sequences are unidentified,[165] but many ZFs also bind to RNA or protein, and some to RNA only.[162,172,173] The classic example is TFIIIA (Chapter 8), which is required for the transcription of 5S rRNA genes and is titrated off DNA by its higher affinity for 5S rRNA, the first demonstration of the regulation of TFs by RNAs.[174,175]

A large fraction of C2H2-ZF TFs have been shown to regulate alternative splicing.[176] A splice variant that introduces three additional amino acids (KTS) between the third and fourth ZFs of the Wilm's tumor protein WT1[d] changes the specificity of the WT1 protein from DNA to spliceosomes,[178] presumably by binding RNA, given that WT1 also contains

[c] So-called because they have a domain shaped like a finger that is structured by a coordinated zinc ion.

[d] Frequently mutated in pediatric kidney tumors and urinogenitary developmental disorders.[177]

an RNA recognition motif[179] and transcription and splicing are coupled.[180] Disturbance of the ratio of +/-KTS isoforms causes a developmental syndrome, affecting kidney and genital development.[181] Both isoforms bind DNA and RNA *in vitro*,[182–184] shuffle between the nucleus and translating polysomes in the cytoplasm,[185] and their subnuclear location is RNase- but not DNase-sensitive.[182]

A 1994 analysis by Yigong Shi and Jeremy Berg of two representative C2H2-ZF proteins, one of which was Sp1 (which controls the expression of many housekeeping, tissue-specific, cell cycle and signaling pathway response genes[186]), showed that they have a higher affinity for RNA-DNA hybrids than for double-stranded DNA and that this increased affinity was strand-specific, i.e., dependent on which strand is RNA.[187]

The C2H2-ZF transcription factor YY1, which regulates the expression of various genes during embryogenesis, cell differentiation and proliferation,[188] binds chromatin in an RNA-enhanced fashion[189] and appears to play a major role in mediating enhancer-promoter loops.[190] YY1 also interacts with an RNA-binding protein involved in splicing regulation, depletion of which attenuates YY1 chromatin binding and YY1-dependent DNA looping and transcription.[29] The ZF-containing TAD insulator CTCF has also been shown to be a high-affinity RNA-binding protein.[65,191–195]

Later studies confirmed that over 800 human proteins bind RNA-DNA hybrids and over 300 prefer binding RNA-DNA hybrids over dsDNA.[196] These observations raise the possibility, if not the likelihood, that trans-acting RNAs are involved in the exposure and selection of genomic TF-binding sites, explaining the differential locus specificity of TF binding and the reason for a loose consensus sequence,[e] as well as enabling directionality of action by strand selection.

RNA-DNA hybrids[f] (which form 'R-loops' with the displaced DNA strand) occur widely throughout the human genome[198,199] and even encompass 8% of the yeast genome.[200] RNA-DNA hybrids are enriched at unmethylated CpG-rich promoters, transcription start sites and regions enriched for activating histone modifications such as H3K4me1/2/3, H3K9ac and H3K27ac.[201] RNA-DNA hybrids regulate genome stability and DNA repair,[197,202,203] promoter-proximal chromatin architecture and cellular differentiation,[204] transcriptional activation[205] and "are enriched at loci with … potential transcriptional regulatory properties … supporting a model of certain transcription factors binding preferentially to the RNA:DNA conformation".[206] The formation and stability of RNA-DNA hybrids are in turn regulated by RNA methylation and other modifications[197,207] (Chapter 17).

Nucleic acid triplex structures[g] (wherein a single stranded DNA or RNA forms 'Hoogsteen' hydrogen bonds with the purine-rich strand of polypyrimidine-polypurine tracts in the major groove of duplex DNA) also occur *in vivo*, as first detected by the sequence-specific binding of RNA to 'native' dsDNA.[214] Triplex-forming sequences are overrepresented in eukaryotic but not bacterial genomes, notably in regulatory regions and promoters.[215–217] Antibody and sequencing studies have also shown that triplex structures abound in eukaryotic chromosomes[218–221] (Figure 16.1). Triplex hotspots targeted by lncRNAs have been proposed to contribute to chromatin compartmentalization in conjunction with 'architectural' TFs such as CTCF[222] and the positions of lncRNA:DNA triplex-forming sites have been shown to be predictors for TADs.[223] Triplex-forming oligonucleotides have been shown to alter cell division, inhibit tumor growth, stimulate recombination and modulate target gene expression.[224–227]

Many lncRNAs, including those expressed from enhancers, have been shown to interact sequence-specifically with DNA[228] to regulate various processes through R-loop or triplex formation, including chromatin architecture, transcription, radiation response, cell proliferation, cell differentiation and organ development, in some cases (at least) intersecting with epigenetic pathways.[204,217,229–240] Triplexes are also involved in small RNA-mediated transcriptional gene silencing.[241]

[e] A subset of which can be specifically addressed by exact match to a transacting RNA, either a small RNA or an RNA sequence within a longer RNA.

[f] Interestingly, the stability of genomic RNA-DNA hybrids *in vivo* is controlled by methylation of the RNA (see below).[197]

[g] There are other alternative and multi-stranded structures in eukaryotic genomes, including Z-DNA (binding domains for which occur in RNA editing enzymes, see Chapter 17), G-quadruplexes, I-motifs and cruciform structures, which, regrettably, despite the availability of specific antibodies, have not been mapped in genome-wide studies of genomic features and their dynamic relationship to cell type.[208–212] Many of these alternative DNA structures are formed by simple sequence repeats, which also abound in the genomes of plants and animals.[213]

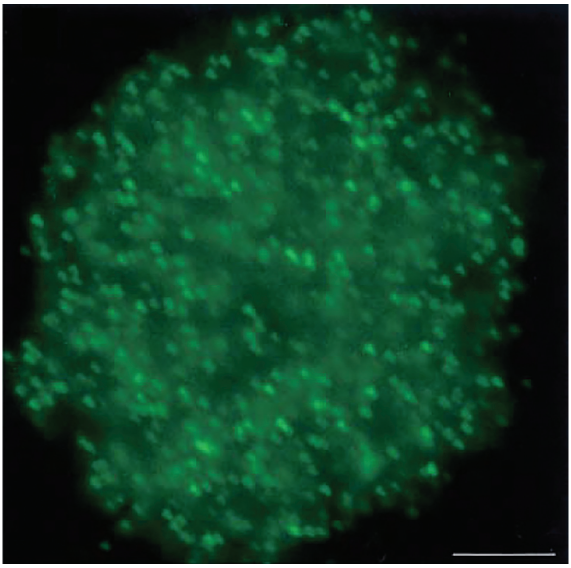

FIGURE 16.1 Triplex-forming DNAs in the interphase nucleus of a human monocytic leukemia cell visualized *in situ* by an anti-triplex monoclonal antibody. The bar represents 5 μm. (Reproduced from Ohno et al.[220] with permission of Springer Nature.)

