Review

Special Focus – Genomic Regulation

Long noncoding RNAs and human disease

Orly Wapinski and Howard Y. Chang

Howard Hughes Medical Institute and Program in Epithelial Biology, Stanford, CA 94305, USA

A new class of transcripts, long noncoding RNAs (lncRNAs), has been recently found to be pervasively transcribed in the genome. Multiple lines of evidence increasingly link mutations and dysregulations of lncRNAs to diverse human diseases. Alterations in the primary structure, secondary structure, and expression levels of lncRNAs as well as their cognate RNA-binding proteins underlie diseases ranging from neurodegeneration to cancer. Recent progress suggests that the involvement of lncRNAs in human diseases could be far more prevalent than previously appreciated. We review the evidence linking lncRNAs to diverse human diseases and highlight fundamental concepts in lncRNA biology that still need to be clarified to provide a robust framework for lncRNA genetics.

A wrinkle in the central dogma

The central dogma of molecular biology posits that genetic information is stored in protein-coding genes [1,2]. This hypothesis considered proteins to be the main protagonists of cellular functions, and RNA to be merely an intermediary between DNA sequence and its encoded protein. Most of the previously known noncoding RNAs (ncRNAs) had infrastructural functions, such as ribosomal RNAs. However, recent advances in the field of RNA biology have challenged the assumed role of ncRNAs [3–7].

RNA molecules both encode sequence information and possess great structural plasticity. RNA can directly interact with DNA and with other RNAs by base pairing – either contiguously or bridged by secondary structures – to form a strong duplex, or in special instances a triplex [8]. Highly structured RNA can also provide docking sites for binding proteins [8]. In addition, RNA has a compact size and significant sequence specificity. Owing to its versatility, RNA is an ideal orchestrator of essential biological networks.

Genome-wide surveys have revealed that eukaryotic genomes are extensively transcribed into thousands of long and short ncRNAs [3–7]; this review focuses on long ncRNAs (lncRNAs) – those greater than 200 nt in length. Many of the identified lncRNAs show spatial- and temporal-specific patterns of expression, indicating that lncRNA expression is strongly regulated [9,10]. Evidence of regulation could suggest that numerous lncRNAs have specific biological functions; alternatively, lncRNAs could be byproducts of other regulatory events, such as those that generate open chromatin to allow cryptic transcription. Even in the latter view, lncRNAs are convenient biomarkers of ongoing regulation. Although only a minority have been characterized in detail, lncRNAs participate in diverse biological processes through distinct mechanisms. Generally, lncRNAs have been implicated in gene-regulatory roles, such as chromosome dosage-compensation, imprinting, epigenetic regulation, cell cycle control, nuclear and cytoplasmic trafficking, transcription, translation, splicing, cell differentiation, and others [3,11–14]. It is now becoming evident that ncRNAs are important transcriptional outputs of the genome.

The relevance of ncRNAs in gene regulation has been rapidly unveiling during the last decade. However, the functional elements in the primary sequence of noncoding genes that determine their role as RNA molecules remain unknown. Protein-coding genes have a defined language with a set of grammatical rules. A unique combination of three nucleotides forms a codon, which when read unidirectionally from 5′ to 3′ translates into a specific amino acid, the most basic component of a protein [1]. Aberrations in codons of a protein-coding gene can be interpreted in terms of the amino acids they encode. We can recognize a mutation in a codon and determine its contribution to a given disease. In contrast to the genetic code for protein synthesis, 'the lncRNA alphabet' – a specific set of RNA sequences or structural motifs important for lncRNA function – remains to be elucidated. By analogy to the way in which protein-coding genes have been studied, this review compiles the first lines of evidence for the involvement of small- and large-scale derangements of lncRNA genes in disease. The use of human genetic studies on lncRNAs could help us to understand the regulatory elements of the noncoding language and will allow us to interpret the contribution of those mutations to the pathogenesis of disease.

Over the past decade multiple studies have identified small- and large-scale mutations affecting noncoding regions of the genome, including chromosomal translocations, copy-number alterations, nucleotide expansions, and single nucleotide polymorphisms (SNPs). When such variation occurs outside of protein-coding genes they are often disregarded. Emerging studies are starting to link distinct types of mutations in lncRNA genes with diverse diseases. However, the precise mechanism by which mutations in lncRNAs contribute to the pathogenesis of disease remains a mystery. Here we review the existing evidence for small- and large-scale mutations in the lncRNA primary sequence

Corresponding author: Chang, H.Y. (howchang@stanford.edu).

 0962-8924/$ – see front matter doi:10.1016/j.tcb.2011.04.001

and discuss the putative mechanisms by which the mutations could contribute to the pathogenesis of a disease. Furthermore, in this review we highlight the importance of future studies on lncRNAs in building the framework necessary to interpret the effect of mutations on lncRNA function and their direct connection to disease.

General mechanisms of lncRNA function implicated in disease

LncRNAs participate in a wide-repertoire of biological processes. Almost every step in the life cycle of genes – from transcription to mRNA splicing, RNA decay, and translation – can be influenced by lncRNAs, as shown in the sections below (Figure 1). Focusing on the distinct mechanisms by which lncRNAs regulate gene expression, we emphasize the effects of variation in lncRNA expression and their impact upon disease. We also highlight the significance of disrupted domains and structural motifs that affect the ability of lncRNA to interact with its partner DNA, RNA, and/or protein for proper function, and discuss how such disruptions could contribute to disease.

