APPLICATIONS OF NEXT-GENERATION SEQUENCING

The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs

Alain Jacquier

Abstract | Over the past few years, techniques have been developed that have allowed the study of transcriptomes without bias from previous genome annotations, which has led to the discovery of a plethora of unexpected RNAs that have no obvious coding capacities. There are many different kinds of products that are generated by this pervasive transcription; this Review focuses on small non-coding RNAs (ncRNAs) that have been found to be associated with promoters in eukaryotes from animals to yeast. After comparing the different classes of such ncRNAs described in various studies, the Review discusses how the models proposed for their origins and their possible functions challenge previous views of the basic transcription process and its regulation.

Small nuclear RNAs
Small RNAs that are involved in precursor mRNA processing.

Small nucleolar RNAs
The functions of these RNAs include RNA cleavage reactions, as well as specifying sites of ribose methylation and pseudouridylation.

Small interfering RNAs
Small antisense RNAs (20–25 nucleotides long) that are generated from specific dsRNAs that trigger RNA interference. They serve as guides for the cleavage of homologous mRNA by the RNA-induced silencing complex.

After entire eukaryotic genomes had been sequenced, techniques that aimed to determine the complete catalogue of transcribed sequences and how they are regulated were among the first large-scale functional genomic approaches that were developed. Until recently, the description of a transcriptome — the entire set of transcripts in a cell — was essentially limited to the characterization of the transcription products of known annotated genes. These products were mainly mRNAs and known stable non-coding RNAs (ncRNAs), such as tRNAs, small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs). However, unexpected levels of complexity then began to emerge, firstly with the discovery of naturally occurring interfering RNAs, such as small interfering RNAs (siRNAs) and microRNAs (for a recent review, see REF. 1). But this was only the tip of the iceberg. The development of high-resolution tiled arrays and, more recently, RNA deep-sequencing and chromatin immunoprecipitation (from which the patterns of chromatin modification give independent clues of transcribed sequences[2]) allowed transcriptome characterization that was not biased by previous genome annotations. These new technologies revealed that the transcription landscape in higher eukaryotes is much more complex than had been anticipated, with a high proportion of transcripts originating from intergenic regions that were previously thought to be silent or in antisense to genes. Transcription that does not map to genes has also been found in yeast. The unanticipated level of complexity has led to the notion of 'pervasive' transcription, which refers to the fact that the transcripts are not restricted to well-defined functional features, such as genes.

Different names have been used in different studies to differentiate the diverse products of pervasive transcription, and these products often represent overlapping populations of transcripts, which adds further complexity. Nonetheless, recurrent patterns of pervasive transcription are beginning to emerge. For example, long ncRNAs (lncRNAs), which can be either intergenic or antisense to genes, are now distinguished from shorter heterogeneous transcripts that have recently been discovered and that cluster at the ends of genes, particularly around the promoter regions. The functional significance of lncRNAs has been discussed extensively in various reviews[3–6]. This Review focuses on the newly identified shorter transcripts, in particular those associated with gene promoters. After discussing and comparing the different types that have been described, I discuss possible mechanisms that have been proposed for their origin. I discuss how the cellular machinery can discriminate between these transcripts and target some for degradation. Finally, I discuss the functional implications of these recent findings for the understanding of fundamental aspects of the transcription process and its regulation.

Unité de Génétique des Interactions Macromoléculaires, Institut Pasteur, Centre National de la Recherche Scientifique URA2171, 25 Rue du Dr Roux, F-75015, Paris, France.
e-mail:
alain.jacquier@pasteur.fr
doi:10.1038/nrg2683

MicroRNA
A form of single stranded RNA typically 20–25 nucleotides long that is thought to regulate the expression of other genes, either by inhibiting protein translation or degrading a target mRNA transcript through a process that is similar to RNA interference.

5′ cap
Eukaryotic mRNA is modified by the addition of an m7G(5′)ppp(5′)N structure at the 5′ terminus. Capping is essential for several important steps of gene expression; for example, mRNA stabilization, splicing, mRNA export from the nucleus and translation initiation.

Progress in identifying novel transcripts

Advances in understanding the complexity of eukaryotic transcriptomes have been driven by methodological breakthroughs. To set the scene for our current picture of a landscape of pervasive transcription, here I introduce the main steps that have been made towards unbiased transcriptomic studies (see BOX 1 for a summary of techniques). The serial analysis of gene expression (SAGE) approach in yeast was probably the first attempt to analyse a transcriptome using an unbiased (that is, annotation independent) method[7]. This pioneering work, which was made possible by the completion of the yeast genome[8], revealed a number of sequence tags that did not match any annotated feature in the genome. At least one-tenth of the intergenic sequences were estimated to exhibit some transcriptional activity. However, because techniques at that time did not allow deep sequencing of the tags, no specific pattern could be recognized for these putative transcripts, and it was not clear whether they represented random transcriptional noise or some degree of experimental noise.

This initial observation in yeast was paralleled by a number of studies in higher eukaryotes that reported a large proportion of transcripts that did not correspond to protein-coding genes. Pioneering genome-scale studies were performed in mice in the RIKEN Mouse Gene Encyclopedia project, which characterized full-length oligo(dT)-primed cDNAs[9,10] (the Functional Annotation of Mouse 3 (FANTOM3) project has characterized more than 43,000 transcription units[11]), and in humans by the use of tiling DNA microarrays[12]. Unexpectedly, these studies revealed that almost half of the poly(A)-tailed RNAs detected were non-protein-coding transcripts that did not match any annotated sequences. These initial observations were originally met with some scepticism, but subsequent independent analyses have confirmed them[13,14].

Using independent experimental approaches, a number of studies have now confirmed the general validity of the concept that in higher eukaryotes the amount of sequence that is transcribed is much greater than would be expected from protein-coding gene repertoires. These studies have used, on a large scale, techniques such as full-length cDNA cloning and sequencing, tiling arrays, determination of sequence tags associated with the RNA 5′ cap structure (the cap analysis of gene expression (CAGE) technique; BOX 1) and/or with RNA 3′ ends, and deep sequencing of RNAs (RNA–seq)[15–17]. Genome-wide techniques have also been developed to test some of the most controversial examples of pervasive transcription. For example, a technique known as asymmetric strand-specific analysis of gene expression (ASSAGE; BOX 1) was developed to assign RNA strandedness unambiguously. This enabled unequivocal description of the antisense transcriptome of human cells; 11% of the tags within an annotated sequence were found in antisense[18]. Also, the specific chromatin signature associated with elongating RNA polymerase II (RNAPII) provided an independent means of identifying long interspersed ncRNAs (lincRNAs)[2]. For higher eukaryotes, a number of recent reviews[19–23] have described the picture of the transcriptome created by these and other studies.

Transcription at gene boundaries in animals

Relatively small RNAs that are substantially enriched at gene boundaries are a prominent novel type of transcript. They were initially revealed by high density tiling arrays[24], and two categories of this novel type have been defined: molecules in the 20 to 200 nucleotide range are called ‘small RNAs’ (sRNAs) and molecules from 200 nucleotides to greater than 1 kb are called ‘long RNAs’ (lRNAs). Note that the distinction of two separate classes is somewhat arbitrary and might reflect technical bias rather than a true biological bimodal length distribution. sRNAs can comprise 10% of the transcription detected in human cells. The sRNAs that cluster at promoters have been called ‘promoter-associated sRNAs’ (PASRs) and those that cluster at the 3′ ends of genes have been called ‘terminator-associated sRNAs’ (TASRs) (see BOX 2 for a summary of the types of transcript discussed).

In animals, a number of recent articles have reported pervasive transcription of sRNAs around promoters, similar to PASR transcription. Studies using independent

Box 1 | Summary of different techniques used for transcriptome analyses

Serial analysis of gene expression (SAGE)
The first technique that was used to analyse transcriptomes in a manner unbiased by annotations. Small cDNA tags (generated by type II restriction enzymes) are concentrated to speed up sequencing[7].

Cap analysis of gene expression (CAGE)
This technique is used to sequence small cDNA tags (similar to SAGE tags) that originate from the capped 5′ end of transcripts[75].

3′ LongSAGE
This is used to determine small cDNA tags (similar to SAGE tags) that originate from the 3′ end of transcripts[42,76].

RNA–seq
A generic term for high-throughput sequencing of cDNAs. There are several variants of the technique that differ by the type of sequencing technologies used and by the way the sequencing primers are added[17,25,35]. The choice of technology has an impact on the length of the reads and on biases relating to the length of the PCR templates being used. The main choice regarding primers is whether they are added before or after cDNA synthesis; this has an impact on tag representation and strandedness determination[17,25,35].

Asymmetric strand-specific analysis of gene expression (ASSAGE)
A variation of RNA–seq in which the RNA has been modified with bisulphite. This changes all Cs to Us, which allows unambiguous strand determination after sequencing[18]. Because the treatment is performed before cDNA synthesis, this approach eliminates artefacts, such as spurious synthesis of second-strand cDNA[77].

Global run-on sequencing (GRO–seq)
This technique generates cDNA tags extended from nascent transcripts synthesized *in vitro* from isolated human nuclei. It allows the mapping of elongating RNA polymerase II[27].

