Ann. N.Y. Acad. Sci. ISSN 0077-8923

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Issue: *DNA Habitats and Their RNA Inhabitants*

Does every transcript originate from a gene?

Carsten A. Raabe[1] and Jürgen Brosius[1,2]

[1]Institute of Experimental Pathology, ZMBE, University of Münster, Münster, Germany. [2]Institute of Evolutionary and Medical Genomics, Brandenburg Medical School (MHB), Neuruppin, Germany

Address for correspondence: Carsten Raabe, Institute of Experimental Pathology, ZMBE, University of Münster, Von-Esmarch-Str. 56, 48149 Münster, Germany. raabec@uni-muenster.de

Outdated gene definitions favored regions corresponding to mature messenger RNAs, in particular, the open reading frame. In eukaryotes, the intergenic space was widely regarded nonfunctional and devoid of RNA transcription. Original concepts were based on the assumption that RNA expression was restricted to known protein-coding genes and a few so-called structural RNA genes, such as ribosomal RNAs or transfer RNAs. With the discovery of introns and, more recently, sensitive techniques for monitoring genome-wide transcription, this view had to be substantially modified. Tiling microarrays and RNA deep sequencing revealed myriads of transcripts, which cover almost entire genomes. The tremendous complexity of non-protein-coding RNA transcription has to be integrated into novel gene definitions. Despite an ever-growing list of functional RNAs, questions concerning the mass of identified transcripts are under dispute. Here, we examined genome-wide transcription from various angles, including evolutionary considerations, and suggest, in analogy to novel alternative splice variants that do not persist, that the vast majority of transcripts represent raw material for potential, albeit rare, exaptation events.

Keywords: stochastic transcription; evolutionary raw material; exaptation; genomic plasticity; coding definition

No correlation between genome size and organismal complexity

Before high-throughput genome and RNA sequencing became standard in molecular genetic research, the immense contribution of non-protein-coding DNA to mammalian genome architecture remained mostly unnoticed. Common views postulated that entire genomes are assemblies of building blocks of protein-coding genes. Hence, the respective genome sizes would correlate proportionally to the actual number of protein-coding genes.

For a long time, it was suggested that biological complexity would relate to the actual number of protein-coding genes per genome. Genome size would provide accurate approximation of organismal complexity.

However, consensus of what biological complexity might be is hard to establish. Most often, the actual number of different tissues or cell types is considered.[1,2] Therefore, multicellular species display higher complexities than unicellular organisms.[1,2]

Today, the de facto sizes of many genomes are known.[1,3] The genomic data, when analyzed across species, revealed contradicting results. The genome size of many amphibians is extraordinary high and ranks well above that of most mammals. Amoebas, for example, would be assigned higher complexity than humans. With genome sizes in excess of 10 times that of the human genome, these unicellular eukaryotes render the underlying concept doubtful. Species that are likely to display comparable organismal complexities exhibit very different genome sizes;[1] for example, the sizes of the plant genomes of *Arabidopsis thaliana* (~125 Mb) and *Zea mays* (~2.3 Gb) vary by a factor of ~20.[4,5]

In general, results of interspecies comparisons are significantly influenced by polyploidy, experimental error, and non-protein-coding DNA content.[1] However, even when corrected for polyploidy, a general connection[6] between genome size and

doi: 10.1111/nyas.12741

organismal complexity might be debatable and hard to establish.[7]

No correlation between protein gene numbers and organismal complexity

Before the first human genome sequence draft became available, predictions of the actual gene number often ranked near 100,000 protein-coding genes. The results of the human genome sequencing project suggested something entirely different. Even lower estimates of about 50,000 protein-coding genes by far exceed real numbers;[4,8–10] and generous predictions delivered fewer than 30,000 protein-coding genes.[11] Interestingly, the number of human protein-coding genes is in the range of the roundworm *Caenorhabditis elegans*.[11,12] Comparisons of various mammalian genomes supported the initial results.[8,13,14] The intuitive assumption that elevated organismal complexity would relate to increasing numbers of protein-coding genes had to be rejected entirely.[6,12] Furthermore, most protein-coding genes display high degrees of sequence conservation across species. De novo generation of novel mRNA-encoding genes out of the neutral, nonfunctional sequence space is relatively rare. Unlike what was expected, speciation events apparently are not accompanied by bulk de novo generation of species-specific protein genes. However, expansion of gene families, as well as exaptations of novel exons by alternative splicing or co-optation of novel regulatory modules, might play more significant roles for protein diversity and differential expression.[15–18]

Genome-wide transcription and change of the protein-centric view of genome evolution

The idea that the vast majority of intergenic "junk" DNA might provide transcriptionally competent templates for RNA polymerase was difficult to comprehend because the bulk of non-protein-coding DNA was widely regarded to be devoid of function.

Apart from the functional transfer RNAs (tRNAs), small nuclear RNAs (snRNAs), and ribosomal RNAs (rRNAs), the vast majority of intergenic non-protein-coding RNA (npcRNA) remained unrecognized for a long time.[19,20] Even the identification of functional small npcRNAs did not lead to substantial changes of existing paradigms: for many investigators, proteins remained the only relevant molecules.

Later, identification of transcribed genomic regions by tiling microarrays and RNA deep sequencing allowed establishment of entire transcriptional landscapes on open platforms.[19,21,22] The data impressively demonstrated that equation of RNA transcription chiefly with protein-coding RNAs was wrong.[21–24] To date, the distinction between functional RNA and by-products of leaky expression or RNA degradation still remains nontrivial. In any case, the original preconceptions, namely that genomes express RNAs most cost effectively and confine transcription to regions for protein-coding and "structural RNAs,"[a] had to be revised.[21,22,24]

A major consequence of genome-wide RNA expression is the formation of interleaved transcriptional units.[25,26] The borders, which would ideally define the local extension of exons, genes, and other functional modules, are fuzzier than anticipated.[25] Accordingly, pervasive transcription demands refined criteria to connect different transcripts to identical genes.[27] One primary transcript can be processed into multiple mRNAs by alternative splicing, yielding variants of one protein, or, in extreme cases, when the open reading frame (ORF) changes at the beginning of transcript, almost completely