LncRNAs involved in epigenetic silencing

The *INK4b/ARF/INK4a* locus encodes three tumor suppressor genes that have been linked to various types of cancers. Inhibitor of cyclin kinase 4b (*INK4b*) is also known as p15/cyclin-dependent kinase inhibitor (*CDKN2B*) and encodes the p15 protein. Inhibitor of cyclin kinase 4a (*INK4a*) is also known as p16/cyclin-dependent kinase inhibitor (*CDKN2A*) and encodes the p16 protein. Both

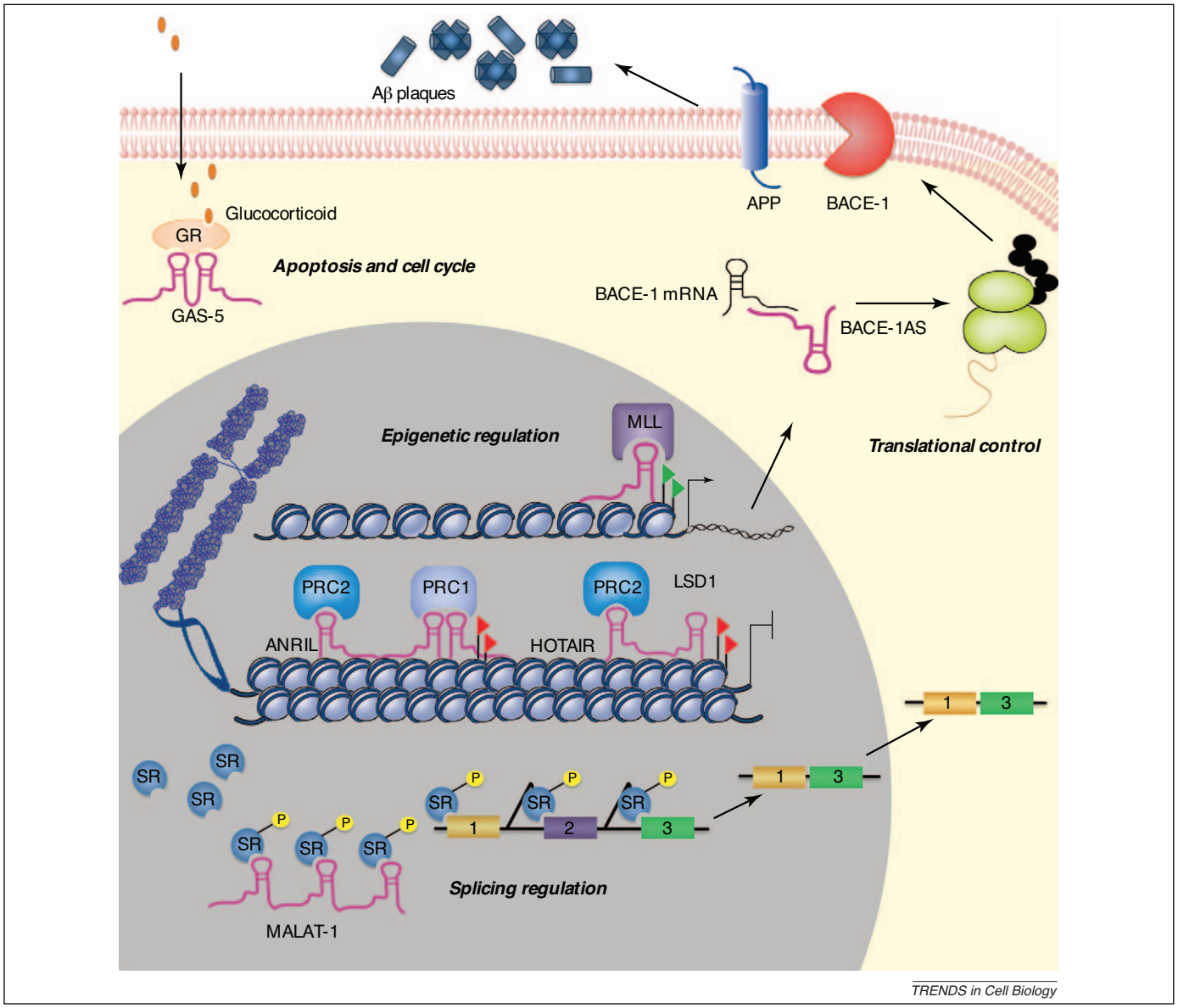

Figure 1. LncRNAs participate in a wide repertoire of biological process. Recent examples of mutated lncRNAs implicated in disease include ANRIL and HOTAIR that bind to chromatin-remodeling complexes PRC1 and PRC2 to alter chromatin and transcription. GAS5 lncRNA acts as a decoy for the GR transcription factor and prevents GR from binding to DNA and transcriptional activation. MALAT1 RNA binds to SR proteins to regulate mRNA alternative splicing, whereas BACE-1AS RNA binds to the complementary BACE-1 mRNA to regulate BACE-1 translation. Red chromatin marks denote transcriptional inhibition. Green chromatin marks denote transcriptional activation. Abbreviations: GR, glucocorticoid receptor; GAS-5, growth arrest-specific 5 ncRNA; Aβ plaques, amyloid-β plaques; APP, amyloid precursor protein; BACE-1, β-site APP-cleaving enzyme; βAPP BACE-1AS mRNA, βAPPBACE-1 antisense ncRNA; MLL, mixed-lineage leukemia; PRC1, polycomb repressive complex 1; PRC2, polycomb repressive complex 2; ANRIL; antisense lncRNA of the *INK4* locus; HOTAIR, HOX antisense ncRNA; MALAT-1, metastasis associated in lung adenocarcinoma transcript; SR, serine/arginine-rich family of nuclear phosphoproteins; P, phosphorylation.

p15 and p16 are involved in cell cycle regulation. Alternative reading frame (ARF) protein participates in the activation of the apoptosis pathway and cell cycle arrest by promoting MDM2 degradation. Within this busy locus is an antisense lncRNA, ANRIL (antisense lncRNA of the *INK4* locus), that spans an estimated region of 30–40 kb [15]. The ANRIL transcript is antisense to *INK4b* and its expression correlates with *INK4a* epigenetic silencing.

