Divergent Transcription: A Driving Force for New Gene Origination?

Xuebing Wu1,2 and Phillip A. Sharp1,3,*

1David H. Koch Institute for Integrative Cancer Research
2Computational and Systems Biology Graduate Program
3Department of Biology
Massachusetts Institute of Technology, Cambridge, MA 02139, USA
*Correspondence: sharppa@mit.edu
http://dx.doi.org/10.1016/j.cell.2013.10.048

The mammalian genome is extensively transcribed, a large fraction of which is divergent transcription from promoters and enhancers that is tightly coupled with active gene transcription. Here, we propose that divergent transcription may shape the evolution of the genome by new gene origination.

Widespread Divergent Transcription

The vast majority of the human genome, including half of the region outside of known genes, is transcribed (Djebali et al., 2012). However, most intergenic transcription activity produces short and unstable noncoding transcripts whose abundances are usually an order of magnitude lower than those from typical protein-coding genes. Except for a few well-studied cases (see review in Gutmann and Rinn, 2012; Lee, 2012; Mercer et al., 2009; Ponting et al., 2009; Rinn and Chang, 2012; Ullitsky and Bartel, 2013; Wang and Chang, 2011; Wei et al., 2011; Wilusz et al., 2009), it’s unclear whether most intergenic transcription is regulated or has cellular function.

Recent evidence has shown that most intergenic transcription occurs near or is associated with gene transcription, such as transcription from promoter and enhancer regions (Sigova et al., 2013). The majority of mammalian promoters direct transcription initiation on both sides with opposite orientations, a phenomenon known as divergent transcription (Core et al., 2008; Preker et al., 2008; Seila et al., 2008). Divergent transcription generates upstream antisense RNAs (uaRNAs, or PROMPTs, promoter upstream transcripts) near the 5’ end of genes that are typically short (50–2,000 nucleotides) and relatively unstable (Flynn et al., 2011; Ntini et al., 2013; Preker et al., 2008, 2011). Similar divergent transcription also occurs at distal enhancer regions, giving rise to RNAs termed enhancer RNAs (eRNAs) (Kim et al., 2010; De Santa et al., 2010). In mouse and human embryonic stem (ES) cells, most long noncoding RNAs (lncRNAs, longer than 100 nucleotides) are associated with protein-coding genes, including ~50% as uaRNAs and ~20% as eRNAs (Sigova et al., 2013). These observations suggest that divergent transcription from promoters and enhancers of protein-coding genes is the major source of intergenic transcription in ES cells.

In the textbook model of a eukaryotic promoter, the directionality is set by the arrangement of an upstream cis-element region followed by a core promoter (Figure 1A). The cis-elements are bound by sequence-specific transcription factors, whereas the core promoter is bound by TATA-binding protein (TBP) and other factors that recruit the core transcription machinery. Most mammalian promoters lack a TATA element (TATA-less) and are CpG rich (Sandelin et al., 2007). For these promoters, TBP is recruited through sequence-specific transcription factors such as Sp1 that bind CpG-rich sequences and components of the TFIID complex that have little sequence specificity. Thus, in the absence of strong TATA elements such as for CpG island promoters, TBP-complexes are recruited on both sides of the transcription factors to form preinitiation complexes in both orientations (Figure 1B). This model is supported by the observation that divergent transcription occurs at most promoters that are associated with CpG islands in mammals, whereas promoters with TATA elements in mammals and worm are associated with unidirectional transcription (Core et al., 2008; Kruesi et al., 2013). In addition, divergent transcription is less common in Drosophila where CpG islands are rare (Core et al., 2012). Since transcription factors with chromatin remodeling potential and transcription activation domains also bind at enhancer sites, it is not surprising that these are also sites of divergent transcription. In fact, promoters and enhancers have many properties in common, and it has been shown recently that many intragenic enhancers can act as alternative promoters producing tissue-specific IncRNAs (Kowalczyk et al., 2012).