A good example, from plants, is the lncRNA APOLO, which coordinates the expression of multiple genes in response to cold through sequence complementarity and R-loop formation, decoys Polycomb and binds transcription factors at the promoter of a master regulator of root hair formation.[242–244] Amazingly, APOLO function can be partly mimicked by the sequence-unrelated lncRNA UPAT, which interacts with orthologous proteins in mammals, indicating conservation of regulatory structures and lncRNA functions across kingdoms.[244]

The enhancer lncRNA KHPS1 forms a triplex with enhancer DNA sequences to activate expression of the neighboring *SPHK1* gene, by evicting CTCF, which insulates the enhancer from the SPHK1 promoter. Deletion of the triplex-forming sequence attenuates SPHK1 expression, leading to decreased cell migration and invasion, and the targeting of KHPS1 lncRNA can be switched by swapping the triplex-forming promoter sequence to other genes.[232,233]

Other classes of ‘transcription factors’ such as Y-box proteins also bind RNA, and known RNA-binding proteins such as hnRNP K (better known as a ‘splicing factor’) also act as transcription factors.[245,246] The ‘paired-box’ transcription factor Pax5, which is a ‘master regulator’ of B-cell development by recruiting chromatin-remodeling, histone-modifying and basal transcription factor complexes to its target genes,[247] is hijacked to the Epstein Barr Virus genome by a viral-encoded non-coding RNA.[248] Another ‘transcription factor’, the nuclear hormone receptor ESR1 (estrogen receptor α), which is commonly activated in breast cancer, is also an RNA-binding protein.[249]

Dual RNA-DNA or ambiguous RNA/DNA-binding proteins also include p53,[172,250] the ‘guardian of the genome’, possibly the most intensively studied gene and protein in human molecular biology, which binds a lncRNA (‘damage-induced noncoding RNA’, DINO).[251] The dual DNA/RNA-binding protein TLS/FUS (Translocated in LipoSarcoma/FUsed in Sarcoma) is allosterically regulated by lncRNA pncRNA-D.[252–254] Even RNA polymerase is regulated by RNAs. In mammals, RNA polymerase II is repressed by short RNA polymerase III transcripts derived from mouse B2 and human Alu repeat (SINE) elements.[255–258] These elements also provide mobile RNA polymerase II promoters.[259]

Importantly, approximately half (~350) of the human C2H2-ZF proteins, many of which are unique to primates, contain a KRAB transcriptional repression domain, which binds TEs,[h] and evolved by recurrent TE capture that partners them with emergent TE-mediated regulatory networks, influencing genomic imprinting, placental growth and brain development.[171,260–270] Most eukaryotic TFs also contain intrinsically disordered domains that overlap their DNA-binding domain and direct their target specificity,[271–275] likely by interaction with guide RNAs (see below).

Presumably, different types of DNA/RNA-binding proteins recognize different types of nucleic acid structures and transact a different type of signal in different contexts within the decisional systems that control cell division and differentiation during development, as well as in physiological responses. The fact that most eukaryotic ‘transcription factors’ have confusing and enigmatic functions attests to the likelihood that they have been interpreted in the wrong conceptual framework, with RNA the missing link.[276]

[h] KZFPs (KRAB domain-containing zinc finger proteins) control the pleiotropic activation of TE-derived transcriptional cis-regulator sequences, some of which are primate-specific, during early embryogenesis, in part through histone H3K9me3-dependent heterochromatin formation and DNA methylation.[260,261] Primate-specific KZFPs also regulate gene expression in neurons.[262]

GUIDANCE OF DNA METHYLATION

Transcriptional gene silencing in fungi and plants by RNA-directed DNA methylation was well established in the 1980s and 1990s (Chapter 12). These studies eventually showed that the enzymes that methylate DNA are directed to their sites of action by small RNAs interacting with the RNAi protein AGO4.[277,278] Small RNAs (miRNAs, siRNAs and piRNAs) also induce site-specific DNA methylation in animals,[52,53,279,280] which again involves Argonaute proteins,[281–285] suggesting that what had originated as an RNA-based mechanism for defense against viruses has been co-opted as a means of genome regulation.[i]

In 2004, Linda Jeffery and Sara Nakielny showed that the *de novo* DNA methylases Dnmt3a and Dnmt3b, but not the maintenance methylase Dnmt1, bind siRNAs with high affinity.[287] Later others reported that Dnmt1 (which restores methylation at hemi-methylated CpG sites after DNA replication) binds lncRNAs to alter DNA methylation patters at cognate loci.[288–291]

Demethylation also appears to be an active process guided by RNAs.[292–295] RNA-directed DNA demethylation has also been reported to involve R-loop formation,[239,294] and recruitment of the TET2 dioxygenase/demethylase (which unlike other TET enzymes does not contain a DNA-binding domain, but does bind RNA[296,297]) by RNAs transcribed from endogenous retroviruses.[298]

Some methyl-CpG-binding proteins, including MeCP2, bind siRNAs and other RNAs, mediated through a domain distinct from the methyl-CpG-binding domain with, interestingly, RNA and methyl-CpG binding being mutually exclusive,[287,299] although there seems to be variations of the regulatory mechanisms, including RNA-mediated recruitment to phase-separated heterochromatin compartments (see below). LncRNAs have also been shown to link DNA methylation with histone modification through triplex formation with target sequences.[300]

GUIDANCE OF HISTONE MODIFICATIONS

A range of histone variants and over 100 different histone modifications are differentially incorporated into nucleosomes located at millions of different positions in different cell types and different stages of development and differentiation (Chapter 14). However, like DNA methylation enzymes, histone-modifying enzymes also have no intrinsic DNA-binding capacity or specificity, which is often assumed to be provided by sequence-specific DNA-binding proteins or transcription factors that interact with them. On the other hand, like DNA methylation enzymes, many histone modification writers and readers contain domains that bind RNA and/or contain RNA-binding modules. These include RNA recognition motifs,[301] chromodomains,[296,302,303] bromodomains,[296] Tudor domains,[304] PRC2 subunits EZH2, EED, Suz12 and Jarid2,[305–310] the H3K20 trimethylase Suv4–20h[60] and other histone-modifying complexess.[311]

RNA binding to histones was first reported in the mid-1960s[312,313] (Chapter 4). Around the turn of the century, a number of groups showed that PcG (Polycomb group) proteins from *C. elegans* and vertebrates also bind RNA, and that this binding is essential for their chromatin localization and repression of homeotic genes.[314–316] In 2005, Renato Paro and colleagues showed that the switch from the silenced to the activated state of a Polycomb response element in the *Drosophila bithorax* locus (Chapter 5) requires non-coding transcription.[317]

In 2007, John Rinn and colleagues showed that lncRNAs transcribed from human homeotic gene loci, like those in *Drosophila*, are expressed along developmental axes and demarcate active and silent chromosomal domains that have different H3K27me3 profiles and RNA polymerase accessibility, the exemplar of which, HOTAIR (Chapter 13), interacts with PRC2 and is required for PRC2 occupancy and histone H3K27 trimethylation at the *HOXD* locus.[318] Other studies showed, for example, that retinoic acid-induced expression of lncRNAs follows the collinear activation state and correlates with loss of Polycomb repression at the *HOXA* locus.[319]

In 2008, the groups of Chandrasekhar Kanduri and Peter Fraser showed that lncRNAs differentially expressed from parentally imprinted loci, specifically the 105 kb Air RNA and the 91 kb Kcnq1ot1 RNA (formerly KvLQT1-AS, Chapter 9), and later others,

[i] Transcriptional gene silencing can be induced by siRNAs in the absence of DNA methylation,[286] indicating that small RNAs participate in other pathways that control chromatin state and architecture.

also bind PRC2 to repress the relevant alleles.[320–322] In the same year, we showed that lncRNAs expressed from the antisense strand of homeotic gene loci during embryonic stem cell differentiation are associated with both the chromatin-activating Trithorax MLL1 complex and activated chromatin containing H3K4me3 marks.[323] Subsequently other lncRNAs (including enhancer RNAs and small RNAs derived from them) were also shown to associate with Trithorax complexes, including RNAs involved in maintenance of stem cell fates and lineage specification (such as Evx1-as and HOTTIP),[14,45,324–334] and even grain yield in rice.[335]

In 2009, Rinn and colleagues surveyed over 3,300 lncRNAs and showed that ~20% (but only ~2% of mRNAs) interact with PRC2, and that others are bound by other chromatin-modifying complexes.[336] Moreover, knocking down a selection of these RNAs caused derepression of genes normally silenced by PRC2.[336] Over 9,000 RNAs bind PRC2 in embryonic stem cells,[306] with many individual cases, including the lncRNAs H19, MEG3, ANRIL[j] and HOTAIR, subsequently characterized in some detail.[303,309,337–342]

RNA has also been shown to be required for PRC2 chromatin occupancy, PRC2 function and cell state definition.[343] Short RNAs transcribed from Polycomb-repressed loci resemble PRC2-binding sites in *Xist*, and interact with PRC2 through its subunit Suz12.[307] PRC2 binds G-quadruplex structures in RNA,[344] which inhibit PRC2 activity and are antagonized by allosteric activation of PRC2 by H3K27me3 and regulators of histone methyltransferases,[345,346] indicating complex decisional transactions.