Tiling arrays
In this technique, cDNA probes are hybridized to DNA microarrays that carry overlapping oligonucleotides that cover the complete genome (or a fraction of a genome). This methodology can confer resolution of a few nucleotides[16].

Chromatin immunoprecipitation (ChIP)
ChIP using antibodies against specific histone modifications reveals modification patterns that are characteristic of promoters[33]. This approach indirectly revealed many unknown non-coding RNAs[2].

Box 2 | Names for non-coding RNAs and their definitions

These are the main types of non-coding RNAs (ncRNAs) discussed in this Review (it is not a comprehensive list of all non-coding RNAs).

TUFs
A generic name for transcripts of unknown function[78].

Small RNAs (sRNAs)
According to REF. 24, sRNAs are defined as any ncRNAs <200 nucleotides.

Long RNAs (lRNAs)
According to REF. 24, lRNAs are defined as any ncRNAs >200 nucleotides.

Long interspersed ncRNAs (lincRNAs)
These RNAs have been identified by tiling microarrays in several mouse and human cell types. They derive from non-coding genomic regions that have transcription-dependent chromatin modifications over a distance of at least 5 kb (REF. 2).

Promoter-associated sRNAs (PASRs), promoter-associated lRNAs (PALRs) and terminator-associated sRNAs (TASRs)
These RNAs have been described in different human cell lines by tiling array analysis[24] or RNA–seq of size-selected RNAs (100–300 nucleotide-long PCR products)[35]. PASRs are <200 nucleotides long; PALRs are >200 nucleotides long.

Transcription start site-associated RNAs (TSSa-RNAs)
Small RNAs described in several mouse and human cell types by RNA–seq analysis of an RNA fraction enriched for transcripts in the 16–30 nucleotide range. Further analyses of a few of these transcripts showed them to be 20–90 nucleotides long[25].

Global run-on sequencing (GRO–seq) tags
RNA tags generated from the human IMR90 cell line by GRO–seq, a methodology that reveals nascent transcripts[27]. This methodology does not provide direct indications on the size of the RNA being transcribed.

Transcription-initiation RNAs (tiRNAs)
Tiny RNAs (modal size of 18 nucleotides) that were identified from human cells, chicken embryos and several *Drosophila melanogaster* tissues by RNA–seq of gel-purified sRNA fractions[26].

Promoter upstream transcripts (PROMPTs)
These unstable human transcripts are stabilized by the depletion of exosome factors in human HeLa cells. They are found on both strands, upstream of promoters. There is currently no indication of their size, but they are likely to be short because they are associated with chromatin modifications that mark promoters but not with those that mark transcriptional elongation[46].

Cryptic unstable transcripts (CUTs)
Budding yeast unstable transcripts that are defined as RNAs that can be identified when the Trf4–Air2–Mtr4p polyadenylation (TRAMP) complex and nuclear exosome factors are mutated. They have been characterized by the large-scale sequencing of 3′ LongSAGE tags[42] and tiling arrays[42,43]. They are heterogeneous at their 3′ ends, and they are usually 200–600 nucleotides long. They are principally found associated with promoters on both strands.

Stable unannotated transcripts (SUTs)
Budding yeast ncRNAs that are as yet unannotated (and hence have unknown function). They are more stable than CUTs as they are detected even in the absence of exosome mutants[43]. They are usually longer than CUTs (median length 761 nucleotides). There is not a clear partition between CUTs and SUTs, and some ncRNAs that have been defined as CUTs in one study[42] have been defined as SUTs in another study[43] (discussed further in the main text).

techniques — deep sequencing of small size-selected RNAs from mice[25,26], chicken embryos or *Drosophila melanogaster* tissues[26] and a novel genome-wide run-on technique (known as global run-on sequencing (GRO–seq); see BOX 1) applied to human cells[27] — described remarkably similar transcript profiles. The pattern is that sRNAs are transcribed from sequences that flank the transcription start sites (TSSs) of active promoters: either they are expressed in the same orientation as the gene and there is a peak of expression approximately 50 nucleotides downstream of the TSSs, or they are expressed in a divergent orientation (that is, in the opposite direction to the gene) and there is a peak of expression around 250 nucleotides upstream of the TSSs (FIG. 1). These transcripts have been called TSS-associated RNAs (TSSa-RNAs)[25] or transcription-initiation RNAs (tiRNAs)[26] (BOX 2). Northern blot analyses of a few TSSa-RNAs have revealed small heterogeneous RNAs, similar to PASRs, in the range of 20 to 90 nucleotides. The size distribution of the tiRNAs, as deduced from RNA–seq, has been reported to be notably smaller (modal size ~18 nucleotides[26]) than the named types. Some of the differences between these categories could simply reflect variations in experimental methods, in particular because the RNA fractions analysed were gel purified to enrich for sRNAs (BOX 2). However, there might be multiple types, and therefore the question remains as to whether the TSSa-RNAs, tiRNAs, PASRs and sRNAs revealed by GRO–seq, which are all small and exhibit remarkably similar distributions (FIG. 1), result from a common or distinct biological mechanisms.

What is the source of promoter-associated ncRNAs?

Although the existence of pervasive transcripts at the 5′ ends of genes seems clear, the source of their transcription remains uncertain. At steady state these transcripts

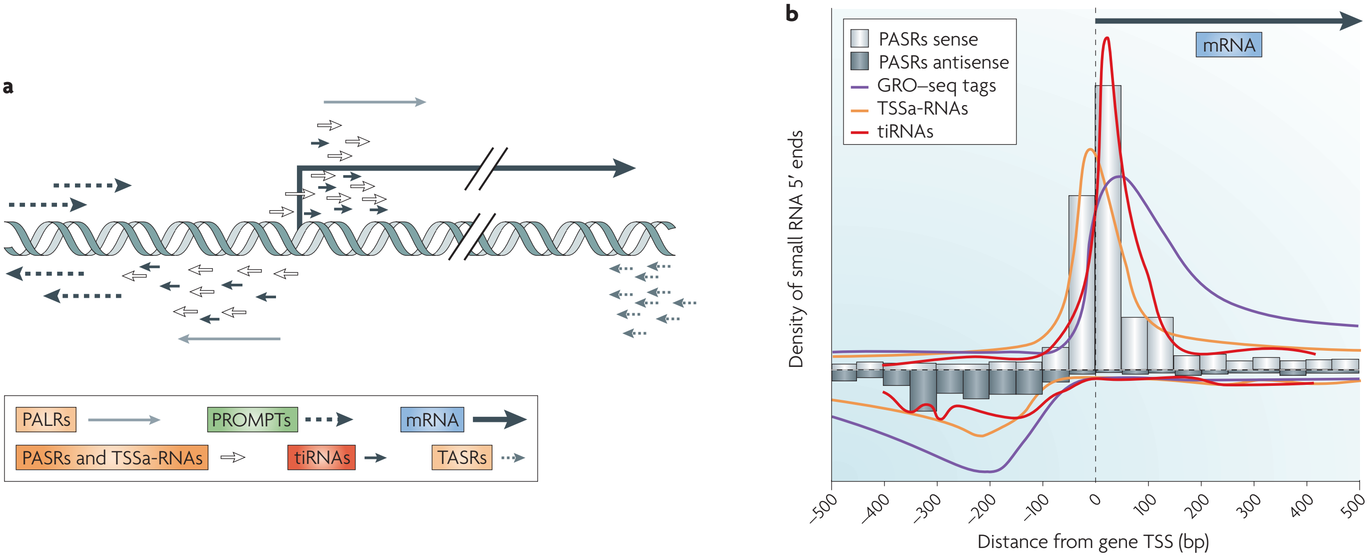

Figure 1 | **Characteristics and distributions of small RNAs found at gene borders in animals. a** | Schematic representation of the different small RNAs associated with promoter or terminator regions in animals. Different classes of RNA are shown by different types of arrow[35,25,46]. For promoter upstream transcripts (PROMPTs), the dashed lines indicate that their characteristics (such as size and heterogeneity) remain unknown. The associated mRNA is shown as a large arrow. **b** | Smoothed distributions of the 5′ ends of promoter-associated small RNAs (PASRs). On the top distribution, sense RNAs relative to the associated mRNA are shown; on the bottom distribution, antisense RNAs relative to the associated mRNA are shown. The zero on the *x* axis represents the position of the mRNA transcription start sites (TSSs). The grey boxes represent the distribution of the PASRs (data from REF. 35), the violet line represents the distribution of global run-on sequencing (GRO–seq) RNA tags (data from REF. 27), the orange line represents the distribution of TSS-associated RNAs (TSSa-RNAs; data from REF. 25) and the red line represents the distribution of transcription-initiation RNAs (tiRNAs; data from REF. 26). PALRs, promoter-associated long RNAs; TASRs, terminator-associated small RNAs.

do not seem to be abundant[24–26], but they are likely to be short-lived molecules. The distribution, size and instability of these sRNAs suggest that at least some might be by-products of so-called 'paused' RNAPIIs. Paused RNAPIIs are engaged with DNA but accumulate between ~20 and 50 nucleotides downstream of some TSSs; however, they retain an elongation potential[28–30]. It is not clear how stable the complexes formed between paused RNAPIIs and the DNA template are and to what extent the peaks of RNAPII binding might reflect repeated rounds of initiation and slower release from the DNA at the site of pausing[27,31]. At a certain frequency, paused RNAPII could dissociate from DNA or prematurely terminate transcription, releasing small transcripts.