The U1-PAS Axis and Gene Maturation

Promoter-proximal noncoding transcription in both yeast and mammals has been shown to be suppressed at the chromatin level, including nucleosome remodeling (Whitehouse et al., 2007), histone deacetylation (Churchman and Weissman, 2011), and gene loop formation (Tan-Wong et al., 2012). We and others recently found that in mammals promoter upstream antisense transcription is frequently terminated due to cleavage of the nascent RNA by the same process responsible for the generation of the poly A tract at the 3’ ends of genes (Almada et al., 2013; Ntini et al., 2013). In both cases, the primary signal directing this process is the poly(A) signal (PAS) motif, AAUAAA or similar (Proudfoot, 2011). Pol II terminates transcription
within several kb after such cleavage (Anamika et al., 2012; Richard and Manley, 2009). Computational analysis showed that relative to the 5′ end of the sense regions, PAS motifs are enriched, whereas potential U1 snRNP-binding sites, or 5′ splice site-like sequences, are depleted in the upstream antisense regions. The binding of U1 snRNP is known to suppress PAS-directed cleavage over regions of thousand nucleotides downstream (Berg et al., 2012; Kaida et al., 2010). Thus, the bias in the distribution of U1 snRNP-binding sites and PAS promotes expression of full-length mRNAs by suppressing premature cleavage and polyadenylation but favors early termination of uaRNAs. This conclusion is strongly supported by the finding that inhibition of U1 snRNP dramatically increased termination and polyadenylation of sense-oriented transcripts in the gene region (Almada et al., 2013).

If the U1-PAS axis defines the length of a transcribed region, then it might be expected that for a typical protein-coding gene (~20 kb) to evolve from intergenic noncoding DNA would involve strengthening of the U1-PAS axis by gaining U1 sites and losing PAS in the sense orientation. Examining the distributions of U1 and PAS sites in bidirectional promoters involving UCSC-annotated mRNA-mRNA, mRNA-IncRNA, and mRNA-uaRNA pairs, we found that IncRNAs showed properties resembling intermediates between mRNA genes and uaRNA regions in terms of the density of U1 sites and PAS sites (Almada et al., 2013). That is, the density of PAS decreases from regions producing uaRNA to IncRNA to mRNA, whereas U1 sites show the opposite trend, consistent with the differences in the length and abundance of these transcripts. We also studied the evolution of the U1-PAS axis in vertebrates, and found that older genes exhibit progressive gain of U1 sites and loss of PAS sites at their 5′ ends. Together these observations suggest that strengthening of the U1-PAS axis may be associated with the origination and maturation of genes.

De Novo Gene Origination from Divergent Transcription

Below we propose a model (Figure 2) arguing that the act of transcription in germ cells strengthens the U1-PAS axis in the upstream antisense region of an active gene, or the associated enhancer regions, creating a feedback loop amplifying transcription activity, which eventually may drive origination of a new antisense-oriented gene (Figure 3).

One consequence of transcription is that it can cause mutations, especially on the coding (nontranscribed) strand. During transcription, transient R loops can be formed behind the transcribing RNA polymerase II, exposing the coding strand as single-stranded DNA, whereas the noncoding strand is base paired with and thus protected by the nascent RNA (Aguilera and García-Muse, 2012). The lack of splicing signals in the divergent transcript also makes it more vulnerable to mutations toward G and T bases (A), which increases the chance of gaining GT-rich sequences such as U1-binding site (5′ splice site [5′ SS]) and also increases the chance of losing A-rich sequences such as PAS, which terminates transcription (B). U1 binding can enhance transcription through promoting transcription initiation and reinitiation, and also inhibiting the usage of nearby PAS.

De Novo Gene Origination from Divergent Transcription

Below we propose a model (Figure 2) arguing that the act of transcription in germ cells strengthens the U1-PAS axis in the upstream antisense region of an active gene, or the associated enhancer regions, creating a feedback loop amplifying transcription activity, which eventually may drive origination of a new antisense-oriented gene (Figure 3).

One consequence of transcription is that it can cause mutations, especially on the coding (nontranscribed) strand. During transcription, transient R loops can be formed behind the transcribing RNA polymerase II, exposing the coding strand as single-stranded DNA, whereas the noncoding strand is base paired with and thus protected by the nascent RNA (Aguilera and García-Muse, 2012). The lack of splicing signals in the divergent transcript also makes it more vulnerable to mutations toward G and T bases (A), which increases the chance of gaining GT-rich sequences such as U1-binding site (5′ splice site [5′ SS]) and also increases the chance of losing A-rich sequences such as PAS, which terminates transcription (B). U1 binding can enhance transcription through promoting transcription initiation and reinitiation, and also inhibiting the usage of nearby PAS.

De Novo Gene Origination from Divergent Transcription

Below we propose a model (Figure 2) arguing that the act of transcription in germ cells strengthens the U1-PAS axis in the upstream antisense region of an active gene, or the associated enhancer regions, creating a feedback loop amplifying transcription activity, which eventually may drive origination of a new antisense-oriented gene (Figure 3).