PRC1 function also appears to be controlled by RNA[303] and PRC1 resides in membrane-less phase-separated nuclear organelles[347] that are likely to be RNA nucleated (see below).

PRC2[k] binds many RNAs 'promiscuously',[348] a description[l] that does not mean 'non-specifically'.[339,349] The association of Polycomb and Trithorax complexes with many RNAs, likely through orthologous domains (see below), is consistent with their function as guide molecules for RNA-directed site-specific histone and DNA modifications. It is also consistent with the fact that Polycomb and Trithorax proteins are involved in many differentiation and developmental decisions, from cell cycle regulation to embryogenesis and body plan specification,[350–355] so their binding to many different guide RNAs would be expected. Again, these include RNA-DNA, RNA-RNA and RNA-protein interactions, recruitment or eviction of histone modifiers and other chromatin-modifying proteins, alteration of DNA topology, allosteric inhibition, and reorganization of phase-separated domains.[342,343,345,346,356–364]

LncRNAs also control switching between Polycomb and Trithorax response elements.[329,356] Other histone modifications are also regulated by lncRNAs,[m] including during memory formation.[366] An intriguing observation, not inconsistent with RNA involvement, is that histone-modifying enzymes, rather than the parental histones, may remain associated with DNA through replication to re-establish the epigenetic information on the newly assembled chromatin.[367]

XIST AS THE EXEMPLAR

While initially thought to be a special case, the best characterized and most illustrative example of the complex interplay between lncRNAs, chromatin structure and gene expression is Xist.[368,369] Xist has eight exons and is 17 kb in length.[370] It has a highly modular structure, including a number of types of conserved 'repeat' sequences, and interacts with over 80 different proteins including cohesin, Polycomb and other chromatin remodelers,[370–374] at low copy number.[375,376]

Xist-mediated silencing of the inactive X chromosome in mammals requires its repeat sequences[377–382] and involves Polycomb recruitment, deacetylation of H3K27ac and H3K27 methylation on the silenced chromosome.[305,368,382–387] Spreading involves the partitioning of chromatin topology[388–392] by the formation of phase-separated domains[376,393,394] and the interaction with RNA-binding proteins with repetitive elements (in particular LINE-1 elements) in the X chromosome to recruit silencing mechanisms targeted to repeats,[379,394–399] as first proposed by Mary

[j] ANRIL also binds PRC1.

[k] For many reasons, PRC2 has been the most intensively studied of all of the histone-modifying complexes with respect to the role of lncRNAs in epigenetic regulation of gene expression.[339]

[l] The presence of intrinsically disordered domains in most proteins involved in development has also been described as conferring promiscuity, a functional trait that allows flexible interactions in regulatory networks (see below).

[m] LncRNAs also control the methylation of a number of non-histone proteins involved in cell signaling, gene expression and RNA processing.[365]

Lyon,[400,401] likely via triplex formation.[402] Xist also acts as a suppressor of hematological cancers[403] and is essential to maintain X-inactivation of immune genes, dysregulated in females suffering systemic lupus erythematosus or COVID-19 infection.[385]

Xist expression and action is controlled and effected by other lncRNAs[369] that are expressed antisense to *Xist* (Tsix, which blocks RepA RNA binding to PRC2[305,404,405]) or from adjacent loci on the active or inactive X chromosomes[406–408] (Figure 16.2). The Jpx lncRNA, whose gene resides ~10 kb upstream of *Xist*, activates *Xist* by regulating CTCF anchor site selection to alter the topography of chromosome loops.[37,409,410] A primate-specific TE-derived lncRNA, XACT, coats active X chromosomes in pluripotent cells and is connected into the pluripotency regulatory network in humans by primate-specific retroviral enhancer.[411] Another lncRNA expressed from the X-inactivation center, Tsx, functions in germ and stem cell development as well as in learning and behavior.[412] The lncRNA Firre, which also contains a number of repeats, one of which interacts with the nuclear matrix factor hnRNPU, anchors the inactive X chromosome near the nucleolus and is required for the maintenance of its repressive H3K27me3 marks.[413] Firre is also required for the topological organization of other chromosomal regions,[414,415] and is involved in other developmental processes including adipogenesis and hematopoiesis.[416]

X-inactivation also involves the RNAi enzyme Dicer[368,417,418] and methylation of *Xist* transcripts,[419] implicating small RNAs, the RNA interference pathway and RNA modifications in a complex set of decisional pathways that control chromatin architecture from yeast to humans.[420,421] Chromosomal dosage compensation in *Drosophila*, which involves global activation of the single X chromosome in males is controlled by the lncRNAs roX1 and roX2,[422] via a conserved predicted stem-loop structure required for histone H4K16 acetylation of the X chromosome[423] and selective X-chromosome subnuclear compartmentalization.[424]

ENHANCER RNAs AND CHROMATIN STRUCTURE

As discussed in Chapter 14, enhancers play a key role in specifying cell identity and are a signature feature of the regulation of gene expression during development.

Enhancers were initially identified by their activity, rather than their physical manifestation, but have been interpreted in terms of the initial speculations about the latter, postulated and promulgated by Mark Ptashne (and widely accepted) to comprise cluster of binding sites for TFs that act at a distance by 'looping' to make contact with the promoters of target genes, some of which can be located hundreds of kilobases distant.[425–431]

However, early studies had shown that lncRNAs are transcribed from enhancer regions in well-studied loci,[432–434] with supporting evidence accumulating

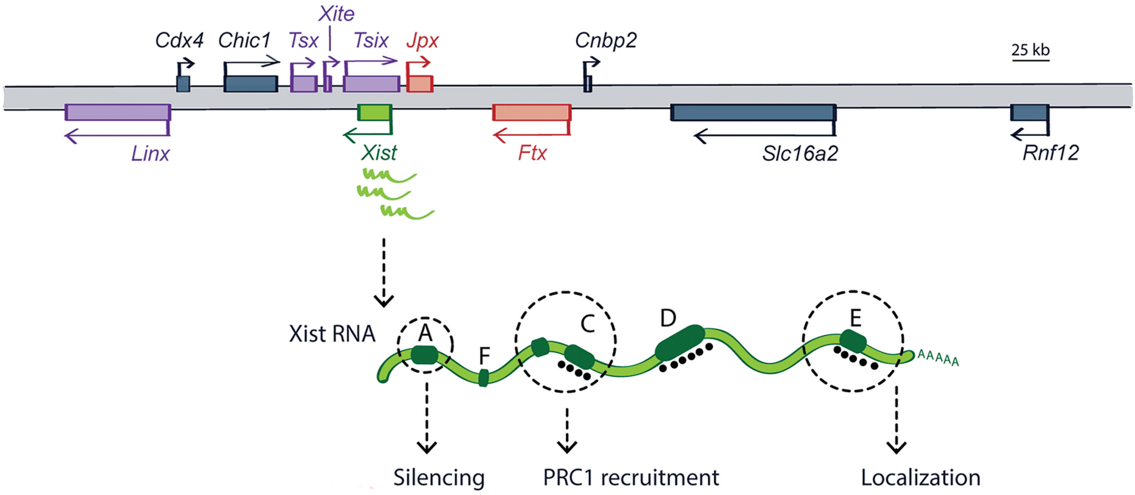

FIGURE 16.2 The organization of the *Xist* locus. The *Xist*, *Tsix*, *Jpx*, *Xite*, *Tsx* and *Ftx* genes specify lncRNAs. (Reproduced from Loda and Heard[399] under Creative Commons attribution license.)

