Transcriptional noise and the fidelity of initiation by RNA polymerase II

Kevin Struhl

Eukaryotes transcribe much of their genomes, but little is known about the fidelity of transcriptional initiation by RNA polymerase II *in vivo*. I suggest that ~90% of Pol II initiation events in yeast represent transcriptional noise, and that the specificity of initiation is comparable to that of DNA-binding proteins and other biological processes. This emphasizes the need to develop criteria that distinguish transcriptional noise from transcription with a biological function.

In the classic view of the transcriptome, RNA polymerase initiates transcription from specific sites that are designed to produce functional RNA products. However, biological processes do not work perfectly, and the fidelity of a biological process is defined by the frequency of an 'incorrect' event as compared with the 'correct' event. For eukaryotic Pol II, 'correct' initiation events lead to discrete RNA species, most of which are messenger RNAs. However, very little is known about how often Pol II initiates transcription 'nonspecifically' from incorrect sites and generates 'transcriptional noise' of no biological significance. The concept of transcriptional noise used here is distinct from noise arising from stochastic variations in the expression level of a given gene in individual cells[1].

The issue of transcriptional noise has become increasingly important, because recent studies in a wide range of eukaryotic organisms indicate that there is far more transcription than expected from the classical view of the transcriptome[2,3]. Here, on the basis of experimental observations, including a recent analysis of genome-wide distribution of Pol II[4], I estimate that only 10% of the elongating Pol II molecules in the yeast *Saccharomyces cerevisiae* are engaged in transcription that initiates from conventional promoters and that the remaining 90% of the elongating Pol II molecules represent transcriptional noise. Furthermore, these calculations suggest that the specificity of Pol II initiation (a ~10^4-fold difference between an optimal site and an average genomic site) is comparable to that of sequence-specific DNA-binding proteins and other biological processes considered to be specific.

Focusing on yeast, there are~20,000 molecules of Pol II in cells of *S. cerevisiae*[5]. Of these, ~60% are hyperphosphorylated on the C-terminal domain and associated with chromatin in a salt-stable manner, indicating that they are in the act of transcriptional elongation[6,7]. The locations of these elongating Pol II molecules in the yeast genome can be estimated from chromatin immunoprecipitation experiments that measure Pol II (and TATA-binding protein, or TBP) occupancy *in vivo*. Preinitiation complex formation, experimentally defined by TBP occupancy, is the limiting step at the vast majority of yeast genes[8,9]. Furthermore, Pol II elongation in wild-type cells occurs rapidly upon initiation and is highly processive, so that Pol II density across a gene is nearly constant[4,10]. As a consequence, and unlike RNA measurements that are confounded by differential RNA stabilities, Pol II density in coding regions and TBP association at promoters are directly linked to the level of transcription *in vivo*.

Where are elongating Pol II molecules in the yeast genome? The calculations below assume that the maximal Pol II density is one Pol II molecule per 100 base pairs (bp). This assumption is strongly supported by the physical size of Pol II and the amount of DNA it protects *in vitro*, and by observations that Pol II can initiate transcription at a maximal rate of once every 5 seconds[11] and elongates at a rate of 20–30 bp per second[10]. Maximal TBP and Pol II occupancy, and hence transcriptional activity, is observed at several heat-shock or galactose-induced genes[8]. However, such highly active genes are expressed only under specific stress conditions that induce transcription of a relatively small number of genes.

Under optimal conditions, the most active genes are the 138 encoding ribosomal proteins and ~50 genes encoding other proteins such as glycolytic enzymes[12]. TBP and Pol II occupancy at such highly expressed genes is only ~25% of the maximal level observed with the best stress-induced genes[8], and this corresponds to one Pol II molecule for every 400 bp. Assuming an average mRNA-coding region of 1 kilobase, there are approximately 500 Pol II molecules associated with these highly active genes. RNA levels of the remaining 5,000–6,000 standard yeast genes vary over a wide range, but together represent ~50% of all Pol II transcription[13], which gives an average value for a standard yeast gene ~1% of that observed for highly active genes. Wild-type cells also express ~1,500 antisense transcripts and other RNAs with complex structures[14], but these are generally expressed at lower levels than standard yeast genes and hence contribute minimally to overall amounts of RNA.

Together, these considerations suggest that, in a snapshot of an average cell, there are ~1,000–1,500 Pol II molecules associated with genes engaged in the production of mRNA.

Kevin Struhl is in the Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA.
e-mail: kevin@hms.harvard.edu

This represents ~10% of all elongating Pol II molecules, and it is unlikely that this number is higher than 20%. Importantly this conclusion is supported by the recent genome-wide analysis of Pol II occupancy which shows that, with the exception of well-transcribed genes, Pol II associates to a comparable extent with the vast majority of the yeast genome, including numerous regions conventionally considered to be transcriptionally inert[4]. This apparently near-constant level of Pol II association does not reflect experimental background or nonspecific association, because it is approximately ten-fold above that observed at certain repressed loci. Lastly, as methylation of histone H3 is directly linked to elongating Pol II and the associated Paf1 complex[15], the presence of elongating Pol II throughout the genome nicely explains why histone methylation occurs at regions conventionally thought to be transcriptionally inert[16].

