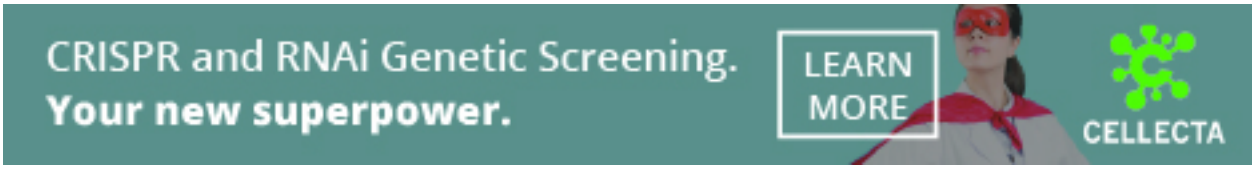
Evidence for variation in abundance of antisense transcripts between multicellular animals but no relationship between antisense transcription and organismic complexity

Miao Sun, Laurence D. Hurst, Gordon G. Carmichael, et al.

Genome Res. 2006 16: 922-933
Access the most recent version at doi:10.1101/gr.5210006



Letter

Evidence for variation in abundance of antisense transcripts between multicellular animals but no relationship between antisense transcription and organismic complexity

Miao Sun,[1] Laurence D. Hurst,[2,4] Gordon G. Carmichael,[3] and Jianjun Chen[1,4]

[1]*Section of Hematology/Oncology, Department of Medicine, University of Chicago, Chicago, Illinois 60637, USA;* [2]*Department of Biology and Biochemistry, University of Bath, Somerset, BA2 7AY, United Kingdom;* [3]*Department of Genetics and Developmental Biology, University of Connecticut Health Center, Farmington, Connecticut 06030-3301, USA*

Given that humans have about the same number of genes as mice and not so many more than worm, what makes us more complex? Antisense transcripts are implicated in many aspects of gene regulation. Is there a functional connection between antisense transcription and organismic complexity, that is, is antisense regulation especially prevalent in humans? We used the same robust protocol to identify antisense transcripts in humans and five other metazoan genomes (mouse, rat, chicken, fruit fly, and nematode), and found that the estimated proportions of genes involved in antisense transcription are highly sensitive to the number of transcripts included in the analysis. By controlling for transcript abundance, we find that the probability that any given transcript is putatively involved in sense–antisense regulation is no higher in humans than in other vertebrates but appears unusually high in flies and especially low in nematodes. Similarly, there is no evidence that the proportion of sense–antisense transcripts is especially higher in humans than other vertebrates in a given subset of transcript sequences such as mRNAs, coding sequences, conserved, or nonconserved transcripts. Although antisense transcription might be enriched in mammalian brains compared with nonbrain tissues, it is no more enriched in human brain than in mouse brain. Overall, therefore, while we see striking variation between multicellular animals in the abundance of antisense transcripts, there is no evidence for a link between antisense transcription and organismic complexity. More particularly, we see no evidence that humans are in any way unusual among the vertebrates in this regard. Instead, our results suggest that antisense transcription might be prevalent in almost all metazoan genomes, nematodes being an unexplained exception.

[Supplemental material is available online at www.genome.org.]

While it appears intuitively reasonable to suppose that organisms differ in their "complexity," this apparently simple assertion begs numerous further questions. One issue is definitional, that is, what is complexity, and how might it be measured? Organismic complexity, it is argued, is a compound term with at least four types being distinguished: nonhierarchical morphological, nonhierarchical developmental, hierarchical morphological, and hierarchical developmental (McShea 1996). According to the complexity in differentiated cell, tissue, and organ types, with or without developed limbs and nervous systems, as well as language ability, and so on, it is a common notion that humans are the most complex species, while mammals are more complex than primitive vertebrates, and vertebrates are more complex than invertebrates.