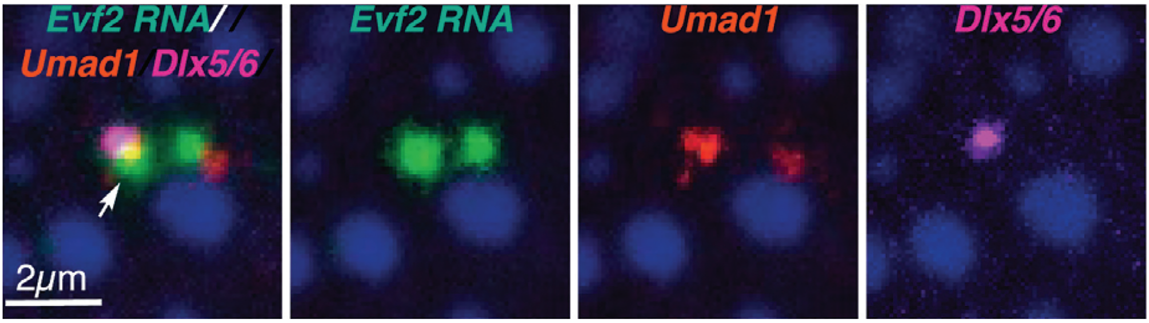

FIGURE 16.3 Cloud formed by the enhancer lncRNA Evf2 and its localization to activated (*Umad1*, 1.6 Mb distant) and repressed (*Akr1b8*, 27 Mb distant) target protein-coding genes. (Reproduced with permission from Cajigas et al[473] with permission of Elsevier.)

with improving transcriptomic and chromatin analysis technologies. Although sometimes referred to as (protein-coding) 'gene deserts',[435] enhancers exhibit the characteristics of *bona fide* genes, including nucleosome-depleted promoter regions that bind transcription factors and the transcription of adjacent sequences.[436–441] Indeed, the epigenetic architecture of, and the features of transcription initiation at, the promoters of conventional protein-coding genes and enhancers are almost indistinguishable.[436,441–443]

Enhancers and 'super-enhancers' are transcribed to produce non-coding RNAs specifically in the cells in which they are active[317,435,437,441–447] and their expression is considered the best molecular indicator of enhancer activity in developmental processes[437,443,448–453] and cancers.[91,454–457]

Enhancers recruit RNA polymerase[458] and produce short unstable bidirectional transcripts ('eRNAs') from their promoters,[437,444,459–462] as do protein-coding genes,[430,463–465] and it is uncertain whether these transcripts play a role in enhancer action or simply mark active promoters and/or reflect promiscuous RNA polymerase initiation at accessible chromatin.[441,461,464,466–468] On the other hand, enhancers also express multi-exonic lncRNAs, the half-life of which is exosome regulated,[466] and many if not most lncRNAs likely derive from enhancers.[441,444,446,469–479] Enhancers with tissue-specific activity are enriched in introns, suggesting that "the genomic location of active enhancers is key for the tissue-specific control of gene expression".[480]

There is good evidence that enhancers regulate chromosome looping and local chromatin reorganization to alter cell fate,[435,441,451,481,482] which is consistent with but does not demonstrate direct contact between enhancer TF-binding sites and the promoters of target genes. It is also consistent with enhancer RNAs organizing looping with target genes, which has been demonstrated in at least two cases,[482,483] and/or the formation of topologically associated, possibly phase-separated, chromatin domains[16] as local hubs of transcription regulation (Figure 16.3). Recent studies suggest that there is no direct contact between TFs bound at the enhancer and the promoter of genes regulated by enhancer action,[484] and that maintenance of enhancer-promoter interactions and activation of transcription are separable events.[485]

At the heart of the debates about the mechanism of enhancer action has been the question of whether the RNAs transcribed from active enhancers are simply a passive by-product of TF occupancy, or whether it is the 'act of enhancer transcription' or the enhancer RNAs themselves that mediate enhancer action.[442,447,486,487] The evidence for these possibilities, which are not mutually exclusive, has been widely canvassed and variously interpreted, with the former initially favored because it fitted the TF paradigm of protein-coding gene regulation and did not require acceptance of large numbers of regulatory RNAs.[441,476,487–489] In line with these preconceptions, some studies have reported that the transcribed enhancer RNA sequences can be partly/substantially (it is difficult to be sure[n]) deleted, truncated or replaced with no obvious effect (see, e.g., [490,491]).

Other studies showed that enhancer RNAs are required for enhancer activity.[92,473,479,482,492–498] For example, deletion in mice of the multi-exonic lncRNA Maenli, which is expressed from an enhancer that controls limb development and is deleted in a human

[n] Due to incomplete characterization of the enhancer RNA/transcription unit.

developmental disorder, recapitulates the human phenotype.[479] Deletion of internal exons of the enhancer-derived lncRNA ThymoD blocks T-cell development and causes developmental malignancies.[482] Truncation of the lncRNA Evf2, which is transcribed from the highly conserved Dlx5/6 enhancer that spatially organizes the expression of a 27 Mb region on chr6 during mouse forebrain development, abrogates the action of the enhancer.[473,499] The latter study also showed that the 5' and 3' ends of the Evf2 enhancer RNA had different functions,[o] and that Evf2 formed an "RNA cloud" encompassing its target genes.[473] It has also been shown that methylation and splicing of enhancer RNAs are required for enhancer function and chromatin organization.[490,500–502]

siRNA-mediated knockdown of enhancer RNA also abrogates or reduces enhancer action, demonstrating the involvement of the RNA.[477,503–507] RNA has been shown to be required for the formation of enhancer-target promoter contacts by the transcription factor YY1[190] (which itself regulates the expression of many lncRNAs[508]) and ectopic expression of enhancer RNAs upregulates expression of the genes normally targeted by the enhancer.[472,509]

Careful analysis shows that the variable phenotypic consequences of enhancer lncRNA knockdown, truncation or ablation depend on the details.[496,510,511] For example, a short deletion of the promoter and first two exons of the 17 kb lncRNA Hand2os1, which is expressed from an enhancer essential for heart morphogenesis, did not produce discernable heart phenotypes, but deletion of exons 4 and 5 caused severe contraction defects in adult heart that worsened with age, and deletion of the entire *Hand2os1* sequence led to dysregulated cardiac gene expression, septum lesion, heart hypoplasia and perinatal death.[510]

It has been shown that enhancer RNAs produced in response to immune signaling bind the bromodomains of BRD4 (and other epigenetic reader bromodomain-containing proteins) to augment BRD4 enhancer recruitment and transcriptional cofactor activity.[512] BRD4 also cooperates with oncogenic fusions of MLL1 to induce transcriptional activation of enhancer RNAs, one of which has been shown to bind histone H4K31ac to promote histone recognition and oncogene transcription.[513] Other mechanisms may involve the interplay between different types of regulatory RNAs and chromatin-associated proteins, as suggested by interactions between NEAT1 and BRD4/WDR5[p] complexes, and enhancer RNAs with cohesin, with effects on specific target genes.[474] Finally, there is also evidence of concerted action of *cis*-acting enhancer RNAs with other transcripts that have trans-acting roles.[514–516]

While there may not yet be universal acceptance, the evidence is accumulating that enhancer RNAs are integral to enhancer function,[517] and that enhancer RNAs are simply a (large) class of lncRNAs that regulate chromatin architecture and the expression of protein-coding and (other) lncRNAs, albeit through physical mechanisms that are not yet well understood, but involve recognition of effector proteins and sequence-specific RNA-DNA contacts via R-loops or triplexes,[17,221,236,518] and formation of topologically associated domains to form developmental stage-specific transcriptional hubs.[431,519,520] It is also evident that transcription itself modulates chromosome topology and phase transition-driven nuclear body assembly.[521–524]

There are ~400,000 enhancers (and ~400,000 differentially accessible chromatin elements,[525,526] which likely correspond to promoters) in the human gen ome.[438,441,444,448,456,460,527–533] This is similar to the number of lncRNAs expressed from the human genome (Chapter 13). Indeed, apart from the fact that they do not encode proteins,[q] enhancers might be properly viewed as genes, which, together with the multitude of other genes expressing functional non-coding RNAs, resolves the G-value enigma.

mRNAs may also have enhancer function,[12,490] which would not be surprising given the interwoven nature of the expression of genetic information during the complex ontogenies that underpin animal and, to a lesser extent, plant development.