PASRs, as revealed by promoter-proximal GRO–seq peaks, seem to be more abundant at highly active promoters with broad TSS regions, such as promoters with high CpG frequencies, than at promoters with a single dominant TSS, which are typically associated with a TATA box[25–27]. This suggests that PASRs might reflect some differences between these two classes of promoters at an early stage of transcription. Importantly, the presence of sRNAs in both the sense and divergent orientation with respect to the gene promoter implies that this early step of the transcription cycle is poorly polarized, with RNAPII being engaged in both directions (see discussion below). This conclusion is further supported by the observation that chromatin modifications associated with promoters flank TSSs in a bimodal distribution, whereas modifications associated with transcription elongation extend unidirectionally from TSSs and across transcribed genes[32–34]. Moreover, a technique designed to determine unambiguously the directionality of transcripts genome-wide also identified a concentration of tags that were divergent from many TSSs in human cells[18]. However, if sense sRNAs are direct by-products of paused RNAPIIs dissociation, their 5′ ends would be expected to match those of precursor mRNAs. However, the majority of their 5′ ends map some distance downstream of the TSS (FIG. 1b), and this observation holds true even when considering only promoters with a single dominant TSS[25,26]. It might be that the sRNA transcripts have undergone some processing and their mapped 5′ ends do not coincide with their TSSs. Processing has indeed been suggested by the analysis of PASRs and promoter-associated lRNAs (PALRs; BOX 2), because these transcripts often overlap, which suggests a possible precursor–product relationship[24,35]. In addition, although RNAPII transcripts often start with a G, there is no bias for the first nucleotide of TSSa-RNAs, which is consistent with the suggestion that the 5′ ends of these sRNAs do not coincide with their primary 5′ ends[25]. However, a notable proportion of PASRs have a 5′ cap[35], as shown by the enrichment of these sRNAs with antibodies against the 5′ cap and by analyses of CAGE tags (which had a distribution pattern similar to that of PASRs) that were generated by a cap-capture technique.

TATA box

A consensus sequence in promoters that is enriched in thymine and adenine residues, and is generally important for the recruitment of the transcriptional machinery.

Cap structures, which are added after the transcription of the first 20 to 30 nucleotides[36], are considered to be the hallmarks of primary RNAPII transcript 5′ ends. However, it has been recently proposed that caps, or similar structures, might be added to cleaved 5′ ends in a secondary reaction[35]. This suggestion is supported by the observation that cap-containing sRNAs and CAGE tags are found in coding sequences and, most importantly, are found more frequently in exons than in introns. This suggests that the majority of these caps were added secondarily after the cleavage of spliced mRNAs. It is therefore not impossible that capped sRNAs have been processed. Their distribution and the presence of a cap could still be consistent with the hypothesis that these transcripts are released from a paused RNAP.

However, there clearly are alternative hypotheses. In particular, sRNAs could be generated by RNAPIIs that initiate transcription in the vicinity of, but not at, the correct mRNA TSS and that terminate prematurely, before they switch to the elongation phase. This model is not exclusive of the pausing model. Indeed, if paused RNAPIIs are released at a certain frequency, one could imagine that they reinitiate in the vicinity of their site of release. Clearly, further work is needed to elucidate the origin(s) of PASRs in animals.

Whatever the detailed biochemical pathway that gives rise to the various sRNAs that cluster around the TSSs of higher eukaryotic genes, one conclusion that their distribution suggests is that many eukaryotic promoters are intrinsically bidirectional; a large proportion of human genes (>50%) are associated with RNAPII engaged in divergent directions and on both sides of TSSs. It is still unknown at which step bidirectional engagement occurs during the initiation of transcription — for example, with respect to the formation of pre-initiation complexes (PICs) (see discussion below). Divergent transcription is found at most active promoters but, for an unknown reason, only the RNAPII engaged in the gene orientation will eventually shift to the elongation phase to generate unidirectional mRNAs.

Pervasive transcription from yeast promoters

In addition to the novel classes of transcripts in animals, recent studies have revealed new types of non-coding transcripts in the yeast *Saccharomyces cerevisiae*. These products of pervasive transcription were first found during analyses of nuclear RNA degradation processes. In the nucleus, 3′ to 5′ exonucleolytic degradation constitutes the major RNA degradation pathway and is performed by the nuclear form of a multifactor complex called the RNA exosome[37]. Efficient RNA degradation by the exosome requires polyadenylation by its associated complex, the Trf4–Air2–Mtr4p polyadenylation (TRAMP) complex[38,39]. This complex participates in the maturation of many stable ncRNAs, such as ribosomal RNAs, snRNAs and snoRNAs[37]. TRAMP inactivation also allowed a novel class of ncRNAs to be discovered. The remarkable feature of this novel class is that the ncRNAs are virtually undetectable in normal cells and are only revealed by compromising the activities of the exosome, the TRAMP complex or both[40]. Therefore they have been called cryptic unstable transcripts (CUTs). CUTs are ~200–600 nucleotide-long ncRNAs. They are capped, heterogeneous in size owing to multiple 3′ ends and extremely unstable. Genome-wide studies with low-resolution DNA arrays have suggested that CUTs are promoter-associated RNAs[41].

Two recent studies have now established the distribution of CUTs in budding yeast at high resolution. In a strain depleted for both exosome and TRAMP components, the first study used immunoprecipitation of nuclear RNA to purify an RNA fraction that was highly enriched for CUTs. This fraction was analysed by tiling arrays and high-throughput sequencing by 3′ LongSAGE (BOX 1) to characterize the heterogeneous 3′ ends of CUTs at nucleotide resolution and, most importantly, to allow discrimination of overlapping transcripts[42]. In an independent study, sensitive tiling array hybridization was used to analyse the transcriptome of a yeast exosome mutant (to study CUTs) and the transcriptomes of yeast grown in various conditions. This latter analysis revealed another class of ncRNAs called stable unannotated transcripts (SUTs)[43].

SUTs are defined as ncRNAs that are more stable than CUTs. They are also longer on average than CUTs (median length 761 nucleotides)[43]. However, there is no strict demarcation between SUTs and CUTs; many transcripts defined as SUTs in one study[43] were identified as CUTs in another study[42]. The tiling array and sequencing analyses confirmed that CUTs (and SUTs) are most often associated with promoters (FIG. 2) and further indicated that CUTs almost exclusively arise from nucleosome-free regions (NFRs). NFRs are found in intergenic spaces, in particular close to the 5′ ends of genes[44], where they mark, in yeast as in mammals, core promoter regions. CUTs therefore usually share an NFR with a gene. Most strikingly, these new data, from which transcript strand specificity could be determined, showed that in more than 78% of cases, CUTs were transcribed in the opposite direction to their associated gene promoters[42,43]. Together, these studies show that more than 30% of yeast promoter regions generate divergent transcripts but, in most cases, transcripts in the non-coding direction are rapidly degraded and are therefore not readily detectable in normal cells.

Where do CUTs come from? The 5′ ends of most sense CUTs that have been mapped are located a few hundred nucleotides upstream of gene TSSs. Therefore they cannot originate from RNAP paused at the TSS. There is no correlation between the strength of promoters and the level of their associated CUTs. However, antisense CUTs often share the same regulation profiles in response to different growth conditions as the mRNA that shares the same NFR. This suggests that they are under the control of the same transcription activators[42,43]. One current model is that several steps of transcription initiation are not highly specific with respect to the polarity of transcription. Pools of general transcription factors, which are recruited by transcription activators, could form alternative PICs in an NFR, but only the PIC located correctly relative to the ORF would give rise to

Pre-initiation complex
This is formed by the general transcription factors that assemble after recruitment by transcription activators. At TATA box-containing promoters, the pre-initiation complexes assemble on the TATA box and position the RNA polymerase for transcription.

Exosome
A protein complex that has 3′ to 5′ exonuclease activity (an additional endonuclease activity has been described). There are two forms of the exosome that differ in their associated co-factors; one complex is nuclear and one is cytoplasmic.

Nucleosome-free regions
These are regions of the chromatin that are depleted from nucleosomes. They are mainly found at gene boundaries, in particular at the 5′ end at which they correspond to the core promoter regions.