Assuming that humans are, in some sense, more complex than mice and flies, the second issue is then biological. What factors underlie the differences in complexity? Following the discovery of the remarkably small number of protein-coding genes in the human genome (Lander et al. 2001; Venter et al. 2001), it was suggested that complexity might arise from alternative splicing (Lander et al. 2001; Venter et al. 2001; Modrek and Lee 2002; Kim et al. 2004b). While no doubt this is true in part, it is remarkable that across a wide span of taxa, there is little difference in the abundance of alternative splicing (Brett et al. 2002; Harrington et al. 2004). What else might underpin the differences in complexity? It has been suggested that the basis of eukaryotic complexity and phenotypic variation may lie primarily in a control architecture composed of a highly parallel system of *trans*-acting RNAs that relay state information required for the coordination and modulation of gene expression (Mattick 2001, 2004, 2005) and that organismal complexity arises from progressively more elaborate regulation of gene expression (Levine and Tjian 2003). Antisense regulation might be a good candidate for such gene control (Mattick 2004, 2005). Indeed, antisense regulation has been suggested to be important to all species from bacteria to humans (Merino et al. 1994; Kumar and Carmichael 1998; Vanhee-Brossollet and Vaquero 1998; Carmichael 2003; Kramer et al. 2003; Lavorgna et al. 2004; Wang and Carmichael 2004).

The majority of natural antisense transcripts are *cis*-encoded ones, which are transcribed from the opposite strand of the same genomic loci from their sense counterparts (Vanhee-Brossollet and Vaquero 1998). Besides the *cis*-encoded antisense transcripts, there is another kind of antisense called *trans*-encoded antisense such as microRNAs (Ambros 2004; Bartel 2004), which are tran-

[4]Corresponding authors.
E-mail jchen@medicine.bsd.uchicago.edu; fax (773) 702-3002.
E-mail l.d.hurst@bath.ac.uk; fax 44 (0)1225-386779.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5210006.

16:922–933 ©2006 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

scribed from different genomic loci than their target genes (Kumar and Carmichael 1998; Vanhee-Brossollet and Vaquero 1998; Chen et al. 2004). To date, 250–300 microRNAs have been identified or predicted in humans and mice, respectively (http://microrna.sanger.ac.uk/sequences/index.shtml; December 2005), and most of them are conserved between these two species; in addition, their putative target sites might also be conserved between mammals (Lewis et al. 2005). Although this might be due to the fact that the conservation across species is often the very reason why they are detected by computational methods, there is no evidence yet that microRNA regulation is much more prevalent in humans than in mice, and thus we focus our study on *cis*-encoded antisense transcription alone. All the antisense transcription/transcripts or sense–antisense pairs mentioned in this study hence refer to *cis*-encoded ones.

Antisense regulation (mediated by *cis*-encoded antisense) has been implicated in many aspects of gene regulation, including translational regulation, genomic imprinting, RNA interference, alternative splicing, X chromosome inactivation, RNA editing, and heterochromatic gene silencing (for reviews, see Kumar and Carmichael 1998; Vanhee-Brossollet and Vaquero 1998; Lavorgna et al. 2004). In addition, our recent studies with regard to antisense intron size and sense–antisense coordinate expression have suggested that antisense regulation is likely to be a common and important gene-regulation mechanism in humans (Chen et al. 2005a,b,c).

Several recent genome-wide analyses have suggested that antisense regulation might be common but have different abundance in many eukaryotic genomes. Notably, very different proportions of sense–antisense (SA) transcripts have been proposed in different genomes, for example, ~5% in rice (*Oryza sativa*) (Osato et al. 2003), 15% in the fruit fly (*Drosophila melanogaster*) (Misra et al. 2002), 15% in the mouse (*Mus musculus*) (Kiyosawa et al. 2003), and 22% (Chen et al. 2004) to even >40% (Cheng et al. 2005) in the human (*Homo sapiens*) genome. If this variation is real, it might suggest a relatively rapid loss and gain of antisense regulation in evolution. In this regard, the disparate figures for mouse and human are especially interesting given that these two species are relatively close evolutionarily (diverged ~75 million years ago [Mya]), and their genomes are well annotated. This would fit with the suggestion, albeit not one commonly considered, that there may be more antisense regulation in humans and this may be related to the greater complexity of humans, particularly to the higher demands of human brain learning, memory, speech, and cognitive capabilities (J.S. Mattick, pers. comm.).

Here we ask whether the differences between humans and other animals are likely to be real. In the aforementioned studies, different protocols were used. If then we apply the same protocol to humans and other animals, do we still observe large differences in the estimated proportion of SA transcripts? In this study, we used the same robust protocol that has been demonstrated to be very specific and efficient (Chen et al. 2004), to identify antisense transcripts in the human and other five metazoan genomes. Although the estimated putative SA pairs are much more common in humans than in mice, rats (*Rattus norvegicus*), chicken (*Gallus gallus*), fruit flies, and nematodes (*Caenorhabditis elegans*), the estimated proportions of genes involved in antisense transcription are seriously affected by the number of transcripts available for analysis. By controlling for transcript abundance, we show that the estimated proportion of SA genes is approximately constant in vertebrates.