A recent study has characterized the mechanism by which the lncRNA ANRIL mediates *INK4a* transcriptional repression in *cis* [16]. ANRIL was shown to interact with the Pc/Chromobox 7 (CBX7) protein, a member of the polycomb repressive complex 1 (PRC1). Because the *INK4b/ARF/INK4a* locus encodes three tumor suppressor genes, their expression must be strongly regulated. Although it remains to be elucidated, altered ANRIL activity might result in dysregulated silencing of the *INK4b/ARF/INK4a* locus, contributing to cancer initiation. Elevated levels of both CBX7 and ANRIL are found in prostate cancer tissues and closely correlate with reduced INK4a levels [16]. Importantly, structural analyses pinpointed the residues in CBX7 required for direct interaction with ANRIL RNA, and point mutations in CBX7 that selectively disrupt RNA binding impaired the ability of PRC1 to repress the *INK4b/ARF/INK4a* locus and restrain cell senescence [16]. Therefore, ANRIL could be an initiating factor in cancer formation by causing abnormal silencing of the *INK4b/ARF/INK4a* locus as postulated by the findings in prostate cancer.

Genome-wide association studies (GWAS) have shown that the intergenic region encompassing ANRIL is significantly associated with increased susceptibility to coronary disease, intracranial aneurysm, type 2 diabetes, as well as several types of cancers [15]. Specific SNPs in and around ANRIL correlate with propensity to develop the diseases listed above [15]. Some of these SNPs directly impact upon enhancer function [17], whereas others also alter the transcription and processing of ANRIL transcripts [18]. Although a direct mechanism underlying the potential role of ANRIL in the disorders listed above is still uncharacterized, this reveals the importance of tightly regulating ANRIL expression and the interaction of ANRIL with CBX7 protein and the target *INK4b/ARF/INK4a* locus.

The lncRNA HOTAIR provides another example of lncRNAs involved in cancer progression by remodeling the chromatin landscape. In breast cancer, increased expression of HOTAIR was reported to correlate with poor prognosis and tumor metastasis [18]. The association of HOTAIR levels with cancer metastasis was described in a cross-sectional study, however, and longitudinal analysis of HOTAIR expression in human cancer progression would provide stronger support for this idea. HOTAIR serves as a modular scaffold by interacting with the polycomb repressive complex 2 (PRC2) and the lysine-specific demethylase 1 (LSD1)–corepressor for element-1-silencing transcription factor (CoREST) complex to silence the *HOXD* loci in *trans* [19]. PRC2 is a histone methyltransferase with activity at H3K27, whereas LSD1 is a histone methyltransferase that recognizes H3K4me3 marks. Using a series of deletion mutants, the domains necessary for HOTAIR interaction with corresponding proteins were mapped to the RNA primary sequence. PRC2 binding mapped to the 5′ end, specifically to the first 300 nt of HOTAIR [19]. By contrast, the LSD1 binding site corresponds to the 3′ end between nt 1500–2146 [19].

Upon increased HOTAIR expression, PRC2 gains chromatin occupancy at novel target sites preventing transcription of several metastasis suppressor genes. Silencing of these metastasis-suppressor genes results in breast cancer metastasis [20]. The link between HOTAIR and metastatic disease depends on the direct interaction between RNA and its protein partner, and the association between RNA and its target DNA sequence. Therefore, altering HOTAIR levels results in enhanced PRC2 repressive activity in an anomalous set of metastasis-suppressor target sites, contributing to breast cancer progression.

ANRIL and HOTAIR act as scaffold molecules by interacting with chromatin modification complexes. In both cases, overexpression of these lncRNAs causes changes to the chromatin landscape that can facilitate cancer initiation and/or progression. The mechanisms by which ANRIL and HOTAIR are altered in disease, whether by primary sequence mutation or other mechanisms, remain to be elucidated.

Splicing regulation by lncRNAs

The lncRNA MALAT-1 (metastasis-associated in lung adenocarcinoma transcript) was identified in an attempt to characterize transcripts associated with early-stage non-small-cell lung cancer (NSCLC) [21]. Two recent studies found that MALAT-1 regulates alternative splicing through its interaction with the serine/arginine-rich (SR) family of nuclear phosphoproteins which are involved in the splicing machinery [22,23]. Because the SR family of proteins affects the alternative splicing patterns of many pre-mRNAs its activity must be tightly regulated. Small changes in SR protein concentration or phosphorylation status can upset the fragile balance that controls mRNA variability between different cells and tissue types [24]. Therefore, the lncRNA MALAT-1 has been suggested to serve as a fine-tuning mechanism to modulate the activity of SR proteins.

MALAT-1 is an abundant ~6.5 kb lncRNA transcribed from chromosome 11q13 and primarily localized in nuclear speckles. MALAT-1 modulates the distribution of pre-mRNA splicing factors to nuclear speckles, and particularly affects the phosphorylation state of SR proteins [23]. In MALAT-1 depleted cells, levels of mislocalized and unphosphorylated SR proteins increase, resulting in a higher number of exon inclusion events [23]. In particular, MALAT-1 is highly abundant in neurons where it regulates synaptogenesis [22] by modulating the activity of neuronal SR splicing factors, thereby regulating the expression of genes involved in synapse formation, density, and maturation [22]. Therefore, MALAT-1 contributes to a broad post-transcriptional gene-regulatory mechanism by coordinating specific mRNA patterning in distinct cell types.