[a]The term "structural RNA" is unfortunate and should be discontinued. It originated during the dark ages when most investigators assumed that, for example, ribosomal RNA (rRNA) had no catalytic function, but merely was the coat hanger or rack for ribosomal proteins (i.e., provided a structure). Also, it was used to distinguish non-protein-coding RNAs from messenger RNAs (mRNAs), as the then known examples of RNAs had intricate structures (transfer RNA and rRNA), while mRNAs by and large were thought to be devoid of structure. Unfortunately, this term persisted up to this day and is used for known non-messenger RNAs, usually those with established function. Ideally, the term RNA should be used for all bona fide RNAs, the term mRNA qualifies those that are templates for translation, and transcript designates all RNA macromolecules including leftovers of RNA processing and degradation. Also, this would dispense of the need for yet another unfortunate term, namely noncoding RNA (ncRNA), as most RNAs bear a code. In order to avoid confusion, we did not implement this terminology throughout the manuscript.

different protein products.[28] Are we still talking about a single gene or are there two different genes residing and even partially overlapping in the same locus? Therefore, the simple criterion of overlapping exons for assignment of different transcripts (encoding products with fundamentally different properties and functions) as a single gene unit is not always straightforward.[26,27] Examples of alternative mRNAs that encode protein variants of different functionalities are known.[29] However, ultra-deep sequencing methods revealed, in addition to functional RNAs, many transcripts that originate from aberrant processing. Often, they share exons with known RNA(s). Accordingly (see above), one has to consider these transcripts as variants of identical genes. Obviously, there is need to separate transcriptional noise or aberrant processing products from functional RNAs, which is difficult to ascertain.

In case of species-specific RNAs, evolutionary analysis usually fails to distinguish novel functional RNAs from transcriptional noise. Low persistence rates further question the functional significance of species-specific RNA candidates and, hence, such transcripts must be considered with caution. Unlike the causal role concepts that were applied by the ENCODE consortium, we questioned whether, for example, identification of transcribed fragments alone is sufficient to indicate function.[30–32] Beyond doubt, transcribed regions might relate to some sort of biochemical mechanism, such as transcriptional interference[33] (see below). However, assigment of biological function requires more, such as—ideally—discernible phenotypes in appropriate knockout (or knockdown) cell or animal models.[34–37]

In summary, alternative protein variants with different functions or even identical proteins with multiple functions (moonlighting) exist.[38] The majority of young transcripts merely represent raw material with the potential to generate evolutionary novelties that, in rare instances, might be recruited (exapted) into a new function.[16,18,39,40] In any event, the distinction between functional genetic novelties and aberrant transcripts is extremely difficult.

The gene definition is growing in complexity

The original gene definition was chiefly based on the structure of bacterial mRNA, and limited to protein-coding genes.[15] In particular, identification of important *cis*-regulatory elements encoded within untranslated regions (UTRs) and *in silico* identification of small npcRNAs lagged behind. This is mainly due to the high diversity of npcRNA-encoded structures. Apart from the entire spectrum of npcRNAs, even promoters and enhancers, which are important regulatory modules for RNA expression, were entirely neglected in the original gene definitions.[15,41]

Introns are parts of genes

The identification of introns ultimately leads to novel gene definitions.[15] The understanding that mature mRNA is a product of primary RNA processing indicated that mRNAs represent stable but derived products of heterogeneous nuclear RNA (hnRNA) transcription. It is intellectually impossible to consider only the final product of RNA processing for gene definition and to exclude large parts of the precursor molecule from which it is derived.[15] Eventually, introns had to be accepted as integral parts of the modern gene concept.[15] In addition, splicing enhancers, which regulate alternative splicing and therefore influence the relative abundance of different protein isoforms, are frequently located within introns and underscore the functional significance of (the originally neglected) non-protein-coding sequence space.[42,43] Alternative splicing might be considered a process that enables the acquisition of initially non-protein-coding intronic sequences into coding DNA sequence (CDS) of mRNAs. Therefore, the strict separation of protein-coding exons and introns is perhaps inappropriate, since this definition fails to account for dynamic exchange between both types of gene modules. In fact, categorical assignments of coding appear to be rather interchangeable, and a priori separation does more harm than good. A prominent example of how junk DNA is converted into novel functional elements is displayed by exonizations of parts of short interspersed repeats (SINEs), such as Alu elements, in primates.[39] Regulated by alternative splicing, non-protein-coding sequences located within introns are potentially exapted as novel exons. Most often, the corresponding new splice variant initially has weak splice sites, and contributes only little to the actual level of gene expression, where the variant is likely neutral or slightly deleterious.[39] However, exceptional

cases where alternative splice variants lead to selective disadvantage have been reported.[39,42] It has to be stressed that this is by no means unique to Alu or any SINE exonization but is rather valid for all hitherto neutrally evolving sequences.

Exaptation events leading to novel mRNA variants by exonization are not irreversible.[16,39] Evolution of novel splice variants is divided into at least two periods of different constraint. Novel exons, shortly after exaptation and incorporation into a preexisting ORF, are mostly devoid of any associated function and subject to loss or positive selection with subsequent purifying/negative selection.[16,39] The second period of exon evolution is likely accompanied by substantially lower mutation rates.[18,39,40] Alternative splicing might potentially increase isoform diversity, but only a minute fraction may persist over time to encode novel or altered functions.[18,39,40] The interplay among exaptation, loss, and persistence illustrates the high flexibility of alternative splicing in novel isoform generation. In cases where alternative splicing would lead to exon combinations that do not provide a suitable ATG start codon to initiate translation, the corresponding transcripts would resemble bona fide candidates for non-protein-coding RNAs. However, RNAs that encode alternative reading frames or even transcripts devoid of any protein-coding capacity are very likely to be involved in different or only marginally related molecular functions. The relevance of alternative transcripts might be best valued when challenged by phylogenetic analysis to delineate signals of selection.[28,30,44] Indeed, increased conservation (negative selection) of sequence elements or alternative ORFs would seemingly argue in support of the hypothetical molecular function.[30] Interestingly, a detailed investigation of mRNAs that harbor dual coding capacity indicated bias toward recent evolution.[28] Therefore, the true functional significance of the overprinted mRNAs is difficult to evaluate.[28,30]

Certainly, the detection of selected function by sequence conservation enriches for relevant candidates and significantly increases the signal-to-noise ratio[30] (see above). Of course, not every regulatory mechanism displays sequence conservation in nucleic acid alignments (see below). In cases where RNA is a rather consequential product of gene regulation, the RNA sequence might not display any conservation. This includes processes that involve promoter occlusion or transcriptional interference (see below).