One consequence of transcription is that it can cause mutations, especially on the coding (nontranscribed) strand. During transcription, transient R loops can be formed behind the transcribing RNA polymerase II, exposing the coding strand as single-stranded DNA, whereas the noncoding strand is base paired with and thus protected by the nascent RNA (Aguilera and García-Muse, 2012). The lack of splicing signals in the divergent transcript also makes it more vulnerable to mutations toward G and T bases (A), which increases the chance of gaining GT-rich sequences such as U1-binding site (5′ splice site [5′ SS]) and also increases the chance of losing A-rich sequences such as PAS, which terminates transcription (B). U1 binding can enhance transcription through promoting transcription initiation and reinitiation, and also inhibiting the usage of nearby PAS.
Figure 3. Divergent Transcription Drives New Gene Origination
(A–F) De novo protein-coding gene origination (A–E), and gene duplication or translocation (F). (A) Divergent transcription of a gene (right dark block) generates divergent noncoding RNA (ncRNA) in the upstream antisense direction, which is terminated by PAS-dependent mechanism (PAS: red bars). (B) Transcription increases G and T frequency on the coding strand, thus increases the chance of encoding a U1 site (blue bar) that suppress a downstream PAS (PAS1), favoring the usage of a downstream PAS (PAS2). (C) Increase in G+T content also increases the chance of losing PAS sites (PAS2) that activates a further downstream site (PAS3) and extends the transcribed region. (D) The longer transcript acquires splicing signals, which makes it more stable and exported to the cytoplasm. (E) The longer transcript encodes a short ORF and the resulting short peptide is selected and fixed in the population and becomes a new protein-coding gene. (F) Gene A is translocated or duplicated into the promoter upstream antisense region of gene B, and evolves into a new gene A'. Thin and thick blocks represent transcribed noncoding and coding regions, respectively.

transcription-coupled repair, and somatic hypermutation pathways in germ cell-transcribed genes, in the absence of selection (Green et al., 2003; McVicker and Green, 2010; Polak and Arndt, 2008).

Accumulation of G and T content on the coding strand will strengthen the U1-PAS axis (Figure 2). A-rich sequences such as PAS (AATAAA) are likely to be lost when the genomic DNA accumulates G and T. In contrast, G+T-rich sequences, such as U1 snRNP-binding sites (e.g., resembling 5’ splice sites, G/GTAAGT and G/GTGAGT), are likely to emerge in these regions. Since promoter-proximal PAS reduces transcriptional activity (Andersen et al., 2012), the loss of PAS and gain of U1 sites should contribute to lengthening of the transcribed region as well as its more robust transcription. The gain of U1 sites could also enhance transcription by recruiting basal transcription initiation factors (Damgaard et al., 2008; Furger et al., 2002; Kwek et al., 2002) or elongation factors (Fong and Zhou, 2001). Therefore a positive feedback loop is formed: active transcription causes the coding strand to accumulate sequence changes favoring higher transcription activity.

As noted above, strengthening of the U1-PAS axis also favors extension of the transcribed region. Being longer gives the transcript several advantages: by chance longer RNAs are more likely to contain additional splicing signals such as a 3’ splice site to become spliced, or binding sites for splicing-independent nuclear export factors, thus escaping nuclear exosome degradation by packaging and exporting to cytoplasm (Nott et al., 2003; Singh et al., 2012). Longer RNAs are also more likely to carry an open reading frame, either generated de novo or by incorporation of gene remnants.

Once in the cytoplasm, the RNA should at some frequency be translated into short polypeptides due to widespread translational activity (Carvunis et al., 2012). Some of the polypeptides may provide advantage to the organism and become fixed in the population, thereby forming a new gene.

Accelerating Other New Gene Origination Processes
In addition to de novo gene origination, the model described above also facilitates new gene origination via other mechanisms in regions of divergent transcription. Tandem duplication, retroposition, and recombination of existing genes or gene fragments are the major mechanisms for new gene origination (Chen et al., 2013; Long et al., 2013). Most duplicated genes or gene fragments are silenced due to the lack of required elements such as a promoter. In contrast, genes or gene fragments inserted into regions of divergent transcription, such as upstream of a promoter or flanking an enhancer, will be transcribed, likely under different regulation than prior to their insertion, and thus could evolve to carry out functions different than the original gene. In support of this, a recent survey of human and mouse genes evolved from “domesticated” transposons (Kalitsis and Saffery, 2009) showed that a significant proportion of them are located in bidirectional promoters. Promoter upstream regions also preferentially accumulate transposable elements, which can carry 5’ splice site sequences that may accelerate the process of new gene origination (Gotea et al., 2013).