In NSCLC metastasizing tumors, MALAT-1 expression is three-fold higher than in non-metastasizing tumors [21]. Furthermore, in patients with stage I disease, MALAT-1 expression is closely correlated with poor prognosis [21].

Although its function is still unknown, the authors suggest that MALAT-1 expression could be a prognostic marker for metastasis and survival of NSCLC patients [21].

Controlled MALAT-1 function is crucial for correct gene expression. Several lines of evidence have implicated MALAT-1 in distinct diseases, emphasizing the importance of MALAT-1 activity. However, our current understanding of the normal function of MALAT-1 remains incomplete. It is believed that MALAT1 serves as a structural docking site for accumulating specific splicing factors, such as phosphorylated SR proteins, and this is somehow necessary for efficient alternative splicing [23]. Outstanding questions remaining include: what are the domains or secondary structures in the MALAT-1 sequence that are required for its interaction with SR proteins? How does MALAT-1 binding to SR affect its function or phosphorylation state? Is there a motif in the primary sequence of MALAT-1 that determines its localization to nuclear speckles? What is the MALAT-1 mechanism of action that, when dysregulated, contributes to disease? Answering these questions will provide a greater in-depth understanding of the normal role of MALAT-1 and how its dysregulation contributes to the pathogenesis of disease.

Translational control by lncRNAs

The antisense lncRNA β-site amyloid precursor protein (APP)-cleaving enzyme (BACE1-AS) is a conserved RNA encoded by chromosome 11q23.3. BACE1-AS is transcribed from the opposite strand to BACE1, an aspartyl protease that cleaves APP at the β-site and results in the production of amyloid β-peptide (Aβ). Both transcripts have an overlap of ~100 nt that maps to exon 6 of the human protein-coding transcript. Accumulation of the Aβ neuropeptide has been implicated in numerous neurological disorders, which emphasizes the importance of regulating BACE1 catalytic activity [25]. Elevated levels of Aβ, BACE1 proteins, as well as BACE1-AS have been detected in subjects with Alzheimer's disease (AD), suggesting that altered BACE1 expression plays a role in the pathogenesis of the disease [25].

BACE1 enzymatic activity is required for normal brain function. However, its expression is tightly regulated by BACE1-AS, a post-transcriptional regulator of the sense BACE1 mRNA [25]. A study using an RNase protection assay showed that both sense (coding) and anti-sense (noncoding) transcripts directly associate and form a duplex to increase the stability of BACE1 mRNA. The authors proposed a model in which dysregulated BACE1-AS expression sets in motion a feed-forward cascade; AD-related cell stress results in the upregulation of BACE1-AS, which in turn increases BACE1 mRNA stability and protein abundance in the brain [25]. Consequently, elevated BACE1 protein levels result in higher APP processivity and toxic accumulation of Aβ plaques.

BACE1-AS is an example of how dysregulated levels of ncRNA play a significant role in the pathogenesis of AD through hybridization with a sense RNA molecule. However, a closer analysis of the mechanism of action of BACE1-AS is needed to reveal the specific domains required for RNA–RNA interaction, whether complementary or not, and to characterize the secondary structure generated by duplex formation. In human brains from subjects with AD there is a strong correlation between the levels of BACE-1AS and AD severity [25]. However, given the complexity of AD, more detailed studies on the mechanism of action of BACE-1AS are needed.

LncRNAs regulating apoptosis and cell cycle control

LncRNAs can also participate in global cellular behavior by controlling cell growth. The lncRNA growth-arrest-specific 5 (Gas5) sensitizes the cell to apoptosis by regulating the activity of glucocorticoids in response to nutrient starvation [26]. Upon cellular stress induced by limited availability of growth factors, Gas5 ncRNA accumulates through a 5′ oligopyrimidine tract that confers RNA stability under these conditions [27]. Gas5 binds to the DNA-binding domain (DBD) of glucocorticoid receptor (GR) where it acts as a decoy and prevents GR interaction with cognate glucocorticoid response elements (GRE). Under normal conditions, GR target genes are involved in apoptosis suppression, such as cellular inhibitor of apoptosis 2 (cIAP2), and inhibit the cell-death executioners caspases 3, 7, and 9 [28]. However, upon growth arrest, Gas5 activation compromises GR ability to bind to the cIAP2 GRE, reducing cIAP2 expression levels and thereby removing its suppressive effect on caspases [26]. Gas5 function is dependent on its direct association with the GR protein. The interaction has been mapped to the GR DBD and a hairpin structure in the lncRNA Gas5 primary sequence containing GRE-like sequences between nt 539–544 and 553–559 [26].

The *Gas5* gene locus has been linked to increased susceptibility to autoimmune disorders, such as systemic lupus erythematosus in the mouse BXSB strain [26]. Because glucocorticoids are powerful immunosuppressants, increased lncRNA Gas5 activity in immune cells could suppress GR-induced transcriptional activity and contribute to the development of autoimmune disease. The introns of *Gas5* also encode multiple CD box snoRNAs that function in ribosomal RNA biogenesis [27], and this potentially complicates the interpretation of genetic association studies.