UTRs and promoters: non-protein-coding regions contribute to genes

Inclusion of non-protein-coding introns in the gene concept also emphasizes the importance of *cis*-acting elements located outside ORF exons.[15] UTRs and upstream promoter elements are considered in current gene definitions. They contain important *cis*-regulatory elements, which enable control of mRNA expression and subcellular localization as well as RNA stability and translation.[15,45–49] Motifs that are located within the UTRs of mRNAs contribute regulatory modules for gene expression. Obviously, the corresponding exons must therefore be included in appropriate gene definitions. As an aside, 5′ UTRs controlled by alternative splicing often might contribute to ORFs and therefore encode not only *cis*-regulatory motifs but also, potentially, functional protein domains.[50]

Enhancers are parts of genes

Unlike the core promoter region, enhancer elements often reside at large distances from the regulated transcript. However, the actual three-dimensional genome organization also depends on chromatin-mediated interactions and might differ remarkably from the linear arrangement of DNA primary sequence. Elements located distantly within the primary sequence space might be closely juxtaposed during transcriptional initiation. Genome structures, as established by various epigenetic regulatory interactions, make important contributions to gene expression. Given the extreme impact of such distal interactions, one must consider enhancer elements as integral parts of genes.[15,51]

Arguably, concepts that include enhancers in existing gene models are complicated by the fact that enhancers often participate in regulation of several genes. But uncoupling of protein-encoded function and gene transcription would be contraindicative.[15] Spatiotemporal control of RNA expression is essential to gene function. Excluding enhancers as gene modules would be as artificial as barring core promoters from genes.[15]

In summary, we would like to put forward that not only the translated or transcribed portions of RNA elements of modern gene definitions but also the entire *cis*-regulatory circuits controlling gene expression must be included in refined

concepts. Certainly, such a widened approach might encounter intellectual inconsistencies when taken to the extreme: *cis*-antisense–encoded npcRNA transcription or even promoter-associated npcRNAs might exert regulatory impact on sense gene expression.[52–55] It is debatable whether or not such entities of riboregulation might have to participate in gene definitions in analogy to enhancers.

Genes are always coding

A valid perquisite of what we call a gene is its inherent coding capacity.[15] "Coding" is utilized synonymously with "functional."[15,56] As the initial analysis of RNA focused on the investigation of mRNAs, the term "coding" became limited to protein-coding only.[15,41,56]

To date, many instances of functional (small and long) non-protein-coding loci, which participate in the regulation of various biological processes, have been reported.[52,57,58] However, because they are devoid of any detectable protein-coding potential, the corresponding RNAs would not fulfill gene definitions. This, however, causes the intellectual dilemma that introns are readily included, but functional RNAs devoid of ORFs would have to be rejected as (part of) a gene.[15] Hence, novel gene definitions also facilitate inclusion of npcRNAs, as already implemented in case of rRNA, tRNA, SRP RNA (signal recognition particle RNA), and others.

What is coding? The semantic approach of Barbieri

Barbieri suggests that, in biology, there are many more organic codes than the best exemplified case of tRNA-encoded functions. Briefly, transfer RNAs are capable of carrying out adaptor functions in translation; they establish linkage between mRNA codons and amino acids, connecting two different molecular entities.[59] Obviously, without adaptor-mediated reactions, mRNA codons would remain entirely meaningless.[59] Barbieri also includes signal transduction and splicing and argues that there are many more organic codes in nature.[59]

Like mRNA codons, words, or for that matter any kind of strings, are per se devoid of meaning. Molecular processes that establish sense in codified reactions are replaced by convention in the case of spoken languages. The adaptor concept enables free evolution of either entity, which is connected via codified reaction. The latter requires that molecular recognitions proceed independently between adaptor and either molecular entity of codified reactions (see below).

Interestingly, even mRNA (alternative) splicing is considered to be a codified reaction.[59,60] The molecular adaptor to ensure coordinated intron removal is the multicomponent spliceosomal machinery.[59] Splice donor and acceptor sites, which represent the two independent molecular entities and are linked through spliceosomal action, unlike translation, reside within the same hnRNA molecule.[59]

Conceptually, even the degeneration of the genetic code is readily incorporated: just as many different words in spoken languages have identical meaning, different codons relate to identical amino acids. On the other hand, mutations of mRNA codons, usually within the first or second positions, generally alter the selected tRNA and lead to malincorporation of amino acids.[59]

Finally, catalyzed reactions are considered different from codified reactions. DNA replication or RNA transcription, for instance, rely on simple polymer matrixes to determine the respective nucleotide to be incorporated. This, however, is different from the bridging-like function of tRNA molecules during translation.[59]

What is coding? The relaxed definition of Trifonov

Trifonov applied more relaxed requirements to define organic codes: any kind of functional sequence motif is regarded to be coding; for instance, rather simple nucleotide strings like miRNA seed sequences or snoRNA antisense boxes, both participating in molecular target recognition, would fully agree with his concept.[15,61,62] The respective motifs are not necessarily organized in the form of linearly encoded sequence strings, but even more complex RNA secondary structures and tertiary interactions within npcRNAs would be consistent with this concept.

Non-protein-coding RNA, noncoding RNA, and non-RNA

Molecular functions exerted by non-protein-coding RNA often involve regulation of translation, RNA stability, and chromatin organization, to name but a few.[52,57,63] Given the astonishing diversity of RNA-encoded functions, general definitions

face the problem of appreciating the abundance of entirely different codes provided by npcRNAs.[41] However, the term non-protein-coding RNA (and, for that matter, even the widespread term noncoding RNA (ncRNA)) does the opposite: instead of trying to provide individual definitions based on actual RNA-encoded functions or to define specific sequence codes, it classifies all RNAs devoid of larger (≥ 100 nt) ORFs as npcRNA.[40,57] On the other hand, the definition suffers from the inherent need to arbitrarily assign a minimal ORF length.[40,57] In particular, products of stochastic transcription (see below) or most transcripts initiated at repetitive elements are certainly devoid of any functional significance.