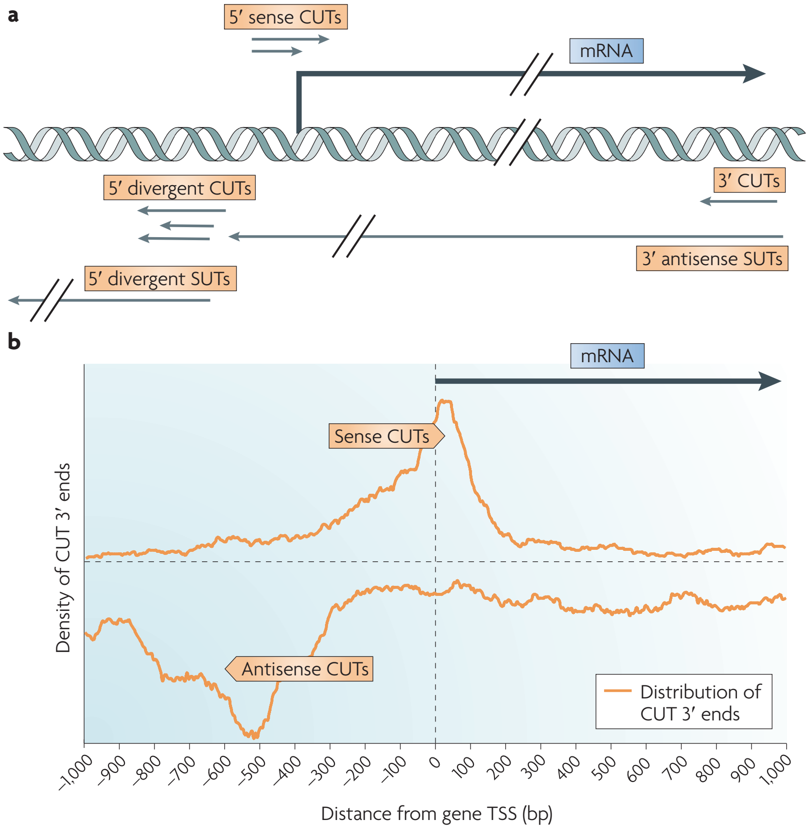

Figure 2 | **Characteristics and distributions of CUTs and SUTs in *Saccharomyces cerevisiae*. a** | Schematic representation of cryptic unstable transcripts (CUTs)[42,43] and stable unannotated transcripts (SUTs)[43] relative to an mRNA. **b** | Distribution of the 3′ ends of CUTs relative to mRNA transcription start sites (TSSs). At the top, sense CUTs relative to the associated mRNAs are shown; at the bottom, antisense (divergent) CUTs relative to the associated mRNAs are shown. The zero on the *x* axis represents the position of the mRNA TSSs. The orange arrows indicate the approximate positions of sense and antisense CUTs.

stable transcripts (FIG. 3). This model has been supported by the observation that mutations in a TATA box reduce the expression of the mRNA as expected but, conversely, greatly enhance the expression of an associated antisense CUT[42]. Therefore the CUT and the mRNA do not issue from the same PIC. Instead, distinct sites of PIC formation must exist for transcription of the mRNA and the CUT. These sites might compete for the same 'pool' of general transcription factors that are recruited by the activators (FIG. 3). Polarity of transcription might then be provided in part by a post-transcriptional quality-control mechanism that targets cryptic transcripts for degradation (see below).

Although this model might account for the majority of CUTs, other mechanisms are also likely to generate CUTs. For example, some CUTs have been found to arise as by-products of unconventional regulation mechanisms, as described below. Finally, although there is substantially less pervasive transcription associated with 3′ NFRs than with 5′ NFRs, the existence of 3′ NFR-associated ncRNAs suggests that the absence of nucleosomes might be sufficient to promote some kind of 'background' RNAPII transcription. However, the transcripts have well-defined TSSs and therefore do not resemble random noise. An intriguing possibility is that 3′ sRNAs in yeast (and TASRs in mammals) might arise from genes adopting a loop structure that juxtaposes promoters and terminators[45].

Similarities and differences between promoter-associated transcripts in animals and in budding yeast. The distribution of CUTs around gene promoters is highly reminiscent of the distribution of PASRs, TSSa-RNAs and tiRNAs (compare FIG. 1 and FIG. 2). However, there are differences that suggest that these pervasive transcripts might not result from exactly the same phenomenon. PASRs and TSSa-RNAs are 50 to 250 nucleotides long and tiRNAs are even shorter, with a size distribution that peaks at 18 nucleotides. By contrast, CUTs are typically heterogeneous in size but are longer on average (200 to 600 nucleotides long; median length around 400 nucleotides). Promoter-associated pervasive transcription gives rise to both sense and antisense transcripts, resulting in a mirror-image-like distribution, but this distribution is not completely equivalent in animals and yeast. In animals, the 5′ ends of pervasive transcripts in the sense orientation with respect to the gene promoter generally map downstream of the TSS of genes, which is consistent with the hypothesis that they are somehow related to paused RNAPIIs (see above). In yeast, by contrast, most mapped sense CUT 5′ ends are located a few hundred nucleotides upstream of gene TSSs and therefore cannot be associated to paused RNAPII. Also, in animals, promoters that efficiently produce PASRs and TSSa-RNAs are strong promoters, which is consistent with the paused polymerase hypothesis, but in yeast there is no correlation between the strength of promoters and the amount of associated CUTs.

However, the depletion of core exosome factors with siRNAs in human cells (all TRAMP–exosome factors are conserved from yeast to humans[37]) allowed an additional class of transcripts to be discovered. These transcripts, like CUTs, are only detectable when exosome activity is compromised. Similar to CUTs, they emanate from a broad region several hundred nucleotides upstream of the promoters — that is, substantially upstream from PASRs and TSSa-RNAs (FIG. 1a). Reverse transcription PCR analyses have shown that transcripts of this class — known as promoter upstream transcripts (PROMPTs)[46] — are transcribed from both strands. Their size is unknown, but the regions from which they are transcribed are bound by RNAPII and are marked by chromatin modifications associated with transcription initiation but not elongation. Since the elongation phase of transcription starts ~50 nucleotides downstream of the TSS, PROMPTs are probably small transcripts in the same size range as PASRs and TSSa-RNAs, and hence are smaller than CUTs. Also, like PASRs but unlike CUTs, the strength of the array hybridization signal for PROMPTs correlates with the strength of the downstream promoters. For example, they are especially prominent at CpG-rich promoters.

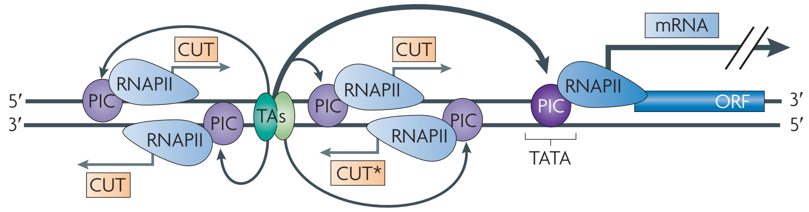

Figure 3 | **Model for the generation of CUTs around gene promoters.** The model presents one possible origin for cryptic unstable transcripts (CUTs). This model is not exclusive of other mechanisms, such as those discussed in the main text for the formation of CUTs associated with unconventional regulation mechanisms. In the mechanism proposed in the figure, general transcription factors are recruited to promoter regions by transcription activators (TAs). This pool of general transcription factors can assemble (curved arrows) pre-initiation complexes (PICs, purple). PICs can recruit RNA polymerase II (RNAPII, cyan shapes). PIC assembly can occur at strong TATA box-binding protein (TBP)-binding sites, such as TATA boxes, leading to the transcription of mRNA. In addition to this strong site of PIC formation (dark purple), cryptic sites can promote the assembly of cryptic PICs (light purple). These PICs can assemble in either orientation and generate CUTs (orange). This model has been supported by experimental data for a CUT at the triosephosphate isomerase 1 (*TPI1*) locus[42]. For this case the configuration is as labelled with an asterisk in the figure: the synthesis of this CUT is co-regulated with the *TPI1* mRNA, which is consistent with the CUT being under the control of the same TAs, and the CUT competes with *TPI1* mRNA for the same pool of transcription factors[42]. ORF, open reading frame.

The promoter-associated pervasive transcripts identified so far might originate from different mechanisms in yeast and animals, but some of the experiments that revealed these sRNAs in animals (deep sequencing of small RNA libraries and GRO–seq) have not yet been performed in yeast. Therefore, it is possible that the equivalent RNAs also exist in yeast. In any event, one important conclusion drawn from both the yeast and animal studies is that many promoter regions generate divergent transcripts and therefore seem to be bidirectional. The model presented in FIG. 3 for the origin of promoter-associated CUTs and the data that support it are consistent with bidirectionality, as they suggest the existence of cryptic sites where PICs can assemble without polarity preference. Hence, bidirectionality could occur at several steps in the transcription cycle: firstly at PIC assembly, which is proposed to account for the existence of divergent CUTs in yeast, and secondly at the step in which the RNAPII is engaged for transcription but has not yet shifted to the full elongation phase, as suggested by the distribution of the PASRs described in animals.

Pervasive transcription and quality control

Degradation of the products of pervasive transcription in yeast. One key question is how the products of pervasive transcription are distinguished from long stable RNAs and degraded. These processes are best understood in yeast. As discussed above, CUTs were identified by mutation of the TRAMP–exosome complex, which is normally involved in their rapid degradation[40]. But how are CUTs distinguished from coding RNAs and sent to this degradation pathway? Part of the answer came from studying the transcriptional termination and maturation of small stable RNAs, such as snoRNAs. It had been previously shown that, when the activity of the nuclear exosome is compromised, snoRNA precursor molecules accumulate in the form of longer transcripts. These transcripts have heterogeneous polyadenylated 3′ ends that resemble CUT 3′ ends[47]. Indeed, polyadenylation of these extended transcripts was found to be, as for CUTs, dependent on TRAMP[39,40]. TRAMP assists the exosome in trimming 3′ extensions as a step of the snoRNA maturation process[48]. Interestingly, this maturation step was found to be coupled to transcription termination.