Gas5 has also been linked with breast cancer because Gas5 transcript levels are significantly reduced compared to unaffected normal breast epithelia [29]. Therefore, Gas5 could act as a tumor suppressor if reduced levels of Gas5 are unable to maintain sufficient caspase activity to activate an appropriate apoptotic response in disease-compromised cells. Furthermore, chromosomal translocations affecting the 1q25 locus containing the *Gas5* gene have been detected in melanoma, B-cell lymphoma, and prostate and breast cancer [30]. In summary, Gas5 regulates apoptosis and potentially human disease development by acting as a transcription factor decoy for steroid hormone receptors. Whether other transcription factors are also regulated by similar RNA-encoded decoys, termed 'ribo-repressors', should be examined in the future.

Another example of a lncRNA involved in cell cycle control is the long intergenic ncRNA p21 (lincRNA-p21), which was identified in an effort to study lincRNAs regulated by p53 [30]. In response to DNA damage, p53 directly induces the expression of lincRNA-p21, a ~3 kb transcript

located in the proximity of the cell cycle regulator gene, *Cdkn1a*. LincRNA-p21 acts as an inhibitor of the p53-dependent transcriptional response by repressing the transcription of genes that interfere with apoptosis. LincRNA-p21 interacts with ribonucleoprotein K (hnRNP-K) and recruits it to repress a host of genes known to be inhibited by p53 expression. Loss of lincRNA-p21 results in hnRNP-K mislocalization and in the loss of association with the promoter regions of p53-repressed genes. A 780 nt region at the 5′ end of lincRNA-p21 is necessary for interacting with hnRNP-K. The protein-interacting domain of lincRNA-p21 retains sequence conservation and is predicted to form a highly stable structure. However, the factors that determine the targeting of specific loci by lincRNA-p21 are still not understood. Although lincRNA-p21 has not been directly associated with disease, we can speculate that loss of function of lincRNA-p21 could be an important factor contributing to cancer initiation because it functions to trigger cell death through the induction of the apoptosis program.

Human genetics of lncRNAs

LncRNAs work as modular molecules with individual domains [20]. The presence of motifs embedded in the lncRNA primary sequence enables the RNA to specifically associate with DNA, RNA, and/or protein. As described above, misexpression of lncRNAs is linked to numerous diseases. However, emerging studies also reveal the presence of large- and small-scale mutations in the lncRNA primary sequence that are highly correlated with disease.

The bulk of sequence mutations in the genome occur in noncoding and intergenic regions [31]. Because a substantial portion of the genome is transcribed [31], mutations are transmitted to the transcriptome, potentially affecting a large number of lncRNAs. However, it has been challenging to determine the contribution of small mutations in lncRNAs to disease because we have yet to characterize how primary sequence translates into lncRNA function. As was the case with studies of protein-coding genes over the past decade, human genetic studies on lncRNAs could help in deciphering the functional rules of the noncoding language (Figure 2).

At present, the effects of mutations in protein-coding genes can be mechanistically linked to disease pathogenesis (Figure 2). Several types of aberrations can disrupt the coding potential of protein-coding genes and these can be classified based on their magnitude. Large-scale mutations consist of whole-gene deletions and amplifications, and chromosomal translocations. Small-scale mutations involve insertions and deletions of a few nucleotides that can alter the coding reading frame and/or result in neutral, silent, or missense, mutations – defined by the replacement of one encoded amino acid by another – or in nonsense mutations leading to truncation of the translation product.

There are many unanswered questions regarding the functional significance of the lncRNA primary sequence. Are there distinct functional motifs embedded in lncRNA primary sequence? How does the primary sequence translate into secondary-structure motifs? Do individual lncRNA domains have independent functions? Does the orientation of the domains have any functional relevance? What is the role of the linker sequence between distinct domains? Do lncRNAs have a 'reading frame'? In other words, are primary- or secondary-structure motifs in lncRNA only functional if presented sequentially from 5′ to 3′ and with predefined spacing, or would any permutations of the arrangements of the motifs suffice, provided that they are all on a long RNA molecule? Once we have mastered the grammar rules that govern this foreign noncoding language we will be able to understand how its disruption can directly contribute to the pathogenesis of disease.

LncRNAs affected by large-scale mutations

Large-scale mutations encompass major chromosomal rearrangements in genomic regions that encode lncRNAs. There has yet to be a genome-wide study that searches for fragile sites containing genomic lncRNA sequences commonly affected in various types of diseases. However, a few independent studies reviewed below have identified lncRNAs that are affected by individual large-scale mutations. A group of small noncoding RNAs – microRNAs – have been strongly associated with common chromosomal aberrations in human leukemias and carcinomas [32]. As with microRNAs, recurring chromosomal aberrations