Given that ~90% of all elongating Pol II molecules do not produce mRNAs corresponding to standard genes, where does this 'junk' transcription initiate? The low efficiency and extent of transcriptional read-through suggest that only a small amount of junk transcription represents imperfect termination of RNAs initiating from conventional promoters. Some junk transcription initiates at cryptic promoters that are located in many mRNA-coding regions. Such cryptic promoters function in wild-type cells[17], and they can generate RNA transcripts at levels comparable to and even higher than those of typical yeast mRNAs in strains lacking a variety of protein complexes that modify chromatin structure[18]. Junk transcription can also initiate at inappropriate positions within intergenic regions, particularly because most intergenic regions are markedly depleted for histones and hence preferentially accessible to the transcription machinery[19,20]. Consistent with this, RNA mapping experiments often reveal many 5′ ends over regions that extend beyond those appropriate for producing the protein encoded by the gene[17]. Importantly, junk transcription generated by any of these mechanisms could occur on either strand and hence be responsible for some (and perhaps many) of the antisense and other noncoding RNAs detected in eukaryotic organisms[2,14].

Why is the majority of junk transcription not reflected in standard RNA analyses? It is highly likely that most RNAs initiated at inappropriate positions are unstable and rapidly degraded by the nonsense-mediated decay[21], exosome[22,23] or other pathways. Indeed, mutations that inactivate these degradation mechanisms result in increased levels of junk transcription. In addition, some junk transcription may be analogous to short 'abortive transcripts' produced by RNA polymerases *in vitro*. Such abortive transcripts would be difficult to detect, even though Pol II would associate near the aberrant initiation site. Importantly, as these degradation mechanisms are unlikely to act instantaneously or be fully efficient, some junk RNA will persist in wild-type cells.

On the basis of the above considerations, the biological specificity, and hence the fidelity, of the Pol II initiation machinery can be estimated. There are ~2,000 nucleotides or potential mRNA initiation sites at an average yeast gene (~6,000 yeast genes in a 12-megabase genome). Given ~10% correct initiation events, initiation from a single correct site in an average gene is ~200-fold more likely than from an average position, although incorrect initiation will not occur equally at all sites. Initiation from typical yeast promoters is ~1% of the maximal level, indicating a ~10^4-fold specificity difference between initiation from a maximally active promoter and from an average sequence. Interestingly, this specificity factor is comparable to that observed for typical sequence-specific DNA-binding proteins binding a high-affinity site, as compared with a nonspecific site. This level of fidelity is also similar to the frequency at which incorrect nucleotides or amino acids are incorporated into growing polymers of RNA and protein, although it is far below the 10^{-8}-fold specificity seen for DNA replication, which depends on proofreading mechanisms. Thus, Pol II initiation from correct promoters seems to be rather error-prone, with a level of fidelity that is roughly comparable to those of other biological processes that are considered to be specific.

As surprisingly high proportions of eukaryotic genomes are transcribed, with numerous noncoding and antisense RNAs being produced[2,14], it is essential to develop criteria that distinguish transcriptional noise from transcription with a biological function. To put it differently, the existence of an RNA does not prove its biological significance, and conversely, the high level of transcriptional noise does not disprove the significance of the observed RNAs. In regard to this, it is important to note that biologically relevant transcription might not necessarily give rise to 'functional' RNA products, but rather generate specific chromatin domains (for example, domains containing methylated histones) that regulate transcription of overlapping or flanking genes. In a related vein, enhancers that regulate distal protein-coding genes might generate junk transcripts in the vicinity of the enhancer; in this case, the junk RNA would essentially represent a marker for the nearby enhancer.

A major difficulty in distinguishing between transcriptional noise and functional transcription is that both are produced by the same Pol II machinery and hence share many characteristics. First, given the basic properties of the Pol II machinery, junk RNAs will have 5′ methyl caps, and they will often have defined 3′ ends that are polyadenylated, in which case they will appear as RNA species of discrete size. Second, although the average level of transcriptional noise and fidelity of Pol II initiation can be estimated, the level of initiation from a given incorrect site could vary over a large range. By analogy with sequence-specific DNA-binding proteins, the level of junk or nonspecific initiation will be influenced by how closely the DNA sequence resembles bona fide core promoter (for example, TATA or initiator sequences) and enhancer elements. Third, as is the case for correct initiation sites and biologically significant RNAs, transcriptional regulatory proteins bound in the general vicinity of an incorrect initiation site can regulate the level of junk RNAs as a function of cell type or environmental conditions. Fourth, as Pol II elongation is mechanistically linked with histone depletion[24,25], transcription from correct initiation sites will reduce nucleosome density throughout the entire transcribed region, thereby increasing DNA accessibility and hence the likelihood of incorrect initiation. As a consequence, transcripts arising from incorrect sense or antisense initiation within coding regions will often be coregulated with the correct transcript. These considerations suggest that the abundance, discrete size and regulatory properties of RNAs observed *in vivo* are inadequate criteria to distinguish biological significance from transcriptional noise.