Termination of the elongated snoRNA precursors relies not on the cleavage and polyadenylation machinery that is involved in mRNA termination but on another complex that also binds to the carboxy-terminal domain (CTD) of RNAPII. This complex contains the Nrd1 and Nab3 RNA-binding proteins as well as the putative RNA helicase Sen1 (REF. 49). Termination of transcription occurs downstream of tetranucleotide motifs, which form binding sites for Nrd1 and Nab3 on the nascent RNA[50]. The arrangement of Nrd1 and Nab3 binding sites is flexible, although clusters of sites seem to be recognized more efficiently. As termination by recognition of Nrd1 and Nab3 sites is not an efficient process, multiple sites are required to achieve complete termination, which results in heterogeneous 3′ ends. The Nrd1–Nab3–Sen1 complex, bound to the RNAPII CTD, directly interacts with the nuclear exosome[51], thereby coupling transcription termination with 3′ to 5′ exonucleolytic trimming by the TRAMP–exosome complex (FIG. 4a). For snoRNAs, this process is controlled by small nucleolar ribonucleoprotein (snoRNP) factors that assemble early during transcription[52,53] and block the exosome, thereby defining the 3′ end of the snoRNAs.

The similarities between the 3′ ends of snoRNA precursors and of CUTs (that is, they are both heterogeneous and polyadenylated by TRAMP) suggested that the same termination–degradation coupled mechanisms might occur for CUTs. Indeed, it was shown that transcription termination of the CUTs also involves the Nrd1–Nab3–Sen1 complex[54,55]. In contrast to snoRNAs, in which the snoRNP factors restrict the exosome activity, CUTs are unprotected and are therefore completely degraded. Interestingly, this coupling between transcription termination and degradation is likely to explain why CUTs are almost undetectable in wild-type cells, as they are degraded as soon as they are synthesized. As the sites recognized by Nrd1 and Nab3 are short, ubiquitous sequences, this process seems to act as a 'default' mechanism that will terminate transcription and trigger degradation of any transcript that is not protected. Nrd1 and Nab3 recognition sites are less frequent in coding sequences (relative to intergenic or antisense sequences), probably as a result of codon usage[55]. ORF sequences are therefore more 'immune' to this degradation process, even though some termination might occur in ORFs, particularly in introns. Also, this degradation process is only efficient during the first few hundred nucleotides of transcription (less than 1 kb[56]), which is the range over which the Pcf11-dependent cleavage and polyadenylation processes that are involved in mRNA transcription termination do not operate efficiently. The Nrd1-dependent process is adapted to the small size of snoRNAs, and therefore the small size of CUTs is likely to result from their mode of transcription termination.

The mode of transcription termination might distinguish CUTs from SUTs. As mentioned above, there is no clear distinction between CUTs and the longer, more stable SUTs. The Nrd1- and Nab3-dependent termination process is not very efficient, therefore some pervasive transcription might be terminated by this pathway in the first few hundred nucleotides of transcription, which will generate CUTs, whereas other transcripts might be terminated at a downstream poly(A) site by a Pcf11-containing cleavage and polyadenylation complex, which will generate SUTs (FIG. 4b). The production of CUTs and SUTs — with their differing lengths and stabilities — could depend on the balance between the two termination processes. The longer

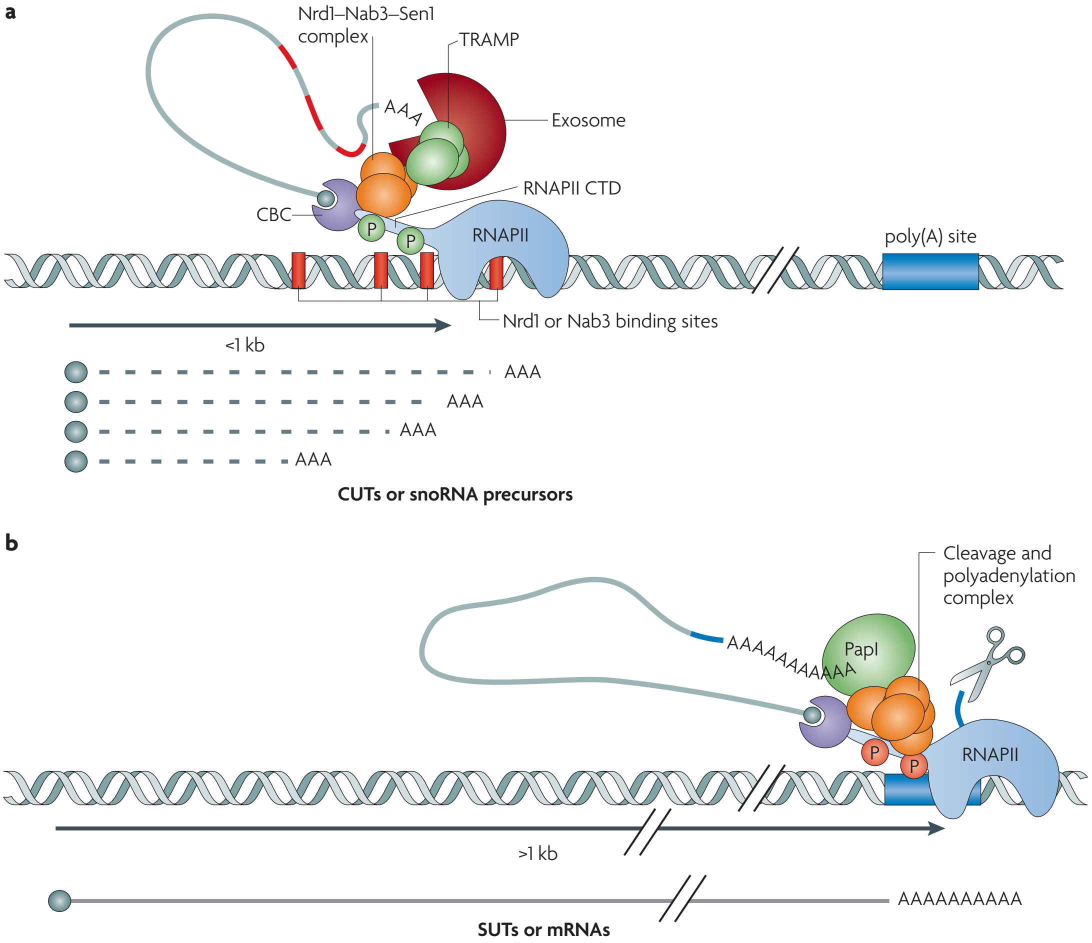

Figure 4 | **The instability of CUTs is linked to their mode of transcription termination. a** | The figure shows how the mode of transcription termination of cryptic unstable transcripts (CUTs) is linked to their rapid turnover. During the first few hundred nucleotides of transcription (<1 kb) the RNA polymerase II (RNAPII) carboxy-terminal domain (CTD) is phosphorylated on serine 5 (green circles labelled with P)[56,79]. This promotes a mode of termination that is specific to small non-coding RNAs, such as small nucleolar RNAs (snoRNAs), small nuclear RNAs and CUTs[49,54,55]. In this mode of termination, the Nrd1–Nab3–Sen1 complex (orange) assembles on the RNAPII CTD after clusters of small tetranucleotide sequences that are recognized by Nrd1 (GUAR) or Nab3 (UCUU) are transcribed. These sequences are shown by red boxes on the DNA and red sections on the RNA (curved grey line). This complex induces transcription termination[49,50,73,80]. The grey circle at the end of the RNA represents the cap, which is bound by the nuclear cap-binding complex (CBC). CBC interacts both with the RNAPII CTD and the Nrd1–Nab3–Sen1 complex. The Nrd1–Nab3–Sen1 complex also physically interacts with the nuclear exosome (shown in dark red)[51], which allows coupling between transcription termination of the CUTs and their degradation by the exosome. This degradation is aided by the oligoadenylation activity of the associated Trf4–Air2–Mtr4p polyadenylation (TRAMP) complex (in green)[38–40]. This process generates small (<1 kb) transcripts (dashed grey lines at the bottom of the figure), which are heterogeneous at their 3′ ends and short lived owing to the coupling between transcription termination and degradation. **b** | If the polymerase does not encounter clusters of Nrd1 or Nab3 sites, the phosphorylation status of the CTD changes as it progresses (red circles labelled with P, which indicate phosphorylation of S3 of the RNAPII CTD). This change makes the polymerase competent for assembly of the cleavage and polyadenylation complex (shown in green for the poly(A) polymerase PapI). This complex promotes transcription termination and RNA polyadenylation of the RNA at poly(A) sites (blue box on the DNA). The RNAs terminated by this mechanism — which can be mRNAs if the RNAs encode proteins or stable unannotated transcripts (SUTs)[43] if they are non-coding — are longer and more stable than CUTs.

polyadenylated SUTs are likely to be exported to the cytoplasm and are probably degraded by the Xrn1-dependent exonucleolytic pathway[57,58].