Results

Differences in frequencies of putative SA gene pairs are not owing to differences in methods to identify SA pairs

We used the same robust protocol (Chen et al. 2004) to identify putative sense–antisense (SA) transcripts in the human, mouse, rat, chicken, fruit fly, and nematode genomes (see Methods). In an experimental validation of SA pairs by orientation-specific RT-PCR, we observed that the specificity of this protocol is >92%; in addition, this protocol also has a higher sensitivity compared with other protocols established previously (Chen et al. 2004). As shown in Figure 1 and Table 1, the putative SA transcripts appear to be much more common in humans (accounting for 22.7% of the whole set of genes) than in mice (11.6%), rats (4.8%), chickens (4.8%), fruit flies (17.2%), and nematodes (0.5%). Thus, the differences that have been reported previously (Misra et al. 2002; Kiyosawa et al. 2003; Chen et al. 2004; Cheng et al. 2005) appear not to be the result of underlying differences in the protocols used. A similar pattern was observed regarding the proportion of qualified transcript sequences (i.e., those that have the correct orientation ensured by our stringent criteria (Chen et al. 2004; see also Methods) estimated to be involved in antisense transcription (Fig. 1), that is, 27.0%, 12.2%, 6.5%, 4.4%, 18.6%, and 0.9% of the qualified transcripts in the human, mouse, rat, chicken, fruit fly, and nematode genome, respectively, may form SA pairs (Table 1). Therefore, the overall proportion of antisense transcripts varies strikingly between the organisms. Most importantly, the pattern appears then to support the possibility that antisense regulation is especially prevalent among human genes. However, there remain further important problems. Notably, the total number of qualified transcript sequences also differs dramatically among the genomes, with threefold to 18-fold more sequences in humans compared with the others (Table 1). Could this explain the apparent enrichment of sense–antisense transcripts in humans?

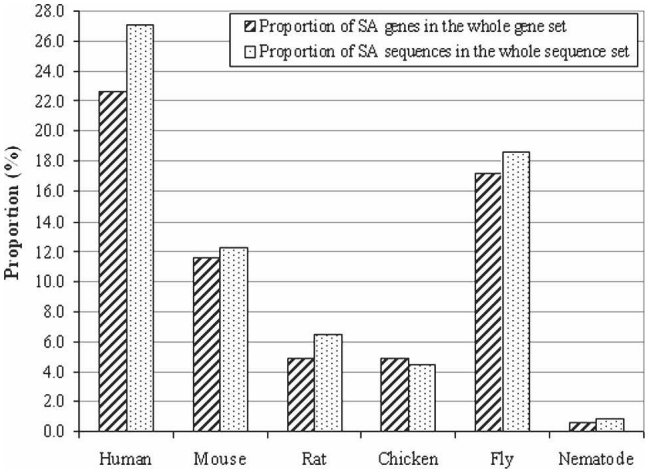


Figure 1. Proportion of sense–antisense (SA) genes or sequences in the whole gene or sequence data set in each genome. One sequence represents one (qualified) transcript sequence, while one gene may contain several (qualified) transcript sequences. For example, as shown in Table 1, a total of 100,444 human SA transcript sequences belong to 6194 SA genes (i.e., ~16 transcript sequences per sense/antisense gene). Humans have the highest proportion of SA genes and sequences, while nematodes have the lowest proportion. See Table 1 for more details.

Table 1. The estimated proportions of SA genes and SA sequences, the distributions of different types of qualified transcript sequences, and the genome compaction among the six individual genomes