For example, antisense promoters in LINE1 elements, which transcribe into their loci of integration, have been reported.[64] Such transcripts, devoid of function and selective pressure, might not even be considered as "noncoding" RNAs, but more appropriately as "non-RNAs," despite their chemical RNA identity (see also below). In addition, the function of transcripts from many bidirectional promoters is not evident.[65]

Finally, the term "npcRNA" serves the underlying concept probably better than anticipated. In particular, the alternative terminology "non-peptide-coding" might not withstand the test of time, as examples of long npcRNAs, which at least occasionally provide templates for active peptide synthesis, have been reported.[66] In that case, the corresponding npcRNAs must be correctly designated as mRNAs. The definitions closely link methods for identification of ORFs and npcRNA candidates.[41,58] It might be worth emphasizing that mRNAs almost always harbor both types of RNA-related codes. Often they are regulated by *cis*-regulatory elements (i.e., riboswitches or thermometers, usually located in 5′ UTRs), which exert riboregulation and integrate cellular stimuli. The 3′ UTRs are also crucial for regulation, including RNA stability and/or translation.[49,55]

Interestingly, various data indicate that sometimes even conserved RNA secondary structures reside within regions that encode functional proteins.[67] In agreement with Trifonov's view, these RNA-encoded motifs are ambiguous, as the same sequence encodes various different functions.[68] The overlap among various functional codes might cause higher than average conservation and also suggests that the actual distinction of coding and non-protein-coding RNAs is fuzzier than initially assumed.

NpcRNAs harbor sequence motifs or higher-order structures, which relate to specific RNA-encoded functions. Besides molecular functions exerted by the RNA itself, many npcRNAs are parts of functional ribonucleoprotein particles (RNPs);[59–61] the formation of RNA structural motifs often requires productive RNA–protein interaction.

Conclusively, the entire spectrum of RNA transcription is subdivided into npcRNAs, which collectively comprise class I RNA transcripts and polypeptide-coding class II RNAs. Also, in light of evolutionary transitions, class II RNAs simply could be considered class I RNAs with the added feature of a translatable ORF.[36,68,69]

Class I RNA transcription encompasses various subclasses of npcRNA. The class itself is further subdivided according to size and RNA-processing patterns.[41] For instance, many long npcRNAs are capped and polyadenylated, and even spliced. Various functionally diverse subclasses of small npcRNAs complete the cellular repertoire of class I transcripts.[41]

A third class could be considered, namely macromolecules that chemically are RNA, yet as a product are devoid of any function. These non-RNAs or class III RNAs could be debris from all sorts of RNA processing, such as intron leftovers, spurious nonfunctional intergenic transcripts, or even products from regulatory processes, including transcriptional interference where the act of transcription itself confers function (e.g., blocking downstream promoters), independent of RNA structure. As is the case in most biological classifications, here, too, it is difficult to define clear borders; rather, these three RNA classes should be viewed as a continuum.

Another difficulty is apparent in the controversial definition of RNAs by size only. Initially, for small RNAs, a general size limit under 500 nt was well accepted in eukaryotes.[36,41] The corresponding size fractions would include prominent small RNAs, such as signal-recognition particle (SRP) RNA (300 nt), small nuclear RNA from fraction K (7SK) RNA (332 nt), and mitochondrial RNA-processing (MRP) RNA (287 nt).[36] The lower end of size ranges is commonly represented by miRNAs, initially classified as tiny RNAs.[36,71] Introduced, among others, by the ENCODE consortium,

different size criteria to define long and small RNA became increasingly popular.[72] Here, RNAs smaller than 200 nt would be regarded as small, and RNAs that display sizes above 200 nt are referred to as long. However, in the light of well-accepted small npcRNAs, the newer definitions are unfortunate. Now, numerous small npcRNAs designated as such in many earlier publications have been shifted into the longer size fraction.[36,72]

Pervasive transcription and transcriptional noise

Based on tiling microarrays, the original analysis aimed at deciphering cellular transcription globally delivered strong indications that almost entire genomes are pervasively transcribed.[21,23,25,73] Unexpectedly, even the large intergenic sequence space (i.e., gene deserts) that was initially considered to be devoid of any functional relevance was subject to active transcription. Literally, almost all nucleotides of genomes are incorporated into at least one primary transcript.[74] In addition to dark matter transcripts, which are initiated outside regions of known genic content and for which no molecular function was predictable, multitudes of *cis*-encoded antisense transcripts add to further layers of complexity.[75–77] Ultimately, the de facto transcriptome is twice the size of the genome. However, questions regarding the functional relevance of mass transcription remain.[78–81] Two mouse knockout models covering a total of ~2.35 Mb deletions revealed no apparent phenotype.[82,83] Notably, such a region must produce at least hundreds of transcripts using conservative measures. At least under conventional laboratory conditions, many dark matter transcripts appear to be nonfunctional.

Occasionally, it is emphasized that many of these so-called dark matter transcripts display patterns of tissue-specific expression and consequently should be functional. Merely based on this observation, functionality of the RNA product is implied. However, a recent study mapping transcriptional start sites reports that few promoters are truly housekeeping.[84] There are numerous promoter-trap studies that show cell-type specific or developmentally regulated expression of reporter genes whose products (β-lactamase, firefly luciferase, etc.) have no function in the respective cell or organism whatsoever.[85,86] Furthermore, stochastic transcription does occur,[87] and spurious transcription factor binding underlying tissue-specific transcription without function is at least equally likely as low-level tissue-specific transcription of a functional RNA.

Initial analysis of genome-wide transcription was mainly carried out on tiling microarrays. Hence, various transcripts might rather be attributed to alternative 5′ UTRs or represent products of alternative 3′ polyadenylation. As 5′ UTRs are often subject to alternative splicing, the actual site of transcriptional initiation could be located distant from the actual CDS sequence. Given the frequent incomplete annotation of RNA 5′ termini, signals detected by tiling microarrays might be misconstrued as intergenic dark matter transcripts.

Alternative polyadenylation of 3′ termini is more prominent than assumed earlier, as more than 60% of all known gene products are alternatively polyadenylated.[88] Novel methods based on deep sequencing confirmed the enormous variation of mRNA polyadenylation during differentiation or in response to environmental stimuli.[89] Occasional read-through beyond transcription terminators could further contribute to the phenomenon of genome-wide transcription (dark matter transcripts or non-RNA).[69,83,90] Interestingly, and in contradiction to pervasive RNA dark matter transcription, recent RNA-Seq data revealed that most transcripts are derived from loci that contain known mRNA-encoding genes.[91] The significance of dark matter transcription has been questioned.[20,91–93] However, various issues regarding experimental design are controversial.[91,94] In any case, the debate emphasizes that a clear distinction between functionally relevant entities and noise (class III/non-RNA) remains nontrivial. This is particularly true when low-abundance transcripts are considered. It is now accepted that RNA polymerase is frequently engaged in rather spurious transcriptional processes, while the functional contribution of stochastic transcription remains debatable. Spurious products might generate significant levels of bulk transcripts. However, when monitored at distinct loci, the abundance of individual transcripts is very low.[87] Genome-wide investigations of RNA polymerase occupancies make it possible to monitor transcribed regions *ex vivo*.[87] One would assume that experiments that record RNA

polymerase occupancies allow more accurate measurements of spurious transcription. This might be the case because the chromatin immunoprecipitation/sequencing (ChIP-Seq) analysis is devoid of disturbing influences such as RNA degradation. In addition, it is possible to specifically investigate preinitiation complex formation in ChIP-Seq experiments based on the genome-wide distribution of complexes containing TATA-box binding protein.[87,95,96] This is meaningful, as initiation is usually the rate-limiting step of RNA transcription.[87] Therefore, it might be possible to analyze the extent and sites of spurious transcription more accurately. In any event, the distinction of spurious transcript and functional young npcRNA is difficult.