Does this model extend to animals? PASRs and TSSa-RNAs are clearly short lived — that is, they are non-abundant at steady state but constitute a major portion of native transcripts. Is their degradation dependent on a pathway that is similar to the CUT degradation pathway? Depletion of core exosome factors by siRNA does not stabilize any transcripts that correspond to PASRs or TSSa-RNAs, which suggests that an exosome-dependent pathway does not degrade these products. It is possible that their degradation is linked to the mechanism that distinguishes mRNA transcription from divergent transcription — that is, in divergent transcription the RNAPII does not shift from the slow initiation phase (paused RNAPII) to the fast elongation phase. However, this mechanism is yet to be characterized.

Is pervasive transcription functional in animals?

Whether pervasive transcription essentially consists of 'futile' background transcriptional noise or has functional significance is a matter of debate. As introduced above, pervasive transcription can, broadly, be categorized as either giving rise to relatively long and stable transcripts (for example, lncRNAs) or to short and unstable transcripts (for example, CUTs, PASRs, TSSa-RNAs, tiRNAs and PROMPTs). There are many examples of lncRNAs being involved in gene regulation, and a number of reviews on the subject are available[4–6], therefore I do not discuss these in detail. One feature is that lncRNAs are involved in various different processes: they can act in *cis* or in *trans*, in sense or antisense, and can function as transcriptional activators or repressors. It is emerging that in many cases they might act through chromatin modification pathways. For example, a recent report that described several thousand human lincRNAs showed that they are bound to diverse chromatin-modifying complexes and are therefore likely to be involved in epigenetic mechanisms, possibly by guiding chromatin-modifying complexes to specific genomic loci[59].

Possible functions for promoter-associated RNAs. The above discussion on the origins and patterns of promoter-associated pervasive transcription suggests that it represents biological noise. It might reflect the intrinsic properties of multistep transcription-initiation processes, with each step being associated with proof-reading or quality-control mechanisms. However, some classes of sRNAs in eukaryotes — notably siRNAs, microRNAs and piwi-interacting RNAs (piwiRNAs) — are known to be functional[1]. Therefore, even if the synthesis of sRNAs, such as PASRs, TSSa-RNAs, tiRNAs or PROMPTs, primarily represents some kind of transcriptional background, these sRNA molecules could have intrinsic functional potential.

In animals, it is uncertain whether promoter-associated transcripts are functional. It has been suggested, for example, that a general function of promoter-associated pervasive transcription might be to help maintain an open chromatin state[60] and/or to keep a pool of RNAPII available that can rapidly be used for mRNA synthesis, which has been suggested to be the function of paused RNAP. Some more specific functions are also now being described. For example, in human cell lines, *cis*-acting heterogeneous RNAs that are ~200–330 nucleotides long and originate from the promoter region of the cyclin D1 (*CCND1*) gene have been shown to bind the translocated in liposarcoma (TLS) repressor. They allosterically activate the repressor and tether it to the *CCND1* promoter to inhibit *CCND1* expression[61]. Also, transfecting a human cell line with synthetic PASRs that mimic naturally occurring sense or antisense PASRs that are associated with the MYC or connective tissue growth factor (CTGF) promoters weakly reduced the expression of the corresponding mRNAs[35]. Other mechanisms have been reported that involve the generation of siRNAs that can either activate[62] or repress[63] transcription, perhaps through the formation of an RNA–DNA triplex[64].

Is pervasive transcription functional in yeast?

In budding yeast, there are now several examples of ncRNAs that are implicated in gene regulation. In most cases, these functional ncRNAs are antisense lncRNAs. Similarly to animal lncRNAs, yeast lncRNAs can be activators or repressors and can act in *cis* or in *trans*, sometimes by chromatin modification[3]. But are small unstable transcripts functional? So far, there is little evidence that divergent CUTs are functional. By contrast, the less frequently occurring sense CUTs have the potential to interfere with the expression of mRNA, as the TSSs and upstream promoter sequences of the CUT and mRNA overlap. Indeed, such transcription interference has already been described for the *SER3* gene, the expression of which is repressed by an upstream transcript called *SRG1*, which overlaps its site of PIC formation[65,66] (FIG. 5a). Strong sense CUTs are often associated with genes from particular pathways, such as glycolysis or the nucleotide biosynthetic pathway[42]. Sense CUTs and their associated mRNAs tend to have reciprocal expression patterns, which is consistent with sense CUTs being involved in gene repression[42] in a similar way to *SRG1*. Note that this mechanism does need to be restricted to sense ncRNAs, and it is possible that some antisense SUTs regulate transcription in a similar manner.

Individual dissection of how and by which factors sense CUTs are regulated with respect to their associated mRNAs might be required to determine how widespread transcriptional interference is. For example, detailed analyses of the regulation of genes that are involved in nucleoside triphosphate biosynthesis (*IMD2* (REFS 67–70), *URA2*, *URA8* and *ADE12* (REF. 71)) uncovered an unexpected mode of regulation. Transcription can start at two alternative sites: when nucleotide concentration is low, transcription initiates at a downstream site and the mRNA is generated, but in the presence of high nucleotide concentrations, transcription initiates at an upstream site, which generates a CUT and therefore results in unproductive transcription. Nrd1 and Nab3 binding sites between the two alternative TSSs are required for rapid degradation of the CUT. Surprisingly, transcription from the two initiation sites is driven by the same PIC. It assembles on a TATA box that drives the synthesis of both the CUT and the mRNA (FIG. 5b). RNAPII acts directly as the sensor of the nucleotide concentration and determines the site at which transcription will initiate[72]. It is therefore unlikely that this unconventional mode of regulation will apply to many other pathways.

Another unconventional mode of regulation that generates a CUT is the autoregulation of Nrd1 synthesis. Nrd1 and Nab3 binding sites are located in the 5′ region of the *NRD1* mRNA sequence and direct premature transcription termination by a mechanism that is sensitive to the amount of Nrd1. This termination generates a CUT with the same TSS as the mRNA[73] (FIG. 5c). It is possible that other genes are also repressed by such a mechanism. In these examples, the CUT molecule itself does not act, even indirectly, in regulation. Instead it is a by-product of an unconventional regulation mechanism. It is important to stress that at present it is not known how widespread these regulatory phenomena are. In addition, other types of regulatory mechanisms that are linked to CUTs are likely to be uncovered. Answering these questions will be an important challenge in the field over the coming years, although answering them is likely to involve labour-intensive dissection of many individual examples.