RNA SCAFFOLDING OF PHASE-SEPARATED DOMAINS

It has been known for many years that there are many ribonucleoprotein (RNP) complexes that exist in defined territories in the nucleus and cytoplasm of eukaryotic cells, prominent examples of which

[o] Evf2, which has a human homolog, also interacts with Sox2 to alter its target specificity, regulates transcription of the homeodomain transcription factors Dlx5 and Dlx6 as well as cohesin binding, and influences chromatin remodeling in the formation of GABA-dependent neuronal circuitry.[122,145,473,499]

[p] WDR5 is a a core subunit of the human MLL1-4 histone H3K4 methyltransferase complexes.

[q] New mechanisms that control the expression and function of RNAs expressed from promoters of protein-coding genes and enhancers are still being identified, affecting splicing, elongation, termination and processing/half-life.[466,534–536]

include nucleoli, spliceosomes, paraspeckles and stress granules,[537–539] none of which are membrane-bound. One of the most important advances of recent years, first canvassed by Harry Walter and Donald Brooks in 1995, and later demonstrated by Clifford Brangwynne, Anthony Hyman and colleagues, is that these and other focal or 'punctate' organelles are phase-separated condensates or 'coacervates'[540,541] that compartmentalize biochemical and regulatory hubs,[542] although not without some controversy and uncertainty.[543]

Phase-separated condensates, which are heterogeneous in constitution and properties, are commonly referred to as 'phase-separated domains' (PSDs). They are also called 'liquid crystal domains', 'liquid droplets', 'biomolecular condensates', 'nuclear clouds' or 'nuclear bodies',[544] and exist in an aqueous state distinct from the surrounding environment, the biological manifestation of liquid or soft matter physics.[545]

In vitro PSDs form spontaneously by association of oppositely charged molecules such as negatively charged RNA interacting with positively charged proteins.[544] *In vivo* they are formed by interactions between RNAs,[r] RNA-binding proteins and proteins with intrinsically disordered domains (IDRs),[27,539,542,544,546–553] explaining the latter's previously mysterious function.[539]

IDRs lack rigid tertiary structure and are characterized by a high proportion of small, polar and positively charged amino acids (arginine, histidine and lysine), often in the form of RGG/RG, histidine-rich domains or other repeats.[554–557] IDRs are promiscuous, i.e., they interact with and are tunable by many partners.[550,556–561] Intrinsically disordered RGG/RG domains mediate specificity in RNA binding,[562,563] and IDRs flank the DNA-binding domains of transcription factors,[273] direct TF binding and, *inter alia*, the temporal regulation of transcription complexes that specify neuronal subtypes.[564]

IDRs and PSDs occur in bacteria and archaea,[565–568] but there have been sharp increases in the fraction of the proteome containing IDRs between prokaryotes, simple eukaryotes and multicellular organisms, and the number of proteins containing IDRs correlates with the number of cell types, suggesting co-evolution of IDR-mediated transactions with developmental complexity.[559,560]

IDRs are usually located at the N- or C-terminal region of the protein. IDRs are present in and essential for the function of nearly all of the proteins involved in animal and plant development,[s] including RNA polymerase, most transcription factors, Hox proteins, histones, histone-modifying proteins, other chromatin-binding proteins, the Mediator complex, RNA-binding proteins, splicing factors, membrane receptors, cytoskeletal proteins and nuclear hormone receptors.[271,272,274,539,551,560,563,570–577]

Surprisingly, the majority of proteins subject to alternative splicing contain IDRs.[560] Moreover, IDRs are overrepresented in alternatively spliced exons subject to tissue- and lineage-specific regulation,[578–581] especially in exons that are alternatively spliced in mammals but constitutively spliced in other vertebrates,[582] which changes the subcellular localization of the isoform and coordinates phase transitions within the cell.[583,584]

IDRs also occur in proteins that are flexibly involved in signal transduction and transport,[539] such as the Ras-GTPase-activating proteins (SH3 domain)-binding proteins G3BP (which forms phase-separated domains),[585] clathrin-mediated endocytosis[586] and synapsins, which are required for the maintenance of synaptic vesicle clusters in neurons by IDR-mediated phase separation.[587]

IDRs are major sites of post-translational modifications and many biological processes, including the regulation of the cell cycle and circadian clocks,[539] have been shown to be dependent on post-translational modification of IDRs.[539,588–590] Post-translational modifications modulate RNA binding[591] and alter the propensity to nucleate PSDs,[592] adding layers of complexity to their interactions and regulation.[539,557,593,594]

The known post-translational modifications not only include the 100 or so found in histones, but also 95 in the IDR of the axonal microtubule–associated protein Tau, which is involved in Alzheimer's and other neurodegenerative diseases.[595] Tandem RNA-binding sites in the RNA-binding protein, TIA-1, facilitate PSD stress granule formation[596] and reduction of this protein protects against Tau-mediated neurodegeneration.[597]

Other proteins involved in neurological functions and disorders, such as TDP-43, ataxin, c9orf72 and FMRP (fragile X mental retardation protein), also contain IDRs that are involved in phase separation,

[r] 'RNA regulates the formation, identity, and localization of phase-separated granules'.[542]

[s] A much higher percentage than found in the rest of the proteome.[272,569]

controlled in part by post-translational modifications.[598,599] For example, the IDR of TDP-43 binds RNAs,[600] and the loss of its RNA-binding ability by mutations or post-translational acetylation leads to its sequestration into PSDs.[601] The IDRs of ataxin mediate formation of neuronal mRNP assemblies, and are essential for long-term memory formation as well as c9orf72-induced neurodegeneration.[602] Neuronal-specific micro-exons overlapping IDRs in the translation initiation factor eIF4G regulate the coalescence of phase-separated granules to repress translation, are misregulated in autism, and their deletion in mice leads to altered hippocampal synaptic plasticity and deficits in social behavior, learning, and memory.[603]

Aberrant promiscuity of IDR-containing proteins (IDPs) and perturbations of PSD formation may underlie the dosage sensitivity of oncogenes and other proteins[604,605] as well as neurodegenerative disorders such as Alzheimer's Disease, Parkinson's Disease, Frontotemporal Dementia, Muscular Dystrophy and Amyotrophic Lateral Sclerosis, where repeat expansions affecting RNA and/or their encoded proteins result in pathological aggregates.[598,606,607]

Mutations in RBPs that cause human monogenic diseases are observed more commonly in IDRs than globular domains,[586,608] indicating that, despite their relatively simple composition, IDRs have strong sequence constraints. It is also clear that RNA nucleates the formation, and is the structural scaffold, of PSDs.[27,542,547,549,551,553,609–614]

PSDs encompass a range of nuclear compartments:[27,615] DNA replication initiation sites;[616] telomeres;[617] centrosomes[618] and meiotic chromosomal pairing foci;[619,620] germ granules;[592,621] nucleoli, Cajal bodies and 'histone locus bodies';[540,622–624] 'extranucleolar droplets';[523] spliceosomes ('nuclear speckles'[t]);[626] specialized spliceosomes (via lncRNAs Gomafu[u] and Malat1);[627–629] paraspeckles[v] (*Neat1*);[630–634] heterochromatin;[609,635] Polycomb bodies;[347] primate-specific nuclear stress bodies;[636,637] nuclear glucocorticoid receptor foci;[638] SARS-Cov2 viral assembly domains;[639] and others, including in plants.[640] They also include cytoplasmic organelles[537,641,642] such as P-granules,[643,644] G-bodies,[645] stress granules,[646] polar bodies (whose formation is dependent on a lncRNA),[647,648] localized mRNP translational assemblies[649,650] and synaptic compartments.[651]