Structural RNA function and conservation

Selection detectable in interspecies comparisons might be the best criterion for identification of functional sequence elements.[30,97–100] The actual functional domain of RNA often depends on complex secondary and tertiary structures, instead of the primary sequence alone.[41]

Novel RNA identification and validation requires fast and accurate *in silico* RNA structure predictions because experimental procedures to decipher RNA secondary folds are (still) far too time consuming for high-throughput analysis. Dimethyl sulfate treatment in combination with subsequent deep sequencing is a promising approach.[101]

Within living cells, various alternative RNA folds coexist in free equilibrium. The relative abundance of individual RNA secondary structures is directly correlated with their thermodynamic stability. RNAs that populate secondary structures of low free Gibbs energy accumulate in higher relative concentration.

To date, RNA secondary structure predictions based on the calculation of free Gibbs energy are suitable for RNAs up to 700 nt.[102] RNAs generally exist as RNP complexes, and protein–RNA interactions inevitably affect *in vivo* folding. Also, RNA secondary structure interactions form during transcription. Therefore, RNA transcripts might populate conformations that are favored in the more limited sequence space of growing RNA chains.[103] Finally, formation of complex RNA secondary structures is time consuming, and one might speculate that the kinetics of RNA polymerase II–catalyzed transcription exerts influences on RNA folds as well. Epigenetic landscapes indicate that specific chromatin modifications are tightly associated with RNA polymerase pausing.[104] This could suggest that even specific RNA secondary structures are subject to co-transcriptional and epigenetic regulation. In summary, actual RNA secondary structures do not necessarily coincide with the most thermodynamically stable conformer.

Detection of purifying selection is not biased toward RNA conformers that display the lowest free energy. Therefore, procedures that integrate phylogenetic signals for identification of conserved RNA structures are the most powerful. Compensating nucleotide exchanges are indicative of RNA secondary structure conservation.[105] Changes that preserve RNA secondary structure are most easily identified in comparisons between RNA homologs of different species.[b] This permits identification of nucleotides engaged in RNA secondary structure as well as domains that are conserved for other reasons. Single-stranded regions of RNA, such as bulges and loops, usually display higher conservation. Such regions are often involved in complex interactions with other RNAs or proteins. For instance, antisense boxes of snoRNA or transacting bacterial small npcRNAs might display compensatory changes in regions involved in target recognition. Searches for such structural constraints are a reliable tool for identification of potential RNA interaction partners and consequently for prediction of function.

Housekeeping npcRNAs and conservation

In particular, npcRNAs that serve important housekeeping functions display higher levels of primary

[b]Compensatory base changes are defined by concomitant alterations in two RNAs that interact with each other by base pairing, or two regions in a single RNA, for maintenance of a conserved higher-order structure. Changes in target and regulatory RNA (e.g., rRNA and snoRNA, or 3′ UTR and miRNA) are being compensated by, for example, replacing a U–A base pair with a C–G base pair. These base exchanges thus preserve the RNA interactions and help to identify potential target sites. Apart from several computational methods that identify regions involved in base pairing interaction, PAR clip and related techniques provide experimental approaches to isolate target RNA. The experimental approaches are of benefit when evolutionary young, not-yet-conserved interactions are analyzed.[106–108]

sequence and secondary structure conservation. In fact, many of the most conserved sequence elements are located within npcRNA.[109] For instance, human and Baker's yeast ribosomal RNAs exhibit sequence conservation of up to 75%.

Long npcRNAs and conservation

Absence of high levels of conservation cast doubt on the functional relevance of many long npcRNAs that in general reveal weak or no conservation.[58] In addition, long npcRNAs are often transcribed at low levels in comparison to mRNAs.[109,110] Apart from the obvious explanation that long npcRNA candidates are transcriptional noise (non-RNAs), one has to consider that the corresponding transcripts are evolutionary young, or that functions of long npcRNAs do not always require high levels of expression and/or conservation.[83] Well-established regulatory processes such as transcriptional interference or promoter occlusion do not entail conservation of RNA primary sequence or secondary structure.[33,54,111] Rather, it is the process of transcription itself that would be subject to purifying selection. Promoter occlusion was originally identified to exert transcriptional control within tandemly arrayed promoters of bacteria and yeast.[33] Even the involvement of long npcRNAs in transcriptional interference was recently reported.[112,113] Elongating RNA polymerases, initiated at promoter upstream sites, proceed through regulatory elements located within downstream promoters.[33] The act of transcription itself limits the productive interaction of RNA polymerase or transcription factors with DNA templates. Overlapping transcription might impair the rate-limiting step of preinitiation complex formation. Obviously, the impact of promoter occlusion is a direct function of relative promoter strength. Suppression is most effective, if mediated by strong, frequently initiating regulatory upstream promoters.[33] In addition, it is suggested that transcriptional pausing at downstream promoters might increase the impact of promoter occlusion.[114,115]

Obviously, it is difficult to find regions of transcriptional co-regulation via sequence alignments. However, identification of conserved promoters would permit *in silico* enrichment of trustworthy candidates. The corresponding *cis*-acting elements are often too small in size to ensure meaningful analysis. High levels of genome-wide transcription suggest that regulation by promoter occlusion is more prominent than anticipated. Therefore, regulatory mechanisms do not depend on the RNA itself and consequently are not detectable in regions of increased conservation.

Apart from promoter occlusion, various reports emphasize that many promoters initiate transcription in a bidirectional manner. Such transcripts initiating on the opposite strand, away from bona fide genes, are less likely to represent functional RNA entities and are considered to be transcriptional noise–generating non-RNAs. Spurious nonfunctional transcripts or those that are generated for transcriptional interference (see above) are clearly RNAs in a biochemical sense, yet they are without function. Nevertheless, such non-RNAs are raw material for potential, albeit rare, future exaptations.[16]

Conservation and/or function?