Item	Human	Mouse	Rat	Chicken	Fly	Nematode
Total genes	27,333	19,100	11,332	7390	10,542	14,406
SA genes	6194	2212	548	356	1814	76
% of SA genes	22.7	11.6	4.8	4.8	17.2	0.5
Total transcript sequences	371,528	110,076	38,909	22,845	39,612	20,194
SA transcript sequences	100,444	13,474	2512	1011	7373	182
% of SA sequences	27.0	12.2	6.5	4.4	18.6	0.9
Total noncoding genes[a]	9131	2341	4250	4048	81	70
Noncoding SA genes	2162	515	323	252	31	10
% of SA genes	23.7	22.0	7.6	6.2	38.3	14.3
Total coding genes	18,202	16,759	7082	3342	10,461	14,336
Coding SA genes	4032	1697	225	104	1783	66
% of SA genes	22.2	10.1	3.2	3.1	17.0	0.5
Total mRNA sequences	115,729	52,404	17,642	13,309	31,225	17,913
(% in the total transcript sequences)	(31.2)	(47.6)	(45.3)	(58.3)	(78.8)	(88.7)
Total sequences with CDS	92,791	63,181	17,382	6183	31,037	17,900
(% in the total transcript sequences)	(25.0)	(57.4)	(44.7)	(27.1)	(78.4)	(88.6)
Genomic DNA size (bp)[b]	3.1×10^9	2.6×10^9	2.7×10^9	9.4×10^8	1.2×10^8	1.0×10^8
Total genes	27,333	19,100	11,332	7390	10,542	14,406
Genomic compaction (gene number/ Mb genome)	8.8	7.3	4.2	7.9	87.9	144.1

[a]Although some of the genes that do not contain CDS (protein-coding) sequences might be unannotated protein-coding genes, the majority of them may represent ncRNA genes.
[b]"Genomic DNA size" refers to the whole length of the genomic DNA sequences available in the annotated chromosomes in the genome version we downloaded and analyzed.

The higher proportion of SA transcripts in humans is owing to greater availability of transcript sequences

To evaluate the effect of transcript abundance on the estimated proportion of SA genes, we randomly selected a set number of sequences from the whole qualified transcript sequence data set for each genome (see Supplemental Table 1) and then followed the same protocol (see Methods) to estimate the proportion of SA genes. For each number of transcripts, we repeated the analysis independently 1000 times and calculated the mean value of SA proportions. As expected, the estimated proportion of SA genes increases as the transcript number increases (Fig. 2A). Notably, with the same numbers of transcript sequences, the estimated proportion of SA genes in humans is about the same as those in mice, rats, and chickens (and perhaps even a little lower). Unexpectedly, the estimated proportions of SA genes are much higher in flies and much lower in nematodes compared with those in the vertebrates. For example, with 20,000 transcript sequences, the estimated proportion of SA genes is 2.05%, 2.90%, 2.74%, 4.22%, 12.09%, and 0.52% in the human, mouse, rat, chicken, fly, and nematode genome, respectively (Fig. 2A; Supplemental Table 1). A similar pattern has also been observed with respect to the effect of abundance of transcript sequences on the proportion of relevant sequences (rather than genes) estimated to be involved in antisense transcription (Fig. 2B; Supplemental Table 1).

As the protocols and sources used to produce cDNAs/ESTs vary greatly between organisms (Harrington et al. 2004), heterogeneities would exist in transcript data sets between different species. As shown in Table 1, the proportion of mRNA sequences in the total qualified transcript sequences dramatically varies from 31.2% in humans to 88.7% in nematodes; the proportion of protein-coding sequences in the total qualified transcript sequences also dramatically varies from 25.0% in humans to 88.6% in nematodes. Would these heterogeneities in sequence sources bias our observations? To address this issue, we used the same procedures as above to analyze the effect of abundance of mRNA sequences alone or of protein-coding sequences alone on the estimated SA proportions. We observe similar patterns to those reported above (Supplemental Fig. 1a–d), indicating that the different abundance of different sorts of sequence data among the organisms (Table 1) does not bias the observations.

If humans have more SA pairs than mice, then how did the novel SA pairs come about? One possibility is that genes orthologous in mice and humans came to overlap in humans but not in mice. An alternative would be that a gene with an ortholog in mice and humans was subject to de novo generation of new overlapping transcripts but only in humans (hence there would be no mouse ortholog for one of the two genes in the SA pair). A third possibility is that both S and A are de novo in humans. To address these possibilities, we identified all the human–mouse ortholog gene pairs in our data set (see Methods). In some cases, one human gene might have several different ortholog genes in mice, and vice versa. To simplify the analysis, we excluded all the one-to-multiple or multiple-to-multiple ortholog pairs (including all the transcripts that belong to these ortholog genes) from the analysis. After such treatment, we identified 11,931 one-to-one human–mouse ortholog gene pairs in our data set. To avoid the potential bias on the analysis of SA proportion due to the difference in transcript number between the paired human and mouse ortholog gene clusters, we further selected a single ortholog transcript sequence with the longest size from each ortholog gene cluster to represent that ortholog gene. Thus, we obtained 11,931 one-to-one human–mouse ortholog transcript sequence pairs, so that humans and mice have the same number of unique ortholog transcripts. We then analyzed the SA pairs formed between these 11,931 one-to-one human–mouse ortholog transcripts in each species. We found that 314 (i.e., 2.6%) of the 11,931 ortholog transcripts form 157 SA pairs in humans, fewer than that in mice (388, i.e., 3.3% of the 11,931 ortholog transcripts form 194 SA pairs). Therefore, under the first possibility,