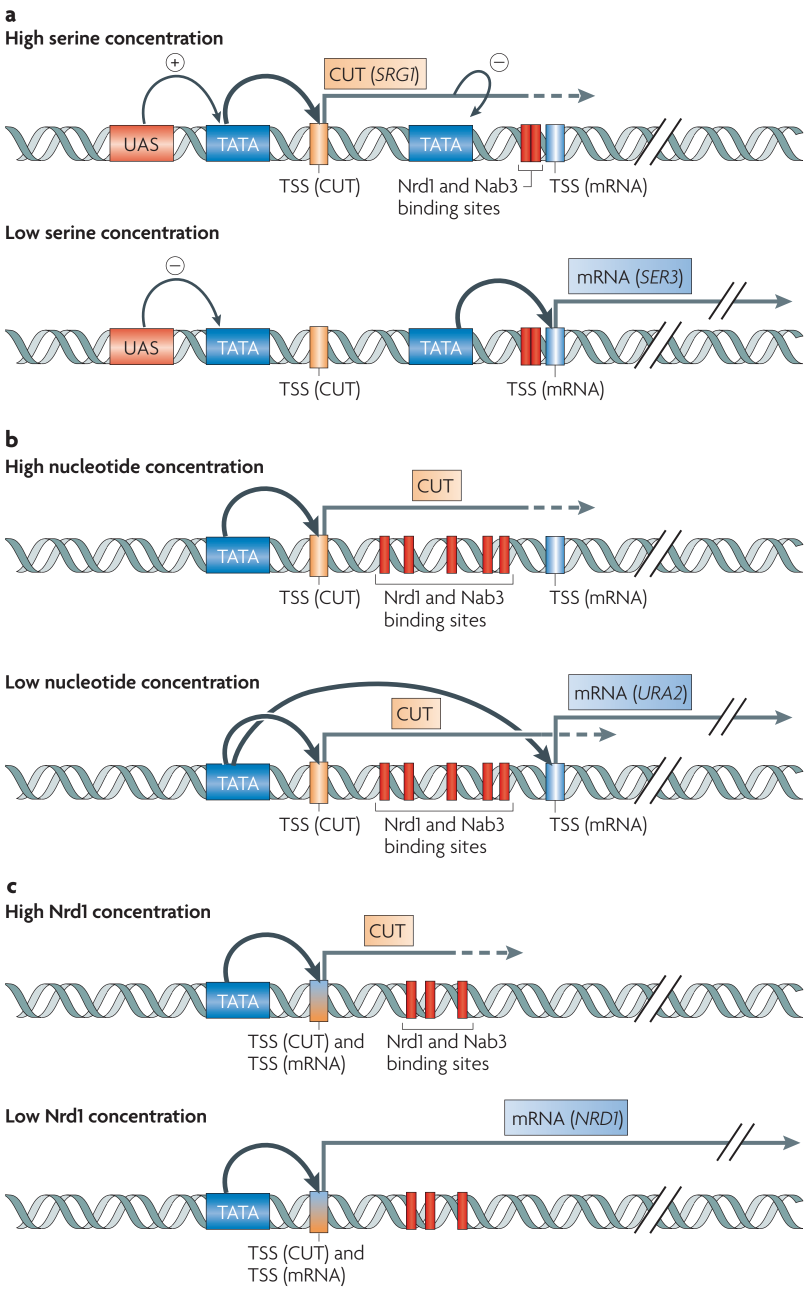

Figure 5 | **Unconventional transcription regulation mechanisms that generate CUTs.** Some unusual regulation mechanisms generate cryptic unstable transcripts (CUTs). Unlike some non-coding RNAs that might directly participate in regulation, these CUTs are thought to be by-products of mechanisms in which they do not play an active part. **a** | The archetypal mechanism is transcription interference, as exemplified by the *SER3* locus. At a high serine concentration, the serine-dependent transcription activator Cha4 binds an upstream activating sequence (UAS) to promote (curved black arrow with + sign) transcription of the non-coding RNA from senescence-related gene 1 (*SRG1*), which has the characteristics of a CUT. *SRG1* transcription overlaps the TATA box and transcription start site (TSS) of the *SER3* mRNA, which inhibits its expression by transcription interference (curved black arrow with – sign). Ectopic expression of *SRG1* has no effect on *SER3* expression. At a low serine concentration, *SRG1* transcription is not induced, which allows constitutive expression of the *SER3* mRNA. Transcription initiation of *SRG1* and the *SER3* mRNA depends on their own, distinct TATA boxes. **b** | Genes involved in the nucleotide synthetic pathway are regulated by a specific mechanism. A single pre-initiation complex (PIC) assembled on a single TATA box can induce transcription initiation from two distinct transcription initiation sites. At a high nucleotide concentration (for example, a high concentration of uracil in the case of *URA2*), the polymerase starts at an upstream site (TSS (CUT)). Transcription then proceeds through a region that is rich in Nrd1 and Nab3 sites, which results in early transcription termination and rapid degradation of the transcript (CUT). At a low nucleotide concentration, the parameters dictating the choice of the TSSs by RNA polymerase II (RNAPII) are modified, resulting in the use of a downstream TSS (TSS (mRNA)), so Nrd1-dependent termination and degradation is avoided and a stable mRNA is produced. **c** | *NRD1* expression is autoregulated. Transcription is initiated from a single PIC and TSS region but goes through a region under the control of Nrd1-induced termination in a manner that is dependent upon the concentration of Nrd1. At a high Nrd1 concentration, transcription terminates early, generating a CUT. At a low Nrd1 concentration, part of the transcription, although it initiates at the same site, escapes Nrd1-dependent termination to generate a full-length *NRD1* mRNA.

Conclusions

The unbiased characterization of transcriptomes has led to the identification of a large repertoire of unexpected transcripts. Many of the individual pervasive transcripts in animals are found only in particular tissues, cell lines, growth conditions or mutant backgrounds[10,13,74]. In addition, independent experiments performed on the same cell lines overlap only modestly in their identification of some sRNAs, such as PASRs[35], showing that sRNA identification is not saturated. Likewise, in yeast, genome-wide analysis of CUTs has so far only been performed under one condition — the exponential growth phase in complete medium of TRAMP–exosome mutants[42,43]. Individual SUTs are often differentially expressed in different conditions, depending on, for example, the carbon source[43]. It is likely that studying mutants in other RNA degradation pathways, or studying different growth conditions or growth phases, will reveal further classes of ncRNAs. The potential involvement of some of these transcripts in gene regulation has opened an unexpected area of investigation that is likely to uncover further novel processes.

The discovery that transcription in eukaryotes might be less specific than previously thought — in that it generates a number of possibly non-functional ncRNAs that must be degraded by secondary quality-control mechanisms — raises the question of the utility of this apparently ‘messy’ process. This might be the price to pay for flexibility and could have two important consequences. First, a lack of rigidity in transcription initiation steps might allow many regulatory processes to act together. For example, broad or specific modes of transcriptional regulation could act at different steps of initiation. Second, a process that is not too rigid might allow scope for rapid evolution. For example, the loose distinction between CUTs and SUTs shows how easy it is to generate new stable transcripts that can be co-opted for a variety of functions. Indeed, eukaryotes show a remarkable ability to generate complex transcriptomes from a limited number of genes by allowing great flexibility in transcription initiation and great malleability in RNA processing. These properties might have been key to the evolutionary success of eukaryotes.

1. Carthew, R. W. & Sontheimer, E. J. Origins and mechanisms of miRNAs and siRNAs. *Cell* **136**, 642–655 (2009).
2. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
3. Berretta, J. & Morillon, A. Pervasive transcription constitutes a new level of eukaryotic genome regulation. *EMBO Rep.* **10**, 973–982 (2009).
4. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: insights into functions. *Nature Rev. Genet.* **10**, 155–159 (2009).
5. Ponting, C. P., Oliver, P. L. & Reik, W. Evolution and functions of long noncoding RNAs. *Cell* **136**, 629–641 (2009).
6. Yazgan, O. & Krebs, J. E. Noncoding but nonexpendable: transcriptional regulation by large noncoding RNA in eukaryotes. *Biochem. Cell Biol.* **85**, 484–496 (2007).
7. Velculescu, V. E. *et al.* Characterization of the yeast transcriptome. *Cell* **88**, 243–251 (1997).
8. Goffeau, A. 1996: a vintage year for yeast and *Yeast*. *Yeast* **12**, 1603–1605 (1996).
9. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).
10. Numata, K. *et al.* Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. *Genome Res.* **13**, 1301–1306 (2003).
11. Hayashizaki, Y. & Carninci, P. Genome Network and FANTOM3: assessing the complexity of the transcriptome. *PLoS Genet.* **2**, e63 (2006).
12. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
13. Ravasi, T. *et al.* Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.* **16**, 11–19 (2006).
14. Kampa, D. *et al.* Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14**, 331–342 (2004).
15. Cloonan, N. & Grimmond, S. M. Transcriptome content and dynamics at single-nucleotide resolution. *Genome Biol.* **9**, 234 (2008).
16. Kapranov, P., Sementchenko, V. I. & Gingeras, T. R. Beyond expression profiling: next generation uses of high density oligonucleotide arrays. *Brief Funct. Genomic Proteomic* **2**, 47–56 (2003).
17. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10**, 57–63 (2009).
18. He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N. & Kinzler, K. W. The antisense transcriptomes of human cells. *Science* **322**, 1855–1857 (2008).
19. Carninci, P. Molecular biology: the long and short of RNAs. *Nature* **457**, 974–975 (2009).
20. Carninci, P., Yasuda, J. & Hayashizaki, Y. Multifaceted mammalian transcriptome. *Curr. Opin. Cell Biol.* **20**, 274–280 (2008).
21. Johnson, J. M., Edwards, S., Shoemaker, D. & Schadt, E. E. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.* **21**, 93–102 (2005).
22. Kapranov, P., Willingham, A. T. & Gingeras, T. R. Genome-wide transcription and the implications for genomic organization. *Nature Rev. Genet.* **8**, 413–423 (2007).
23. Ponjavic, J. & Ponting, C. P. The long and the short of RNA maps. *Bioessays* **29**, 1077–1080 (2007).
24. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484–1488 (2007).
The first description of promoter-associated ncRNAs in mammals.
25. Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849–1851 (2008).
A description of divergent small pervasive transcripts at gene promoters.
26. Taft, R. J. *et al.* Tiny RNAs associated with transcription start sites in animals. *Nature Genet.* **41**, 572–578 (2009).
27. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).
This study produced a genome-wide map of run-on transcripts in mammals that reveals bidirectional transcription at gene promoters.
28. Muse, G. W. *et al.* RNA polymerase is poised for activation across the genome. *Nature Genet.* **39**, 1507–1511 (2007).
29. Zeitlinger, J. *et al.* RNA polymerase stalling at developmental control genes in the *Drosophila melanogaster* embryo. *Nature Genet.* **39**, 1512–1516 (2007).
30. Gilmour, D. S. Promoter proximal pausing on genes in metazoans. *Chromosoma* **118**, 1–10 (2009).
31. Margaritis, T. & Holstege, F. C. Poised RNA polymerase II gives pause for thought. *Cell* **133**, 581–584 (2008).
32. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
33. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**, 77–88 (2007).
34. Kim, T. H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436**, 876–880 (2005).
35. Fejes-Toth, K. *et al.* Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* **457**, 1028–1032 (2009).
36. Rasmussen, E. B. & Lis, J. T. *In vivo* transcriptional pausing and cap formation on three *Drosophila* heat shock genes. *Proc. Natl Acad. Sci. USA* **90**, 7923–7927 (1993).
37. Houseley, J. & Tollervey, D. The many pathways of RNA degradation. *Cell* **136**, 763–776 (2009).
38. Vanacova, S. *et al.* A new yeast poly(A) polymerase complex involved in RNA quality control. *PLoS Biol.* **3**, e189 (2005).
39. LaCava, J. *et al.* RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. *Cell* **121**, 713–724 (2005).
40. Wyers, F. *et al.* Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121**, 725–737 (2005).
This study provided the first description of CUTs in yeast. In addition, in parallel with other reports, this article describes the novel poly(A) polymerase-containing complex TRAMP and its role in exosome-mediated nuclear RNA degradation.
41. Davis, C. A. & Ares, M. Jr. Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **103**, 3262–3267 (2006).
42. Neil, H. *et al.* Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* **457**, 1038–1042 (2009).
43. Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033–1037 (2009).
References 42 and 43 report the genome-wide description of CUTs in yeast. In addition, the second article describes a novel type of more stable pervasive transcripts called SUTs. One main conclusion of both reports is that widespread bidirectional promoters generate most pervasive transcripts in yeast.
44. Lee, W. *et al.* A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genet.* **39**, 1235–1244 (2007).
45. O’Sullivan, J. M. *et al.* Gene loops juxtapose promoters and terminators in yeast. *Nature Genet.* **36**, 1014–1018 (2004).
46. Preker, P. *et al.* RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322**, 1851–1854 (2008).
47. van Hoof, A., Lennertz, P. & Parker, R. Yeast exosome mutants accumulate 3′-extended polyadenylated forms of U4 small nuclear RNA and small nucleolar RNAs. *Mol. Cell Biol.* **20**, 441–452 (2000).
48. Allmang, C. *et al.* Functions of the exosome in rRNA, snoRNA and snRNA synthesis. *EMBO J.* **18**, 5399–5410 (1999).