It has to be emphasized that conservation is not necessarily synonymous with function: for example, an Mb-sized mouse knockout model, where thousands of multispecies conserved regions were deleted, did not reveal detectable phenotypes.[82,83] Of course, there are many possible explanations why knockout mouse models fail to display visible effects. For example, phylogenetically conserved snoRNA genes were deleted in yeast, but only mild phenotypes were observed.[116] Phenotypical effects might only become obvious after several generations.[117] Alternatively, it is known that genomes often encode functional redundancy. Hence, severe phenotypes might only be detectable in a multiple-knockout model. In addition, screens conducted under laboratory conditions often are inappropriate for the identification of many physiological or behavioral effects.[118] On the other hand, various npcRNAs that display very weak conservation serve important functions. The X chromosome–specific Xist RNA participates in transcriptional inactivation of one X chromosome and compensates for the higher X chromosome–linked gene dosage in female mammals. Although this long RNA (17 kb in humans) is rather weakly conserved, it nonetheless exerts an important molecular function.[119]

Various reports emphasize the importance of widespread *cis*-antisense transcription in bacteria and eukaryotes.[54,76] Mechanistically, this type of regulation might be involved in some form of transcriptional interference.[54,75] The regulatory potential

relies on the process of RNA transcription and does not depend on any RNA structural codes. Alternatively, RNA duplexes, which are the theoretical products of sense and antisense RNA interaction, might represent molecular targets for endonucleolytic processing. The latter mechanism has been demonstrated to act in various bacteria.[54] Ultimately, irrespective of either pathway, the corresponding target RNA is downregulated. Mutational analysis of such interactions is almost impossible, as each change is automatically compensated for on the opposite strand. In addition, the function of the sense RNA might be affected by mutational change. Therefore, phylogenetic analysis also fails to establish functional significance of *cis*-antisense transcripts. Searches for signals of evolutionary conservation might miss a large number of functional RNAs, but certainly this strategy offers strong support for the validation of npcRNA candidates.

Indeed, there are no ultimate criteria that would enable identification of functional RNAs beyond doubt.[83] RNAs might be conserved or display more patchwork-like conservation, as is the case for many functional long npcRNAs (see above). The most appropriate criterion for functional RNA candidates is evolutionary persistence even though it would exclude young, yet functionally significant RNAs.[30] Similar to what has been argued in case of exapted exons, nonfunctional or neutral elements or RNAs are unlikely to persist over time.[39,40] In fact, the vast majority of novel exons or transcripts might be lost over time.[39,40,69,83]

Conflicts of interest

The authors declare no conflicts of interest.

Acknowledgments

This paper is based on introductory chapters of the Ph.D. thesis of Carsten A. Raabe (2014) "Experimental RNomics in Parasites," at the University of Münster. Due to the limit of allowed references, we apologize to those authors whose contributions we failed to cite.

References

1. Taft, R.J., M. Pheasant & J.S. Mattick. 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* **29:** 288–299.
2. Adami, C. 2002. What is complexity? *Bioessays* **24:** 1085–1094.
3. Gregory, T.R. 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol. Rev. Camb. Philos. Soc.* **76:** 65–101.
4. Schnable, P.S. *et al.* 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326:** 1112–1115.
5. Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408:** 796–815.
6. Liu, G., J.S. Mattick & R.J. Taft. 2013. A meta-analysis of the genomic and transcriptomic composition of complex life. *Cell Cycle* **12:** 2061–2072.
7. Palazzo, A.F. & T.R. Gregory. 2014. The case for junk DNA. *PLoS Genet.* **10:** e1004351.
8. Lander, E.S. *et al.* 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
9. Eddy, S.R. 2001. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* **2:** 919–929.
10. Liang, F. *et al.* 2000. Gene index analysis of the human genome estimates approximately 120000 genes. *Nat. Genet.* **25:** 239–240.
11. Claverie, J.M. 2001. Gene number. What if there are only 30000 human genes? *Science* **291:** 1255–1257.
12. Harrison, P.M. *et al.* 2002. A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Res.* **30:** 1083–1090.
13. Brosius, J. 2005. Echoes from the past–are we still in an RNP world? *Cytogenet. Genome. Res.* **110:** 8–24.
14. Waterston, R.H. *et al.* 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
15. Brosius, J. 2009. The fragmented gene. *Ann. N. Y. Acad. Sci.* **1178:** 186–193.
16. Brosius, J. & S.J. Gould. 1992. On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA." *Proc. Natl. Acad. Sci. U. S. A.* **89:** 10706–10710.
17. Brosius, J. 1991. Retroposons–seeds of evolution. *Science* **251:** 753.
18. Schmitz, J. & J. Brosius. 2011. Exonization of transposed elements: a challenge and opportunity for evolution. *Biochimie* **93:** 1928–1934.
19. St Laurent, G., Y. Vyatkin & P. Kapranov. 2014. Dark matter RNA illuminates the puzzle of genome-wide association studies. *BMC Med.* **12:** 97.
20. Ponting, C.P. & T.G. Belgard. 2010. Transcribed dark matter: meaning or myth? *Hum. Mol. Genet.* **19:** R162–R168.
21. Kapranov, P. *et al.* 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296:** 916–919.
22. Bertone, P. *et al.* 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306:** 2242–2246.
23. Cheng, J. *et al.* 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308:** 1149–1154.
24. Kapranov, P. & G. St Laurent. 2012. Dark matter RNA: existence, function, and controversy. *Front. Genet.* **3:** 60.
25. Kapranov, P., A.T. Willingham & T.R. Gingeras, 2007. Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet.* **8:** 413–423.