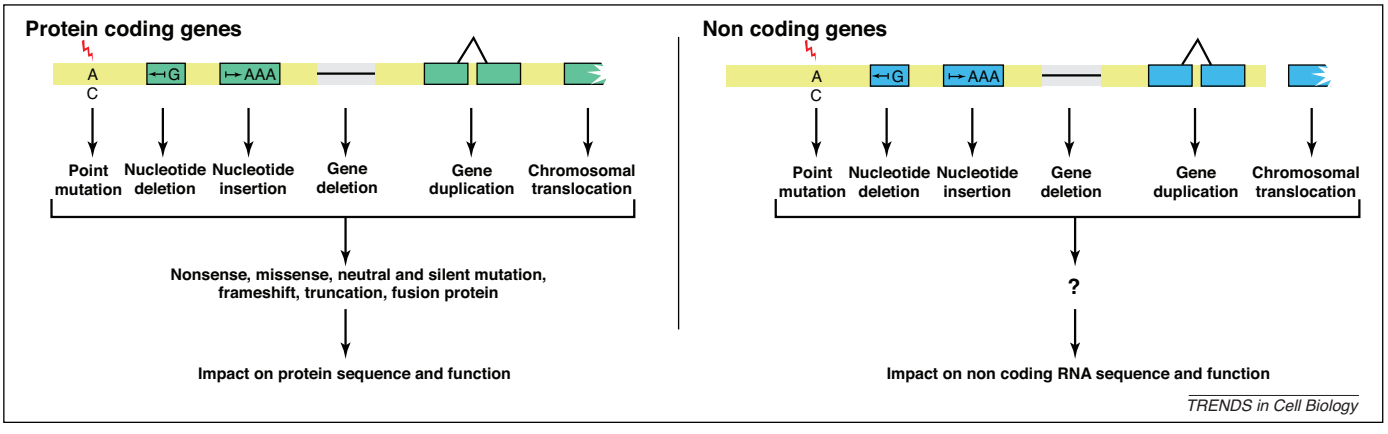

Figure 2. Conceptual framework of the effects of mutations in protein-coding genes compared to those affecting lncRNAs. Although the effects of large-scale rearrangements and single point mutations of protein-coding exons can be predicted rationally based upon the genetic code for protein synthesis, no comparable framework currently exists for lncRNAs. The study of human variations in lncRNA structure and expression, particularly in association with disease, could help us to understand the key functional elements in lncRNAs.

affect the expression of many lncRNAs and could contribute to the development of particular cancers and other diseases [32].

A balanced chromosomal translocation (1;11)(q42.1; q14.3) strongly associated with schizophrenia and other neuropsychiatric disorders in a large Scottish family directly affects two genes on chromosome 1, *DISC1* and *DISC2* (disrupted in schizophrenia 1 and 2) [33]. *DISC1* is a protein-coding gene and *DISC2* encodes an antisense noncoding RNA. Although a link between the *DISC* locus and psychiatric illness has been established, the normal functions of DISC1 and DISC2 remain to be explained. Nevertheless, it is speculated that DISC2 could act as a riboregulator of *DISC1* such that disruption by a chromosomal break could lead to *DISC1* dysregulation. Further studies have identified a large number of SNPs in the *DISC1* genomic sequence that are associated with schizophrenia spectrum disorder, perhaps as a result of disrupted regulation by DISC2 [34,35].

In addition, a large germline deletion (403231 bp) encompassing the *INK4/ARF* locus and the ANRIL lncRNA has been associated with hereditary cutaneous malignant melanoma (CMM) and neural system tumors (NST) syndrome [15]. Based on this large chromosomal deletion, ANRIL was identified as a key player in the development of and hereditary predisposition to cancer [15]. Both of these examples illustrate a potential mechanism by which lncRNAs that are affected by major chromosomal rearrangements have been implicated in disease.

Moreover, microsatellite expansions in the primary sequence of lncRNA genes have been linked to spinocerebellar ataxia type 8 (SCA8) [36–38]. There are two genes transcribed in opposite directions: the protein-coding gene *ATXN8* and the antisense ncRNA gene *ATXN8OS*, which are both affected by a common $(CTG)_n$ expansion. However, the expression of *ATXN8OS* appears to correlate most strongly with the toxic phenotype of the disease, perhaps by affecting the localization and activity of splicing factors. NcRNA transcripts with the trinucleotide expansion accumulate in the nucleus and trigger alternative splicing changes that affect GABA-A transporter 4 (GAT4/Gabt4) expression, resulting in loss of GABAergic inhibition. Although ATXN8OS is clearly associated with the neurodegenerative disorder SCA8, the precise effect of the nucleotide expansion in its function remains to be further studied. The repeat expansion could alter the ATXN8OS 'reading frame' and/or disrupt the formation of a motif that is functionally important for an RNA-binding protein. Therefore, alterations to the primary sequence of the ncRNA ATXN8OS can cause major defects in the cellular behavior of neurodegenerative diseases.

LncRNAs and small-scale mutations

Several lines of evidence suggest that SNPs residing in the key regulatory location of an RNA molecule can severely disrupt its function. Recently, a study examined the structural impact of disease-associated SNPs in the 5′ and 3′ untranslated regions (UTRs) of genes [31]. The algorithm used by this study identified regulatory regions that are structurally affected by SNPs, which are associated with hyperferritinemia cataract syndrome, β-thalassemia, cartilage–hair hypoplasia, retinoblastoma, chronic obstructive pulmonary disease (COPD), and hypertension [31]. The top SNP candidates of this study with a P value <0.1 were present in regulatory regions of RNAs with affected RNA structure, such as open reading frames, protein-binding elements, internal ribosome-entry sites, and others. As illustrated with the ferritin light-chain coding RNA (FTL), four distinct SNPs were able to alter the structure of regulatory elements in its 5′ UTR, resulting in the abrogation of regulatory protein-binding partners [31]. The implications of this study are powerful because they suggest that SNPs could be one of the mechanisms by which disrupted structural motifs in noncoding portions of RNAs can lead to disease [31].