49. Steinmetz, E. J., Conrad, N. K., Brow, D. A. & Corden, J. L. RNA-binding protein Nrd1 directs poly(A)-independent 3′-end formation of RNA polymerase II transcripts. *Nature* **413**, 327–331 (2001).

50. Carroll, K. L., Pradhan, D. A., Granek, J. A., Clarke, N. D. & Corden, J. L. Identification of *cis* elements directing termination of yeast nonpolyadenylated snoRNA transcripts. *Mol. Cell Biol.* **24**, 6241–6252 (2004).

51. Vasiljeva, L. & Buratowski, S. Nrd1 interacts with the nuclear exosome for 3′ processing of RNA polymerase II transcripts. *Mol. Cell* **21**, 239–248 (2006).

52. Ballarino, M., Morlando, M., Pagano, F., Fatica, A. & Bozzoni, I. The cotranscriptional assembly of snoRNPs controls the biosynthesis of H/ACA snoRNAs in *Saccharomyces cerevisiae*. *Mol. Cell Biol.* **25**, 5396–5403 (2005).

53. Yang, P. K. *et al.* Cotranscriptional recruitment of the pseudouridylsynthetase Cbf5p and of the RNA binding protein Naf1p during H/ACA snoRNP assembly. *Mol. Cell Biol.* **25**, 3295–3304 (2005).

54. Arigo, J. T., Eyler, D. E., Carroll, K. L. & Corden, J. L. Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. *Mol. Cell* **23**, 841–851 (2006).

55. Thiebaut, M., Kisseleva-Romanova, E., Rougemaille, M., Boulay, J. & Libri, D. Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the Nrd1–Nab3 pathway in genome surveillance. *Mol. Cell* **23**, 853–864 (2006).
References 54 and 55 show that in yeast, the transcription of CUTs terminates by the same mechanism as the transcription of snoRNAs and that this mode of transcription termination is responsible for their rapid degradation by the nuclear exosome.

56. Gudipati, R. K., Villa, T., Boulay, J. & Libri, D. Phosphorylation of the RNA polymerase II C-terminal domain dictates transcription termination choice. *Nature Struct. Mol. Biol.* **15**, 786–794 (2008).

57. Thompson, D. M. & Parker, R. Cytoplasmic decay of intergenic transcripts in *Saccharomyces cerevisiae*. *Mol. Cell Biol.* **27**, 92–101 (2007).

58. Lee, A., Hansen, K. D., Bullard, J., Dudoit, S. & Sherlock, G. Novel low abundance and transient RNAs in yeast revealed by tiling microarrays and ultra high-throughput sequencing are not conserved across closely related yeast species. *PLoS Genet.* **4**, e1000299 (2008).

59. Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl Acad. Sci. USA* (2009).
This paper extends the known repertoire of lincRNAs to ~3,300 by analysing chromatin state maps in human cells. It shows that these RNAs, which are conserved across mammals, are associated with chromatin-modifying complexes, supporting the idea that they are involved in epigenetic regulatory mechanisms.

60. Preker, P., Nielsen, J., Schierup, M. H. & Jensen, T. H. RNA polymerase plays both sides: vivid and bidirectional transcription around and upstream of active promoters. *Cell Cycle* **8**, 1106–1107 (2009).

61. Wang, X. *et al.* Induced ncRNAs allosterically modify RNA-binding proteins in *cis* to inhibit transcription. *Nature* **454**, 126–130 (2008).

62. Morris, K. V., Santoso, S., Turner, A. M., Pastori, C. & Hawkins, P. G. Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. *PLoS Genet.* **4**, e1000258 (2008).

63. Han, J., Kim, D. & Morris, K. V. Promoter-associated RNA is required for RNA-directed transcriptional gene silencing in human cells. *Proc. Natl Acad. Sci. USA* **104**, 12422–12427 (2007).

64. Martianov, I., Ramadass, A., Serra Barros, A., Chow, N. & Akoulitchev, A. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* **445**, 666–670 (2007).

65. Martens, J. A., Wu, P. Y. & Winston, F. Regulation of an intergenic transcript controls adjacent gene transcription in *Saccharomyces cerevisiae*. *Genes Dev.* **19**, 2695–2704 (2005).
A landmark article that, with reference 66, analyses the first example of gene regulation by transcription interference in budding yeast.

66. Martens, J. A., Laprade, L. & Winston, F. Intergenic transcription is required to repress the *Saccharomyces cerevisiae SER3* gene. *Nature* **429**, 571–574 (2004).

67. Jenks, M. H., O'Rourke, T. W. & Reines, D. Properties of an intergenic terminator and start site switch that regulate *IMD2* transcription in yeast. *Mol. Cell Biol.* **28**, 3883–3893 (2008).

68. Kopcewicz, K. A., O'Rourke, T. W. & Reines, D. Metabolic regulation of *IMD2* transcription and an unusual DNA element that generates short transcripts. *Mol. Cell Biol.* **27**, 2821–2829 (2007).

69. Kuehner, J. N. & Brow, D. A. Regulation of a eukaryotic gene by GTP-dependent start site selection and transcription attenuation. *Mol. Cell* **31**, 201–211 (2008).

70. Steinmetz, E. J. *et al.* Genome-wide distribution of yeast RNA polymerase II and its control by Sen1 helicase. *Mol. Cell* **24**, 735–746 (2006).

71. Thiebaut, M. *et al.* Futile cycle of transcription initiation and termination modulates the response to nucleotide shortage in *S. cerevisiae*. *Mol. Cell* **31**, 671–682 (2008).

72. Kwapisz, M. *et al.* Mutations of RNA polymerase II activate key genes of the nucleoside triphosphate biosynthetic pathways. *EMBO J.* **27**, 2411–2421 (2008).

73. Arigo, J. T., Carroll, K. L., Ames, J. M. & Corden, J. L. Regulation of yeast *NRD1* expression by premature transcription termination. *Mol. Cell* **21**, 641–651 (2006).

74. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).

75. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA* **100**, 15776–15781 (2003).

76. Wei, C. L. *et al.* 5′ Long serial analysis of gene expression (LongSAGE) and 3′ LongSAGE for transcriptome characterization and genome annotation. *Proc. Natl Acad. Sci. USA* **101**, 11701–11706 (2004).

77. Perocchi, F., Xu, Z., Clauder-Munster, S. & Steinmetz, L. M. Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. *Nucleic Acids Res.* **35**, e128 (2007).

78. Gingeras, T. R. Origin of phenotypes: genes and transcripts. *Genome Res.* **17**, 682–690 (2007).

79. Vasiljeva, L., Kim, M., Mutschler, H., Buratowski, S. & Meinhart, A. The Nrd1–Nab3–Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. *Nature Struct. Mol. Biol.* **15**, 795–804 (2008).

80. Carroll, K. L., Ghirlando, R., Ames, J. M. & Corden, J. L. Interaction of yeast RNA-binding proteins Nrd1 and Nab3 with RNA polymerase II terminator elements. *RNA* **13**, 361–373 (2007).

Acknowledgements

I thank Frank Fenerbach for critical reading of the manuscript. This work was supported by the Agence Nationale de la Recherche grant 'CUTreg'.

DATABASES

Entrez Gene: http://www.ncbi.nlm.nih.gov/gene
CCND1 | *SER3*
UniProtKB: http://www.uniprot.org
Nab3 | Nrd1 | Sen1

FURTHER INFORMATION

Alain Jacquier's homepage:
http://www.pasteur.fr/recherche/unites/Gim
FANTOM3 project: http://fantom.gsc.riken.jp/3

ALL LINKS ARE ACTIVE IN THE ONLINE PDF