It has been proposed that lncRNAs play a central role in organizing the three-dimensional genome,[521] including the formation of spatial compartments and transcriptional condensates[610,614,652–655] (Figure 16.4) and hence the four-dimensional patterns of gene expression during differentiation and development.[w] It has been shown that phase separation drives chromatin looping[657] and is required for the action of enhancers and super-enhancers;[551,610,658–660] that transcription factors activate genes through the phase-separation capacity of their activation domains by forming PSDs with RNA polymerase II;[661,662] that Mediator and RNA polymerase II associate in transcription-dependent condensates;[658,661–663] that phase separation of RNA-binding protein promotes polymerase binding and transcription;[664] and that PSDs scaffolded by lncRNAs, including repeat-derived RNAs, mediate heterochromatin formation,[32,609,614,665–669] euchromatin formation,[670] nucleolar structure,[671–674] splicing[675] and DNA damage repair.[676–678]

For example, it has been shown that the cytoplasmic lncRNA NORAD, which is induced by DNA damage and required for genome stability, prevents aberrant mitosis by sequestering Pumilio proteins (which bind many RNAs to regulate stem cell fate, development and neurological functions[680]) into PSDs via multiple repeats.[681–684] Lack of NORAD accelerates aging in mice.[685] Similarly sequestration of the double-strand beak enzyme RAG1 into nucleoli modulates V(D)J recombination activity.[686] Many natural antisense lncRNAs with embedded mammalian interspersed repeats are overrepresented at loci linked to neurodegeneration and/or encoding IDPs.[687]

The PARP1 superfamily,[x] one of the most abundant proteins in the eukaryotic nucleus,[688–690] which catalyzes the polymerization of ADP-ribose units and attachment of poly (ADP-ribose) polymers to arginines in target proteins,[y] including histones, for DNA repair, stabilization of replication forks

[t] p53 target gene association with nuclear speckles is driven by p53.[625]

[u] Japanese for 'spotted pattern'.

[v] Involving RNA-protein and RNA-RNA interactions.[613,630,631]

[w] PSDs may also serve to reduce noise in biological signal processing and control.[656]

[x] There are 18 members of the PARP superfamily encoded in the human genome. PARP2 is also involved in chromatin modification, PARP3 is a core component of centrosomes, and PARP4 is associated with vault particles (Chapter 8).[688]

[y] Interestingly, a similar enzyme in bacteriophage has been reported to add entire RNA chains to a host ribosomal protein to modulate the phage replication cycle.[691]

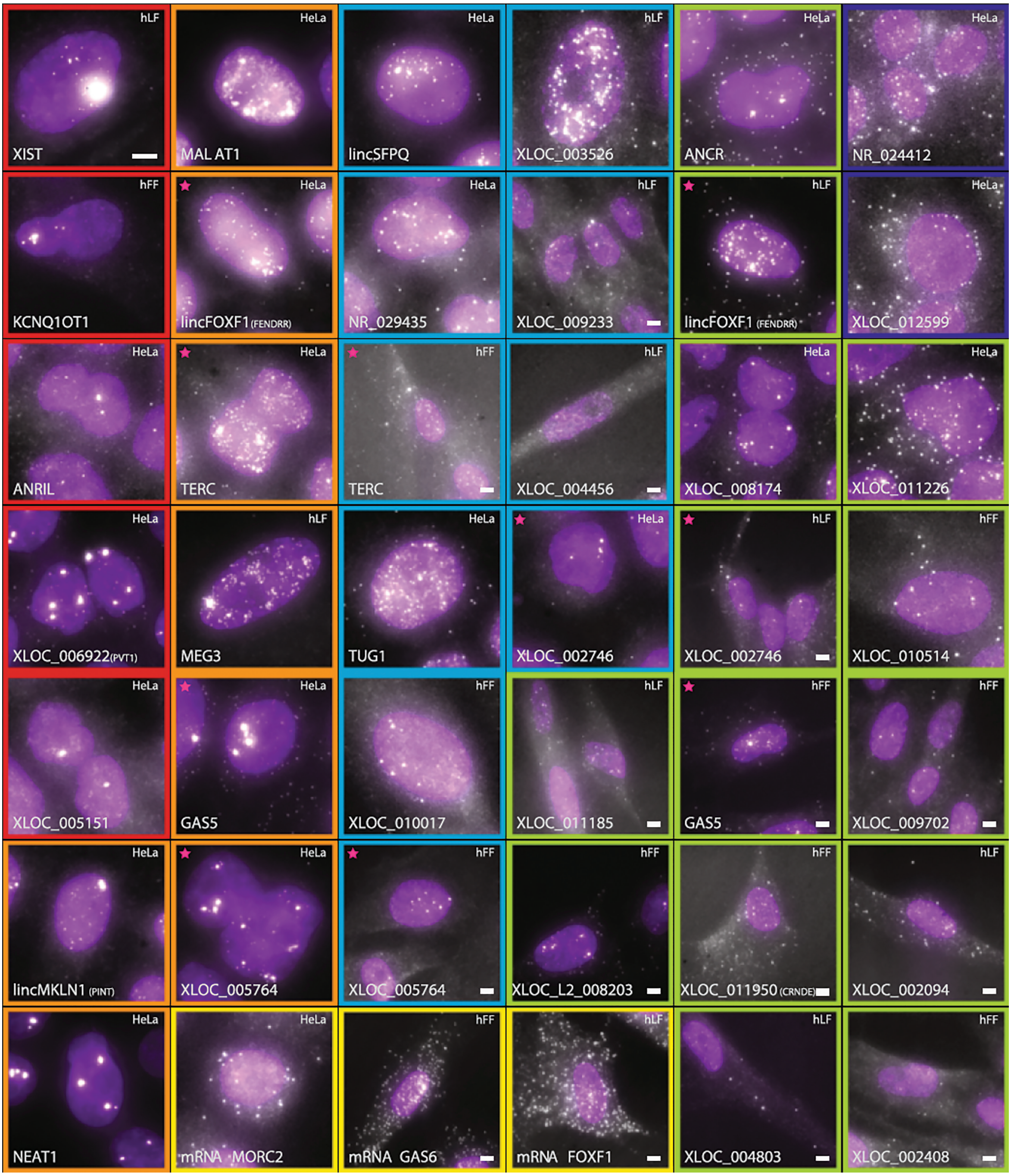

FIGURE 16.4 Subcellular and subnuclear localization of RNAs in punctate domains. (Reproduced with permission from Cabili et al.[679] with the permission of the authors under Creative Commons attribution license.)

and the modification of chromatin, also binds lncRNAs,[692–695] regulates RNA metabolism[696,697] and modulates the phase-separation properties of RNA-binding proteins.[698,699]

Many lncRNAs appear to be localized to defined nuclear and cytoplasmic foci that resemble liquid droplets,[336,414,473,679,700–702] and a genome-wide study identified hundreds of non-coding RNAs forming nuclear compartments near their transcriptional loci, in dozens of cases guiding cooperating proteins into these 3D compartments and regulating the expression of genes contained within them [228,614] (Figure 16.5).