New Gene Origination from Enhancers
Similar to promoters, enhancers are also divergently transcribed, and, as a result, new genes might originate at enhancer regions through the same mechanism described above. The possibility of enhancer-derived new genes has not been previously discussed. Manual inspection of a list of 24 hominoid-specific de novo protein-coding genes (Xie et al., 2012) revealed that MYEOV (myeloma overexpressed), a gene implicated in various types of cancer (Janssen et al., 2000, 2002; Leyden et al., 2006; Moss et al., 2006), is likely derived from an intergenic enhancer in mouse. The mouse syntenic region of MYEOV is within a 5 kb region about 100 kb away from any gene, but covered by intensive H3K4me1 marks, diagnostic of an enhancer, and
positive for Mediator binding in mouse ES cells, as well as nascent transcription signals (GRO-seq) indicating divergent transcription, all indicating this region is an active enhancer in mouse ES cells. Further analysis is needed to firmly establish the role of enhancer transcription in the origination of the MYEOV gene. For example, it will be interesting to examine the evolutionary dynamics of the spatial and functional relationship between the enhancer/MYEOV locus and the corresponding target gene.

**Predictions and Supporting Evidence**

A recent comparative analysis of human-mouse gene annotations detected over a thousand IncRNAs annotated in the upstream antisense region of human genes, whereas IncRNAs divergent from the corresponding mouse protein-coding genes could not be detected (Gotea et al., 2013). This observation suggests that promoter divergent transcription could be capable of generating a large number of primate-specific transcripts. Another study (Xie et al., 2012), identified 24 hominoid-specific de novo protein-coding genes in human, five of which derive from bidirectional promoters ($p < 0.01$, compared to shuffled gene positions), confirming promoter divergent transcription as an important source of de novo gene origination, and enhancer transcription may drive the origination of other new genes, as noted above.

An important feature of genes originated in the proposed model is that both the new gene and the ancestral gene are likely to be expressed in germ cells. This is because for the transcription-induced G and T bias to accumulate and spread in a population, these mutations should occur in germ cells. A prediction of the model is that new genes are preferentially expressed in germ cells, or tissues with high fraction of germ cells. Consistent with this, previous reports showed that lineage-specific genes in human, fly, and zebrafish genomes are preferentially expressed in reproductive organs or tissues, such as testis (Clark et al., 2007; Levine et al., 2006; Tay et al., 2009; Yang et al., 2013). Moreover, divergent gene pairs in the human genome are enriched for housekeeping genes, such as DNA repair and DNA replication genes (Adachi and Lieber, 2002) that are actively transcribed in germ cells. In addition, the strand bias of G and T content correlates with germ cell but not somatic tissue gene expression levels (Majewski, 2003).

The model could explain the origin of divergent protein-coding gene pairs separated by less than 1 kb (usually less), which account for 10% of human protein-coding genes (Adachi and Lieber, 2002; Li et al., 2006; Piontkivska et al., 2009; Trinklein et al., 2004; Wakano et al., 2012; Xu et al., 2012), far higher proportion than would be expected if genes were randomly distributed in the genome. The model proposed here provides a natural explanation for the evolutionary origin of these gene pairs. It is likely that many more genes originated from divergent transcription, with the bidirectional organization having been disrupted by transposon insertion, recombination, or other genome rearrangement events. The model also predicts that divergent gene pairs commonly have unrelated functions, although they frequently might share coexpression. Except for a few cases, such as histone gene pairs and collagen gene pairs that are likely results of tandem duplication, the majority of divergent gene pairs in the human genome do not share higher functional similarity compared to random gene pairs (Li et al., 2006; Xu et al., 2012). For example, 35 of the 105 annotated DNA repair genes have bidirectional promoters, making DNA repair the most over-represented pathway for genes involved in bidirectional promoters, yet all 35 DNA repair genes are paired with non-DNA repair genes (Xu et al., 2012). Similarly, genes coding subunits of protein complexes are enriched in bidirectional pairs in human, yet none of these pairs code for two subunits of the same complex (Li et al., 2006). A similar observation has been reported for yeast and is consistent with the argument that the bidirectional conformation reduces expression noise and is not strongly selected for share functionality (Wang et al., 2011).