26. Gingeras, T.R. 2007. Origin of phenotypes: genes and transcripts. *Genome Res.* **17:** 682–690.
27. Gerstein, M.B. *et al.* 2007. What is a gene, post-ENCODE? History and updated definition. *Genome Res.* **17:** 669–681.
28. Liang, H. & L.F. Landweber. 2006. A genome-wide study of dual coding regions in human alternatively spliced genes. *Genome Res.* **16:** 190–196.
29. Lo, W.S. *et al.* 2014. Human tRNA synthetase catalytic nulls with diverse functions. *Science* **345:** 328–332.
30. Graur, D. *et al.* 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol.* **5:** 578–590.
31. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.
32. Stamatoyannopoulos, J.A. 2012. What does our genome encode? *Genome Res.* **22:** 1602–1611.
33. Shearwin, K.E., B.P. Callen & J.B. Egan. 2005. Transcriptional interference: a crash course. *Trends Genet.* **21:** 339–345.
34. Skryabin, B.V. *et al.* 2007. Deletion of the MBII-85 snoRNA gene cluster in mice results in postnatal growth retardation. *PLoS Genet.* **3:** e235.
35. Skryabin, B.V. *et al.* 2003. Neuronal untranslated BC1 RNA: targeted gene elimination in mice. *Mol. Cell. Biol.* **23:** 6435–6441.
36. Brosius, J. 2012. "RNAissance." In *From Nucleic Acids Sequences to Molecular Medicine, RNA Technologies*, V.A. Erdmann & J. Barciszewski, Eds.: 1–18. Berlin: Springer.
37. Sauvageau, M. *et al.* 2013. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* **2:** e01749.
38. Jeffery, C.J. 2003. Moonlighting proteins: old proteins learning new tricks. *Trends Genet.* **19:** 415–417.
39. Krull, M., J. Brosius & J. Schmitz. 2005. Alu-SINE exonization: en route to protein-coding function. *Mol. Biol. Evol.* **22:** 1702–1711.
40. Krull, M. *et al.* 2007. Functional persistence of exonized mammalian-wide interspersed repeat elements (MIRs). *Genome Res.* **17:** 1139–1145.
41. Brosius, J. & H. Tiedge. 2004. RNomenclature. *RNA Biol.* **1:** 81–83.
42. Wang, Y. *et al.* 2012. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nat. Struct. Mol. Biol.* **19:** 1044–1052.
43. Mattick, J.S. 1994. Introns: evolution and function. *Curr. Opin. Genet. Dev.* **4:** 823–831.
44. Baertsch, R. *et al.* 2008. Retrocopy contributions to the evolution of the human genome. *BMC Genomics* **9:** 466.
45. Araujo, P.R. *et al.* 2012. Before it gets started: regulating translation at the 5′ UTR. *Comp. Funct. Genomics* **2012:** 475731.
46. Chatterjee, S. & J.K. Pal. 2009. Role of 5′- and 3′-untranslated regions of mRNAs in human diseases. *Biol. Cell* **101:** 251–262.
47. Mignone, F. *et al.* 2002. Untranslated regions of mRNAs. *Genome Biol.* **3:** REVIEWS0004.
48. Pesole, G. *et al.* 2000. The untranslated regions of eukaryotic mRNAs: structure, function, evolution and bioinformatic tools for their analysis. *Brief Bioinform.* **1:** 236–249.
49. Zhao, W. *et al.* 2014. Massively parallel functional annotation of 3′ untranslated regions. *Nat. Biotechnol.* **32:** 387–391.
50. Singer, S.S. *et al.* 2004. From "junk" to gene: curriculum vitae of a primate receptor isoform gene. *J. Mol. Biol.* **341:** 883–886.
51. Maston, G.A., S.K. Evans & M.R. Green. 2006. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* **7:** 29–59.
52. Su, W.Y., H. Xiong & J.Y. Fang. 2010. Natural antisense transcripts regulate gene expression in an epigenetic manner. *Biochem. Biophys. Res. Commun.* **396:** 177–181.
53. Taft, R.J. *et al.* 2009. Evolution, biogenesis and function of promoter-associated RNAs. *Cell Cycle* **8:** 2332–2338.
54. Georg, J. & W.R. Hess. 2011. cis-antisense RNA, another level of gene regulation in bacteria. *Microbiol. Mol. Biol. Rev.* **75:** 286–300.
55. Fabian, M.R., N. Sonenberg & W. Filipowicz. 2010. Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.* **79:** 351–379.
56. Trifonov, E.N. 2011. Thirty years of multiple sequence codes. *Genomics Proteomics Bioinformatics.* **9:** 1–6.
57. Maxwell, E.S. & M.J. Fournier. 1995. The small nucleolar RNAs. *Annu. Rev. Biochem.* **64:** 897–934.
58. Ulitsky, I. & D.P. Bartel. 2013. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154:** 26–46.
59. Barbieri, M. 2003. *THE ORGANIC CODES, An Introduction to Semantic Biology*. Cambridge: Cambridge University Press.
60. Black, D.L. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72:** 291–336.
61. Trifonov, E.N. 1989. The multiple codes of nucleotide sequences. *Bull. Math. Biol.* **51:** 417–432.
62. Tollervey, D. & T. Kiss. 1997. Function and synthesis of small nucleolar RNAs. *Curr. Opin. Cell. Biol.* **9:** 337–342.
63. Lin, D. *et al.* 2008. Translational control by a small RNA: dendritic BC1 RNA targets the eukaryotic initiation factor 4A helicase mechanism. *Mol. Cell. Biol.* **28:** 3008–3019.
64. Speek, M. 2001. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol. Cell. Biol.* **21:** 1973–1985.
65. Neil, H. *et al.* 2009. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* **457:** 1038–1042.
66. Banfai, B. *et al.* 2012. Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.* **22:** 1646–1657.
67. Chen, H. & M. Blanchette. 2007. Detecting non-coding selective pressure in coding regions. *BMC Evol. Biol.* **7**(Suppl 1): S9.
68. Trifonov, E.N., Z. Volkovich & Z.M. Frenkel. 2012. Multiple levels of meaning in DNA sequences, and one more. *Ann. N. Y. Acad. Sci.* **1267:** 35–38.