In principle, evolutionary conservation is a useful criterion, as by definition, conservation is a measure of some kind of biological function. As a group, noncoding RNAs in mammalian cells are more conserved than expected by chance. However, this indicates only that some noncoding RNAs are biologically meaningful, and it leaves open the possibility that others (perhaps even the majority) of noncoding RNAs are transcriptional noise. Conversely, there might be little evolutionary conservation if the true biological function associated with an RNA is not the product itself but rather the act of transcription (and associated chromatin modifications), or if the RNA is the byproduct of a nearby enhancer that controls distal genes. Thus, the absence of evolutionary conservation is not a reliable indicator of transcriptional noise. Lastly, it would be useful to have an experimental measurement of transcriptional noise that is not confounded by the possibility that the observed RNAs are biologically significant. By definition, transcription

(and protein binding) from evolutionarily unrelated DNA is noise, so analysis of cells containing large regions of evolutionarily unrelated DNA (for example, *Escherichia coli* DNA in human cells) will be of interest.

It is clear that eukaryotic organisms express many more biologically significant RNAs than were expected according to the classical view of the transcriptome. However, the relative proportions of biologically significant noncoding RNAs and transcriptional noise are unknown. From a whole-genome perspective, it will be a challenge to distinguish the RNAs that are in some way meaningful to the organism from those that arise from the imperfect fidelity of the Pol II transcription machinery.

ACKNOWLEDGMENTS

I thank R. Kornberg, Z. Moqtaderi, M. Oettinger, A. Sandelin, J.Svejstrup & J. Wade for helpful discussions and comments on the manuscript.

COMPETING INTERESTS STATEMENT

The author declares that he has no competing financial interests.

1. Raser, J.M. & O'Shea, E.K. *Science* **309**, 2010–2013 (2005).
2. Willingham, A.T. & Gingeras, T.R. *Cell* **125**, 1215–1220 (2006).
3. Carninci, P. *et al. Science* **309**, 1559–1563 (2005).
4. Steinmetz, E.J. *et al. Mol. Cell* **24**, 735–746 (2006).
5. Borggrefe, T. *et al. J. Biol. Chem.* **276**, 47150–47153 (2001).
6. Svejstrup, J.Q. *et al. Proc. Natl. Acad. Sci. USA* **94**, 6075–6078 (1997).
7. Otero, G. *et al. Mol. Cell* **3**, 109–118 (1999).
8. Kuras, L. & Struhl, K. *Nature* **399**, 609–612 (1999).
9. Li, X.-Y. *et al. Nature* **399**, 605–609 (1999).
10. Mason, P.B. & Struhl, K. *Mol. Cell* **17**, 831–840 (2005).
11. Iyer, V. & Struhl, K. *Proc. Natl. Acad. Sci. USA* **93**, 5208–5212 (1996).
12. Holstege, F.C. *et al. Cell* **95**, 717–728 (1998).
13. Warner, J.R. *Trends Biochem. Sci.* **24**, 437–440 (1999).
14. David, L. *et al. Proc. Natl. Acad. Sci. USA* **103**, 5320–5325 (2006).
15. Hampsey, M. & Reinberg, D. *Cell* **113**, 429–432 (2003).
16. Ng, H.H., Dole, S. & Struhl, K. *J. Biol. Chem.* **278**, 33625–33628 (2003).
17. Miura, F. *et al. Proc. Natl. Acad. Sci. USA* **103**, 17846–17851 (2006).
18. Workman, J.L. *Genes Dev.* **20**, 2009–2017 (2006).
19. Sekinger, E.A., Moqtaderi, Z. & Struhl, K. *Mol. Cell* **18**, 735–748 (2005).
20. Yuan, G.-C. *et al. Science* **309**, 626–630 (2005).
21. Baker, K.E. & Parker, R. *Curr. Opin. Cell Biol.* **16**, 293–299 (2004).
22. LaCava, J. *et al. Cell* **121**, 713–724 (2005).
23. Wyers, F. *et al. Cell* **121**, 725–737 (2005).
24. Schwabish, M.A. & Struhl, K. *Mol. Cell. Biol.* **24**, 10111–10117 (2004).
25. Kristjuhan, A. & Svejstrup, J.Q. *EMBO J.* **23**, 4243–4252 (2004).