X-chromosome dosage compensation in *Drosophila* requires the formation of a phase-separated coacervate by the lncRNAs roX1 and roX2 interacting with the IDR of a specific partner protein (MSL2, 'male sex lethal 2'). Moreover, replacing the IDR of the mammalian ortholog of MSL2 with that from *Drosophila* along with expression of roX2 is sufficient to nucleate ectopic dosage compensation in mammalian cells, showing that the roX–MSL2 IDR interaction is the primary determinant for compartmentalization of the X chromosome, and a likely exemplar of lncRNA-IDR interactions in general.[424]

As further evidence that the eukaryotic nucleus (and indeed the eukaryotic cytoplasm) is finely organized, and that many more phase-separated domains remain to be discovered, a recent report has shown that two related RNA modification enzymes that normally reside (in one case) in the nucleolus and (in the other) in an unknown cytoplasmic domain proximal to mitochondria, both relocate upon nerve cell depolarization to different small unknown punctate nuclear domains.[703]

The complexity is extraordinary and literature on this topic is burgeoning, but it is now clear that PSDs comprise a major and until recently unappreciated fine-scale and dynamic spatial regulation of subcellular and chromatin organization, "the active chromatin hub", first proposed by Wouter de Laat and Frank Grosveld in 1993 based on the study of globin enhancers,[704] which "unifies the roles of active promoters and enhancers".[520] It has also been proposed, with experimental support, that "ribonucleoprotein complexes can act as block copolymers to form RNA-scaffolding biomolecular condensates with optimal sizes and structures in cells".[634]

AN ADDITION TO THE ANCIENT RNA WORLD HYPOTHESIS

The ability of RNA to nucleate phase-separated domains adds a third dimension to its role in the origin of life.[705] While it has been widely accepted that RNA was likely the primordial informational and catalytic molecule of life, its advent would also have enabled the formation of a pre-cellular phase-separated privileged environment wherein organic reactions could be concentrated and evolve. Indeed, compartmentalized RNA catalysis has been demonstrated in membrane-free coacervate protocells.[706,707]

The development of RNA-nucleated coacervates likely involved its interaction with positively charged (particularly arginine-rich) disordered proteins.[708]

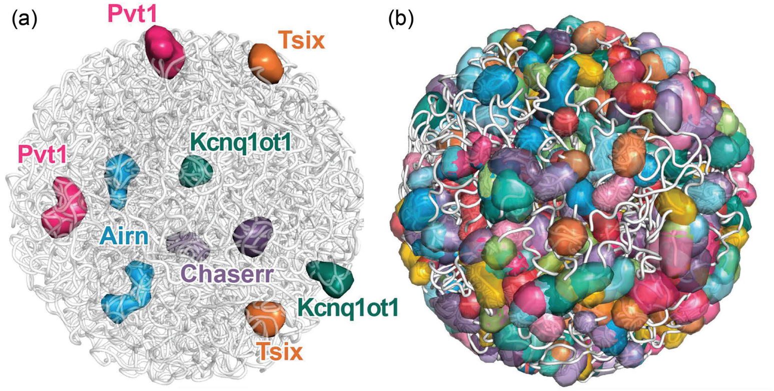

FIGURE 16.5 RNA promotes the formation of spatial compartments in the nucleus. (a) A 3D space filling nuclear structure model of selected lncRNAs. (b) A 3D space filling nuclear structure model of 543 lncRNAs that display at least 50-fold enrichment in the nucleus. Each sphere corresponds to a 1 Mb region or larger where each lncRNA is enriched. (Reproduced from Quinodoz et al.[614] with permission of Elsevier.)

Intrinsically disordered proteins are encoded by the most ancient codons and appear to be the first polypeptides, likely to have functioned initially as chaperones, with catalysis transferred first from RNA to ribonucleoprotein complexes and then to proteins,[554,555,557] a process that may have been interactive.[709]

STRUCTURE-FUNCTION RELATIONSHIPS IN LNCRNAS

The length of lncRNAs varies enormously although in many cases their true length and structure are unknown due to their cell-type specificity and low representation in RNA sequencing datasets. However, high depth sequencing has shown that most are multi-exonic,[710] and some are over 100 kb in length (post splicing), so-called macroRNAs, which have a mean length of 92 kb and are predominantly localized in the nucleus.[711]

Why are lncRNAs so long? The likely answer is that they contain a set of modular domains for binding proteins and guiding them to target sequences in DNA or (other) RNAs[228,712–714] (Figure 16.6).

First, although rapidly evolving (under relaxed structure-function constrains and positive selection for adaptive radiation), lncRNAs exhibit common motifs and motif combinations across vertebrates,[715] and at least 18% of the human genome is conserved at the level of predicted RNA structure.[716] For example, it has been shown that conserved pseudoknots in lncRNA MEG3 are essential for stimulation of the p53 pathway.[717]

Second, similar and potentially paralogous predicted RNA structures occur at many places throughout the genome.[718–720]

Third, lncRNAs are enriched for repeat sequences, which have highly non-random distributions in them.[131,681,721] A notable feature of *Xist*, for example, is that its most highly conserved sequences are the repeat elements, whereas its unique sequences have evolved rapidly,[722] and many of its biological functions, including PRC2 binding, are mediated through its modular repeat elements.[371,379,380,387,397,723]

Many of the lncRNAs referred to earlier have also been shown to be modular, with common features being rapid sequence evolution and structural divergence while retaining related functions, sometimes across large evolutionary distances, and the use of TE-derived sequences as protein-binding domains.[410,721,724,725]

Indeed, TE-derived sequences and tandem repeats participate in many RNA-protein interactions,[34,675,726,727] which leads to the reasonable conclusion that repeat sequences act as RNA-, DNA-, and protein-binding domains that are the essential components of lncRNA function,[721] and that TEs are key building blocks of lncRNAs[397,728] (as well as fulfilling many other modular functions in gene control and gene expression; Chapter 10). Transposition is an efficient means of mobilizing functional cassettes[270] and allowing evolution to explore phenotypic space by modulation of the epigenetic control of developmental trajectories. As pointed out by Neil Brockdorf, “tandem repeat amplification has been exploited to allow orthodox RBPs [RNA binding proteins] to confer new functions for Xist-mediated chromosome inactivation … with potential generality of tandem repeat expansion in the evolution of functional long non-coding RNAs”.[397]

Fourth, many lncRNAs bind chromatin-modifying proteins, transcription factors, nuclear matrix proteins and RNA-binding proteins, in those cases that are well studied, like Xist, roX and HOTAIR, to exert functional consequences.[309,311,342,371,727] It has also been reported that an mRNA can act as a scaffold to assemble adaptor protein assemblies to regulate intracellular transport.[729]

Fifth, chemical probing has shown that lncRNAs, including Xist, physically have a modular structure,[371,374,730,731] and the chemical data matches that predicted by evolutionary conservation of secondary structure, validating both.[374]

Finally, the extensive alternative splicing of lncRNAs strongly imputes a modular structure[710,732–735] and alternative splicing has, unsurprisingly, been shown to alter the function of lncRNAs.[735–738]

If the complex ontogeny of a human requires a large number of guide RNAs then it is not surprising that many have similar protein-binding modules, with variation in repertoire and genomic target specificity, which may only require short stretches of nucleotide complementarity, given the high strength of RNA-RNA and RNA-DNA interactions.[739] These modules may also include enzymes that have adjunct roles in epigenetic transactions: for example, the developmentally regulated lncRNA *H19* binds to and inhibits S-adenosylhomocysteine hydrolase, a feedback inhibitor of DNA methylyltransferases.[740] The alternative splicing of lncRNA exons (which itself must be epigenetically controlled[741]) permits