GWAS studies have shown that SNPs in noncoding regions associate with higher susceptibility to diverse diseases. When comparing Finnish subjects with type 2 diabetes and normal glucose-tolerant controls, a large number of SNPs were identified in the *INK4/ARF* loci that were associated with increased risk of type 2 diabetes [39]. The chromosome region in which the SNPs were characterized harbors the protein-coding genes *CDKN2a* (*INK4a*) and *CDKN2b* (*INK4b*). Both these genes are located adjacent to the gene encoding lncRNA ANRIL, and therefore the SNPs could also affect ANRIL. In a separate GWAS study, distinct SNPs were associated with susceptibility to coronary artery disease and atherosclerosis and ANRIL associated with the high-risk haplotype [40]. Further characterization of the identified polymorphisms showed that SNPs could disrupt ANRIL splicing, resulting in a circular transcript that is resistant to RNase R digestion [18]. These novel circularized transcripts affect ANRIL normal function and influence INK4/ARF expression. It is likely that many more lncRNAs are affected by SNPs located in noncoding genomic regions. Indeed, a recent study of leukemias and colorectal cancers identified both germline and somatic mutations in lncRNA genes [41].

The current state of the lncRNA field is principally supported by evidence from changes in lncRNA expression that are associated with disease. However, genetic studies on lncRNA sequence may distinguish the specific contribution of large- and small-scale mutation to lncRNA function. Once our understanding of lncRNA language is clarified, we will be able to classify diseases based on the identified mutations and their effect on lncRNA function.

LncRNA protein-binding partners affected in disease

In addition to lncRNAs themselves, mutations in protein binding partners of lncRNAs have been identified as drivers of diverse disorders, suggesting that these diseases could result from defective ribonucleoprotein (RNP) complexes.

Recently, the development of several different neurodegenerative disorders, including spinocerebellar ataxia (SCA), amyotrophic lateral sclerosis (ALS), fragile X, and others, has been shown to be modulated by RNA-binding proteins [42]. Dysregulated accumulation of misfolded and/or mutated proteins is a common feature of these diseases. In the case of ALS, the RNA and DNA-binding protein TDP-43 was recently reported to harbor multiple mutations, all of which contribute to the

neurodegenerative phenotype [42]. It is currently believed that the mutant TDP-43 proteins are more prone to aggregate, and this might inhibit the normal function of TDP-43 to process RNA.

Similarly, the RNA-binding protein FUS/TLS, known as FUS (fused in sarcoma) or TLS (translocation in liposarcoma) has been recently implicated in ALS as well as in multiple polyglutamine diseases, where it seems to play a role in premature degeneration of motor neurons [43,44]. FUS/TLS has structural similarities to TDP-43, and FUS/TLS also functions in transcription and RNA processing. In surveys of familial ALS cases, a series of dominant missense mutations in FUS/TLS were identified [43,44]. FUS/TLS is a common fusion protein frequently translocated in human cancers but has only recently been associated with neurodegenerative disorders – where RNA-binding proteins are now being discovered to play central roles. Importantly, FUS/TLS participates in transcriptional regulation via lncRNAs [45]. Upon DNA damage, several single-stranded sense and antisense ncRNAs transcribed from the 5′ regulatory regions of the cyclin D1 gene (*CCND1*) associate with FUS/TLS, which subsequently recruits and inhibits the activity of CREB-binding protein (CBP) and p300 histone acetyltransferase on the target gene *CCND1* [45]. For both TDP-43 and FUS/TLS, a key outstanding question is the identity of potential lncRNAs and mRNAs in neurons that are selectively affected by mutant versions of the proteins, and this could shed light on the basis of neurodegeneration.

Another example of an RNA-binding protein mutated in disease is the fragile X mental retardation protein (FMRP) [46]. Mutations of this protein are responsible for fragile X syndrome in which the localization and translation of neuronal mRNAs is defective. Several lines of evidence indicate that FMRP modulates the transport of neuronal transcripts to dendrites for local protein synthesis by associating with the rodent lncRNA BC1 and its primate homolog BC200, which exhibit complementarity to FMRP target mRNAs [47–49]. BC200 RNA levels are significantly upregulated in the brains of human subjects diagnosed with AD [50], and upregulation is accompanied by mislocalization of BC200 RNA to neuronal cell bodies instead of to dendritic spines. Although BC200 misexpression and localization are biomarkers of AD, whether BC200 contributes to AD pathogenesis is not yet clear.

Concluding remarks

The discovery of dysregulated lncRNAs represents a new layer of complexity in the molecular architecture of human disease. However, there are still many gaps in our current understanding of lncRNA function. The triplet nature of the genetic code has been established for protein-coding genes, but the language and regulatory elements of non-coding genes remain a great mystery.

In general, it has been shown that misexpression of lncRNAs contributes to numerous diseases. In addition, several lines of evidence have suggested that even small-scale mutations, such as SNPs, can affect lncRNA structure and function. However, future studies are needed to elucidate the mechanism by which disease-causing mutations in lncRNA functional motifs can affect its regulatory domains and compromise its ability to interact with other molecules, thereby contributing to the pathogenesis of disease. Further study of lncRNA motifs could yield new RNA-based targets for the prevention and treatment of human disease.

Acknowledgments

This work is supported by the National Institutes of Health (R01-CA-118750) and the California Institute for Regenerative Medicine. H.Y.C. is an Early Career Scientist of the Howard Hughes Medical Institute.