The lack of functional relatedness is also illustrated by the parallel evolution of bidirectional promoters of RecQ helicases (Piontkivska et al., 2009). The five RecQ paralogs were duplicated early during metazoan evolution, yet all evolve to have divergent partners in human. However, these partner genes showed no functional or sequence similarity with each other (Piontkivska et al., 2009), suggesting parallel and independent origination of new genes from all five promoters.

**Impact on Genome Organization and Evolution**

Divergent transcription likely facilitates the rearrangement events that reshape the genome and also introduces unique features into genome organization, including the sharing of promoters, physical linkage in three-dimensional space, and coexpression of distal genes.

Although vertebrates share most of their genes, the genomic position and orientation of specific genes differ significantly due to genome rearrangement events, such as translocation, recombination, and duplication followed by the loss of the original copy. The survival of the gene or gene fragments at the new position can be facilitated by divergent transcription as discussed above. The role of divergent transcription in preserving the function of the new gene copy is likely significant, given that translocation preferentially occurs near active promoters (Chiarle et al., 2011; Klein et al., 2011). The correlation between transcription and translocation could potentially increase the chance that the translocated gene is still expressed and thus functional, therefore reducing the cost of translocation. For example, although ~40% of human protein-coding genes can be traced back to fish, fewer than 7% (83/1,262) of human bidirectional gene pairs are also bidirectional in the fish genome (Li et al., 2006), suggesting that most human bidirectional gene pairs formed with young genes, or by bringing together old genes through translocation facilitated by divergent transcription.

In addition to bidirectional organization, spatial and functional coupling between distal gene pairs would be introduced through new gene origination from enhancer transcription. Due to the tight coupling between gene transcription and enhancer transcription, an enhancer-derived new gene will share a significant coexpression pattern with the old gene, despite the distance in the linear genome. Such coupled transcription of distal gene pairs brought together by chromatin interactions could contribute to the formation of transcription factories, nuclear foci where multiple genes are transcribed together without the requirement of shared function (Edelman and Fraser, 2012; Sutherland and Bickmore, 2006). The existence of transcription...
factories has been supported by increasing evidence, including in vivo live imaging (Ghamari et al., 2013) and chromatin interaction mapping (Li et al., 2012). These are probably related to super-enhancers where many genes that are coordinately expressed are associated with a common enhancer region (Lovén et al., 2013; Whyte et al., 2013). Overlaying comparative genomics analysis onto high-throughput chromatin interaction mapping data across multiple species (Dixon et al., 2012; Li et al., 2012) may help to reveal the evolutionary origin of transcription factories.

Conclusions
In conclusion, we propose that divergent transcription at promoters and enhancers results in changes of the transcribed DNA sequences that over evolutionary time drive new gene origin in the transcribed regions. Although the models proposed here are consistent with significant available data, systematic tests of these models await further advances such as in-depth characterization of additional genomes and experiments designed to test specific hypotheses. Over evolutionary times, genes formed through divergent transcription can be shuffled to other locations losing their evolutionary context. We envision future studies will uncover more functional surprises from divergent transcription, and illuminate how intergenic transcription is integrated into the cellular transcriptome.

ACKNOWLEDGMENTS
We thank all the Sharp lab members, especially Andrea Kriz, Anthony Chiu, Albert Almada, Mohini Jangi, and Jesse Zamudio for comments and Christopher Burge, Qifang Liu, Jianrong Wang, and Jeremy Wilusz for critical reading of the manuscript. This work was supported by United States Public Health Service grants R01-GM34277 and R01-CA133404 from the National Institutes of Health (P.A.S.) and partially by the Koch Institute Support (core) grant P30-CA14051 from the National Cancer Institute. X.W. is a Howard Hughes Medical Institute International Student Research fellow.

REFERENCES

et al. (2012). Intragenic enhancers act as alternative promoters. Mol. Cell
Stanley, J.A., McGowan, S.J., De Gobbi, M., Hosseini, M., Vernimmen, D.,
Kowalczyk, M.S., Hughes, J.R., Garrick, D., Lynch, M.D., Sharpe, J.A., Sloane-
mammalian genomes. BMC Genomics
Condensin controls recruitment of RNA polymerase II to achieve nematode
X-chromosome dosage compensation. Elife
Leyden, J., Murray, D., Moss, A., Arumuguma, M., Doyle, E., McEntee, G.,
Proudfoot, N.J. (2011). Ending the message: poly(A) signals then and now.