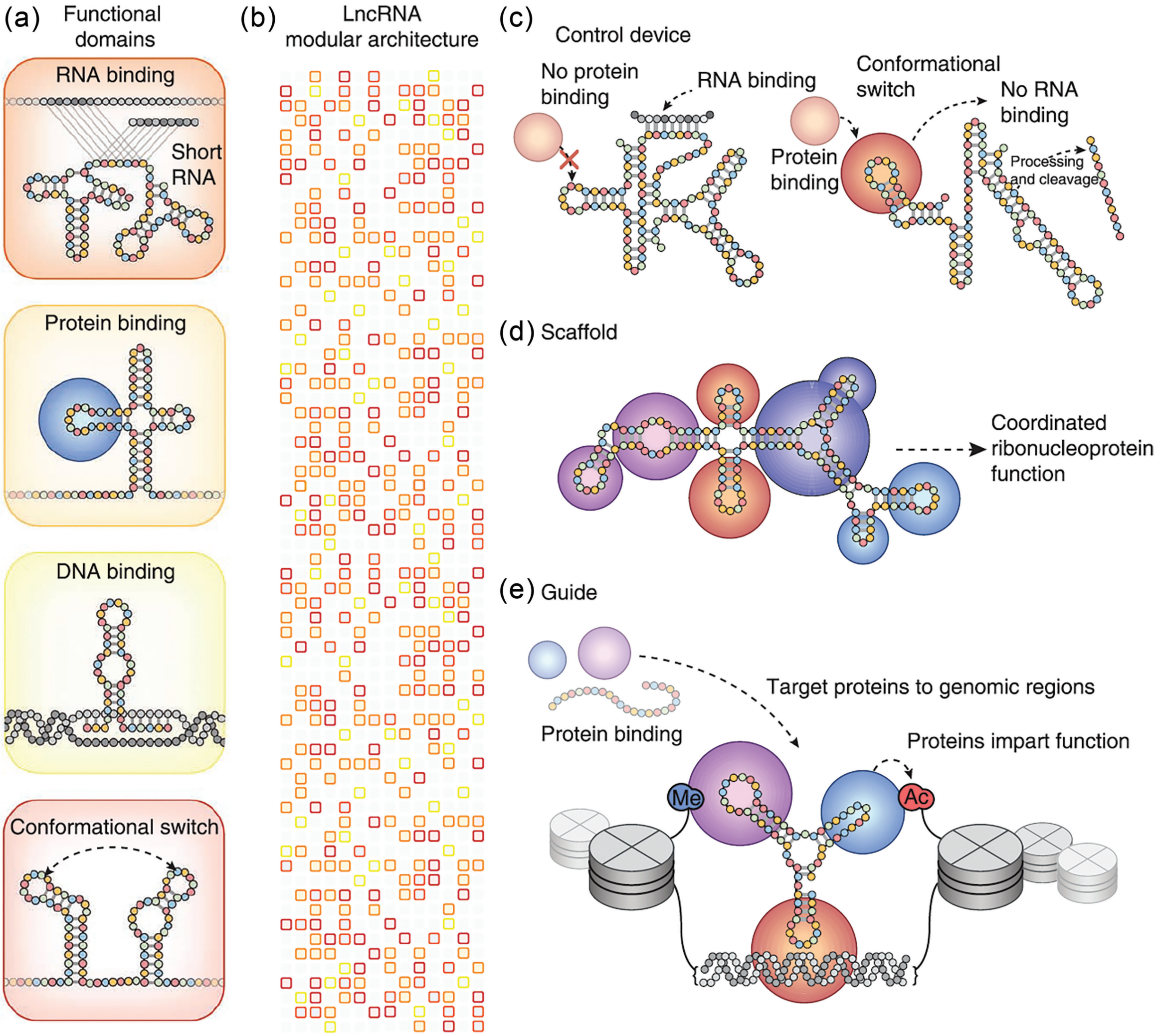

FIGURE 16.6 The modular domain structure and interactions of lncRNAs. (Reproduced from Mercer and Mattick.[714])

selection of specific protein-binding modules and target sites for feed-forward control of protein and regulatory RNA gene expression at many different loci and ultimately cell fate (hold, divide or differentiate) decisions at every stage of developmental ontogeny. There is no cogent model for such fine control by proteins alone.

The challenge now is to determine the repertoire of RNA structures using RNA folding programs, evolutionary conservation, physical and chemical analyses, and machine learning.[374,716,718–720,742–747] The challenge is to determine which RNA structures bind which proteins or which DNA or RNA targets, with a range of techniques becoming available to map RNA localization and interactions,[13,371,518,614,748–755] so that lncRNA biology can be parsed and understood, and thereby construct an expanded Rfam ('RNA family') database,[756] like the Pfam protein domain database[757] that has proved so useful in identifying protein function.

A NEW VIEW OF THE GENOME OF COMPLEX ORGANISMS

It is increasingly evident that lncRNAs, enhancers, topologically associated chromatin domains, transposon-derived sequences and other repeats, chromatin-remodeling and epigenetic information are merging into the same conceptual and mechanistic space.

Imagine the versatility and temporal precision that could be achieved if the locus specificity or local accessibility of proteins that control gene expression during development is guided by stage-specific modular RNAs. The strength of RNA is its potential to address targets through sequence-specific duplex or triplex base-pairing while at the same time recruiting and directing effector proteins to specific genomic locations.

We propose that regulatory RNAs, including the 'repeat' sequences within them, are the evolutionarily and developmentally flexible platforms expressed from the genome in an unfolding symphony to direct and execute the extraordinarily complex decisions required for the precise ontogeny of trillions of differentiated cells in a human, and similarly in other mammals, vertebrates, invertebrates and plants. That is, the epigenetic marks that control development, physiological adaptations and brain function are positioned and controlled by RNAs (among their many other functions in cell biology and gene regulation), and that the proportion of the genome devoted to specifying regulatory and architectural RNAs increases with developmental and cognitive complexity.[758,759]

Small RNAs (miRNAs, sgRNAs, etc.) are simple sequence-specific guides for a single type of effector, such as RISC or CRISPR/Cas. LncRNAs not only have target sequence specificity but are also scaffolds for a range of proteins, notably chromatin-modifying complexes, with both targets and cargoes being four-dimensionally regulated by alternative splicing of lncRNA exons in a feed-forward cascade that directs the next cell fate decision during development. This is a highly efficient system that, like RNAi and CRISPR but in a far more sophisticated and modular manner, directs generic protein effectors to their sites of action.

The misleading historical perspective on the relationship between RNAs and proteins was best expressed by Ewa Grzybowska and colleagues, who concluded: "The current perception of RNA-protein interactions is strongly biased toward a protein-centric approach, in which proteins regulate the expression and activity of RNA, not the other way around."[563]

FURTHER READING

Bah A. and Forman-Kay J.D. (2016) Modulation of intrinsically disordered protein function by post-translational modifications. *Journal of Biological Chemistry* 291: 6696–705.

Cosby R.L., et al. (2021) Recurrent evolution of vertebrate transcription factors by transposase capture. *Science* 371: eabc6405.

Davidovich C., Zheng L., Goodrich K.J. and Cech T.R. (2013) Promiscuous RNA binding by Polycomb repressive complex 2. *Nature Structural & Molecular Biology* 20: 1250–7.

Guttman M. and Rinn J.L. (2012) Modular regulatory principles of large non-coding RNAs. *Nature* 482: 339–46.

Hahn S. (2018) Phase separation, protein disorder, and enhancer function. *Cell* 175: 1723–5.

Kapusta A., et al. (2013) Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLOS Genetics* 9: e1003470.

Kim T.-K., Hemberg M. and Gray J.M. (2015) Enhancer RNAs: A class of long noncoding RNAs synthesized at enhancers. *Cold Spring Harbor Perspectives in Biology* 7: a018622.

Mercer T.R. and Mattick J.S. (2013) Structure and function of long noncoding RNAs in epigenetic regulation. *Nature Structural & Molecular Biology* 20: 300–7.

Niklas K.J., Dunker A.K. and Yruela I. (2018) The evolutionary origins of cell type diversification and the role of intrinsically disordered proteins. *Journal of Experimental Botany* 69: 1437–46.

Polymenidou M. (2018) The RNA face of phase separation. *Science* 360: 859–60.

Quinodoz S.A., et al. (2021) RNA promotes the formation of spatial compartments in the nucleus. *Cell* 184: 5775–90.

Sabari B.R., et al. (2018) Coactivator condensation at super-enhancers links phase separation and gene control. *Science* 361: eaar3958.

Staby L., et al. (2017) Eukaryotic transcription factors: Paradigms of protein intrinsic disorder. *Biochemical Journal* 474: 2509–32.

Trizzino M., et al. (2017) Transposable elements are the primary source of novelty in primate gene regulation. *Genome Research* 27: 1623–33.

Uversky V.N. (2016) Dancing protein clouds: The strange biology and chaotic physics of intrinsically disordered proteins. *Journal of Biological Chemistry* 291: 6681–8.