References

1 Crick, F.H. *et al.* (1961) General nature of the genetic code for proteins. *Nature* 192, 1227–1232
2 Yanofsky, C. (2007) Establishing the triplet nature of the genetic code. *Cell* 128, 815–818
3 Guttman, M. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458, 223–227
4 Guttman, M. *et al.* (2010) *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* 28, 503–510
5 Marques, A.C. and Ponting, C.P. (2009) Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol.* 10, R124
6 Ponting, C.P. *et al.* (2009) Evolution and functions of long noncoding RNAs. *Cell* 136, 629–641
7 Zhao, J. *et al.* (2010) Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell* 40, 939–953
8 Mattick, J.S. (2003) Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* 25, 930–939
9 Dinger, M.E. *et al.* (2008) Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res.* 18, 1433–1445
10 Mercer, T.R. *et al.* (2008) Specific expression of long noncoding RNAs in the mouse brain. *Proc. Natl. Acad. Sci. U.S.A.* 105, 716–721
11 Mattick, J.S. (2009) The genetic signatures of noncoding RNAs. *PLoS Genet.* 5, e1000459
12 Mattick, J.S. and Makunin, I.V. (2006) Non-coding RNA. *Hum. Mol. Genet.* 15, R17–29
13 Qureshi, I.A. *et al.* (2010) Long non-coding RNAs in nervous system function and disease. *Brain Res.* 1338, 20–35
14 Wilusz, J.E. *et al.* (2009) Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.* 23, 1494–1504
15 Pasmant, E. *et al.* (2010) ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J.* 25, 444–448
16 Yap, K.L. (2010) Molecular Interplay of the non coding RNA ANRIL and methylated histone H3 Lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell* 38, 662–674
17 Harismendy, O. *et al.* (2011) 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature* 470, 264–268
18 Burd, C.E. *et al.* (2010) Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. *PLoS Genet.* 6, e1001233
19 Tsai, M.C. *et al.* (2010) Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 329, 689–693
20 Gupta, R.A. *et al.* (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464, 1071–1076
21 Ji, P. *et al.* (2003) MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* 22, 8031–8041
22 Bernard, D. *et al.* (2010) A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. *EMBO J.* 29, 3082–3093
23 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
24 Long, J.C. and Caceres, J.F. (2009) The SR protein family of splicing factors: master regulators of gene expression. *Biochem. J.* 417, 15–27
25 Faghihi, M.A. *et al.* (2008) Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat. Med.* 14, 723–730

26 Kino, T. *et al.* (2010) Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci. Signal.* 3, ra8
27 Smith, C.M. and Steitz, J.A. (1998) Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5'-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. *Mol. Cell Biol.* 18, 6897–6909
28 Webster, J.C. *et al.* (2002) Dexamethasone and tumor necrosis factor-alpha act together to induce the cellular inhibitor of apoptosis-2 gene and prevent apoptosis in a variety of cell types. *Endocrinology* 143, 3866–3874
29 Huarte, M. *et al.* (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* 142, 409–419
30 Mourtada-Maarabouni, M. *et al.* (2009) GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer. *Oncogene* 28, 195–208
31 Halvorsen, M. *et al.* (2010) Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet.* 6, e1001074
32 Calin, G.A. *et al.* (2007) Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell* 12, 215–229
33 Millar, J.K. *et al.* (2000) Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Hum. Mol. Genet.* 9, 1415–1423
34 Devon, R.S. *et al.* (2001) Identification of polymorphisms within Disrupted in Schizophrenia 1 and Disrupted in Schizophrenia 2, and an investigation of their association with schizophrenia and bipolar affective disorder. *Psychiatr. Genet.* 11, 71–78
35 Ekelund, J. *et al.* (2004) Replication of 1q42 linkage in Finnish schizophrenia pedigrees. *Mol. Psychiatry* 9, 1037–1041
36 Mutsuddi, M. and Rebay, I. (2005) Molecular genetics of spinocerebellar ataxia type 8 (SCA8). *RNA Biol.* 2, 249–252
37 Moseley, M.L. *et al.* (2006) Bidirectional expression of CUG and CAG expansion transcripts and intranuclear polyglutamine inclusions in spinocerebellar ataxia type 8. *Nat. Genet.* 38, 758–769
38 Daughters, R.S. *et al.* (2009) RNA gain-of-function in spinocerebellar ataxia type 8. *PLoS Genet.* 5, e1000600
39 Scott, L.J. *et al.* (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316, 1341–1345
40 Broadbent, H.M. *et al.* (2008) Susceptibility to coronary artery disease and diabetes is encoded by distinct, tightly linked SNPs in the ANRIL locus on chromosome 9p. *Hum. Mol. Genet.* 17, 806–814
41 Wojcik, S.E. *et al.* (2010) Non-coding RNA sequence variations in human chronic lymphocytic leukemia and colorectal cancer. *Carcinogenesis* 31, 208–215
42 Lagier-Tourenne, C. and Cleveland, D.W. (2009) Rethinking ALS: the FUS about TDP-43. *Cell* 136, 1001–1004
43 Kwiatkowski, T.J. *et al.* (2009) Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science* 323, 1205–1208
44 Vance, C. *et al.* (2009) Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. *Science* 323, 1208–1211
45 Wang, X. *et al.* (2008) Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature* 454, 126–130
46 Bagni, C. and Greenough, W.T. (2005) From mRNP trafficking to spine dysmorphogenesis: the roots of fragile X syndrome. *Nat. Rev. Neurosci.* 6, 376–387
47 Zalfa, F. *et al.* (2005) Fragile X mental retardation protein (FMRP) binds specifically to the brain cytoplasmic RNAs BC1/BC200 via a novel RNA-binding motif. *J. Biol. Chem.* 280, 33403–33410
48 Johnson, E.M. *et al.* (2006) Role of Pur alpha in targeting mRNA to sites of translation in hippocampal neuronal dendrites. *J. Neurosci. Res.* 83, 929–943
49 Khanam, T. *et al.* (2006) Poly(A)-binding protein binds to A-rich sequences via RNA-binding domains 1+2 and 3+4. *RNA Biol.* 3, 170–177
50 Mus, E. *et al.* (2007) Dendritic BC200 RNA in aging and in Alzheimer's disease. *Proc. Natl. Acad. Sci. U.S.A.* 104, 10679–10684