69. Brosius, J. 2014. The persistent contributions of RNA to eukaryotic gen(om)e architecture and cellular function. *Cold Spring Harb. Perspect. Biol.* **6:** a016089.
70. Brosius, J. 2001. tRNAs in the spotlight during protein biosynthesis. *Trends Biochem. Sci.* **26:** 653–656.
71. Ruvkun, G. 2001. Molecular biology. Glimpses of a tiny RNA world. *Science* **294:** 797–799.
72. Djebali, S. *et al.* 2012. Landscape of transcription in human cells. *Nature* **489:** 101–108.
73. Jensen, T.H., A. Jacquier & D. Libri. 2013. Dealing with pervasive transcription. *Mol. Cell.* **52:** 473–484.
74. Birney, E. *et al.* 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447:** 799–816.
75. Osato, N. *et al.* 2007. Transcriptional interferences in cis natural antisense transcripts of humans and mice. *Genetics* **176:** 1299–1306.
76. Vanhee-Brossollet, C. & C. Vaquero. 1998. Do natural antisense transcripts make sense in eukaryotes? *Gene* **211:** 1–9.
77. Dahary, D., O. Elroy-Stein & R. Sorek. 2005. Naturally occurring antisense: transcriptional leakage or real overlap? *Genome Res.* **15:** 364–368.
78. Louro, R., A.S. Smirnova & S. Verjovski-Almeida. 2009. Long intronic noncoding RNA transcription: expression noise or expression choice? *Genomics* **93:** 291–298.
79. Baker, M. 2010. MicroRNA profiling: separating signal from noise. *Nat. Methods* **7:** 687–692.
80. Polev, D. 2012. Transcriptional noise as a driver of gene evolution. *J. Theor. Biol.* **293:** 27–33.
81. Vickers, K.C. *et al.* 2015. Mining diverse small RNA species in the deep transcriptome. *Trend Biochem. Sci.* **40:** 4–7.
82. Nobrega, M.A. *et al.* 2004. Megabase deletions of gene deserts result in viable mice. *Nature* **431:** 988–993.
83. Brosius, J. 2005. Waste not, want not–transcript excess in multicellular eukaryotes. *Trends Genet.* **21:** 287–288.
84. Forrest, A.R. *et al.* 2014. A promoter-level mammalian expression atlas. *Nature* **507:** 462–470.
85. Friedrich, G. & P. Soriano. 1991. Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes Dev.* **5:** 1513–1523.
86. Stanford, W.L., J.B. Cohn & S.P. Cordes. 2001. Gene-trap mutagenesis: past, present and beyond. *Nat. Rev. Genet.* **2:** 756–768.
87. Struhl, K. 2007. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.* **14:** 103–105.
88. Derti, A. *et al.* 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res.* **22:** 1173–1183.
89. Wilkening, S. *et al.* 2013. An efficient method for genome-wide polyadenylation site mapping and RNA quantification. *Nucleic Acids Res.* **41:** e65.
90. Richard, P. & J.L. Manley. 2009. Transcription termination by nuclear RNA polymerases. *Genes Dev.* **23:** 1247–1269.
91. vanBakel, H. *et al.* 2010. Most "dark matter" transcripts are associated with known genes. *PLoS Biol.* **8:** e1000371.
92. Robinson, R. 2010. Dark matter transcripts: sound and fury, signifying nothing? *PLoS Biol.* **8:** e1000370.
93. Flintoft, L. 2010. Transcriptomics: throwing light on dark matter. *Nat. Rev. Genet.* **11:** 455.
94. Clark, M.B. *et al.* 2011. The reality of pervasive transcription. *PLoS Biol.* **9:** e1000625; discussion e1001102.
95. Kuras, L. & K. Struhl. 1999. Binding of TBP to promoters in vivo is stimulated by activators and requires Pol II holoenzyme. *Nature* **399:** 609–613.
96. Li, X.Y. *et al.* 1999. Enhancement of TBP binding by activators and general transcription factors. *Nature* **399:** 605–609.
97. Necsulea, A. *et al.* 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505:** 635–640.
98. Gerstein, M.B. *et al.* 2014. Comparative analysis of the transcriptome across distant species. *Nature* **512:** 445–448.
99. Ho, J.W. *et al.* 2014. Comparative analysis of metazoan chromatin organization. *Nature* **512:** 449–452.
100. Boyle, A.P. *et al.* 2014. Comparative analysis of regulatory information and circuits across distant species. *Nature* **512:** 453–456.
101. Rouskin, S. *et al.* 2014. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505:** 701–705.
102. Mathews, D.H., W.N. Moss & D.H. Turner. 2010. Folding and finding RNA secondary structure. *Cold Spring Harb. Perspect. Biol.* **2:** a003665.
103. Pan, T. & T. Sosnick. 2006. RNA folding during transcription. *Annu. Rev. Biophys. Biomol. Struct.* **35:** 161–175.
104. Gilchrist, D.A. & K. Adelman. 2012. Coupling polymerase pausing and chromatin landscapes for precise regulation of transcription. *Biochim. Biophys. Acta* **1819:** 700–706.
105. Woese, C.R. *et al.* 1980. Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Res.* **8:** 2275–2293.
106. Hafner, M. *et al.* 2012. Genome-wide identification of miRNA targets by PAR-CLIP. *Methods* **58:** 94–105.
107. Hafner, M. *et al.* 2010. PAR-CliP: a method to identify transcriptome-wide the binding sites of RNA binding proteins. *J. Vis. Exp.* pii: 2034.
108. Jaskiewicz, L. *et al.* 2012. Argonaute CLIP–a method to identify in vivo targets of miRNAs. *Methods* **58:** 106–112.
109. Dinger, M.E. *et al.* 2009. Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. *Brief Funct. Genomic Proteomic* **8:** 407–423.
110. 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.
111. Martens, J.A., L. Laprade & F. Winston. 2004. Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. *Nature* **429:** 571–574.
112. Ard, R., P. Tong & R.C. Allshire. 2014. Long non-coding RNA-mediated transcriptional interference of a permease gene confers drug tolerance in fission yeast. *Nat. Commun.* **5:** 5576.
113. Kornienko, A.E. *et al.* 2013. Gene regulation by the act of long non-coding RNA transcription. *BMC Biol.* **11:** 59.
114. Palmer, A.C., J.B. Egan & K.E. Shearwin. 2011. Transcriptional interference by RNA polymerase pausing and dislodgement of transcription factors. *Transcription* **2:** 9–14.

115. Palmer, A.C. *et al.* 2009. Potent transcriptional interference by pausing of RNA polymerases over a downstream promoter. *Mol. Cell* **34:** 545–555.
116. Thompson, J.R. *et al.* 1988. Sequence and genetic analysis of a dispensible 189 nucleotide snRNA from Saccharomyces cerevisiae. *Nucleic Acids Res.* **16:** 5587–5601.
117. Badis, G., M. Fromont-Racine & A. Jacquier. 2003. A snoRNA that guides the two most conserved pseudouridine modifications within rRNA confers a growth advantage in yeast. *RNA* **9:** 771–779.
118. Lewejohann, L. *et al.* 2004. Role of a neuronal small non-messenger RNA: behavioural alterations in BC1 RNA-deleted mice. *Behav. Brain Res.* **154:** 273–289.
119. Gendrel, A.V. & E. Heard. 2014. Noncoding RNAs and epigenetic mechanisms during X-chromosome inactivation. *Annu. Rev. Cell Dev. Biol.* **30:** 561–580.