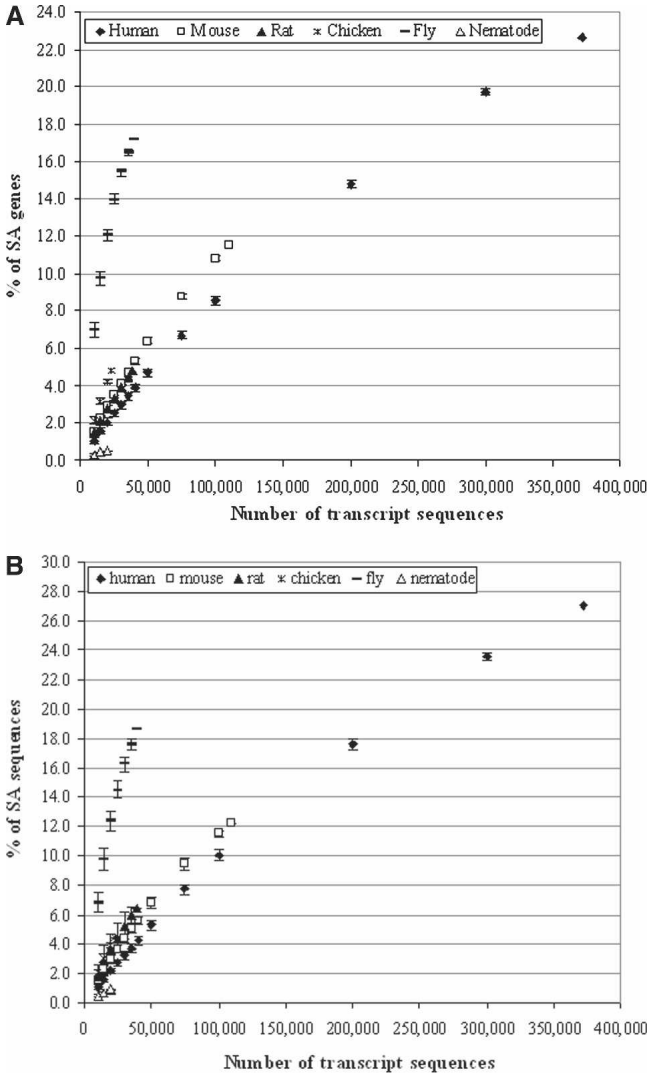


Figure 2. Relationship between the number of qualified transcripts and the estimated proportion of SA genes or sequences. We randomly selected a set number of sequences from the whole qualified transcript sequence data set for each genome (see Supplemental Table 1) and then followed the same procedure (see Methods) to estimate the SA gene/sequence proportion. For each number of transcripts, we repeated the analysis independently 1000 times. The mean values with their standard deviation (mean ± SD) are shown. "% of SA genes/sequences" refers to the proportion of genes/sequences involved in putative antisense transcription (i.e., forming putative SA pairs) in a given gene/sequence data set. (*A*) Relationship between the number of qualified transcripts and the estimated proportion of SA genes. (*B*) Relationship between the number of qualified transcripts and the estimated proportion of SA sequences.

humans do not have more SA pairs formed between the ortholog genes than mice.

To test whether the one-to-one human–mouse ortholog genes form more SA pairs with nonortholog transcripts in humans than in mice, we randomly selected a set number of nonortholog transcripts (those that do not belong to any one-to-one, one-to-multiple, or multiple-to-multiple ortholog pairs) and combined them with the fixed number of 11,931 one-to-one human–mouse ortholog transcripts in each species; we then clustered them and detected the SA pairs formed between ortholog and nonortholog transcripts. As shown in Figure 3A, with the same numbers of nonortholog transcripts, the percentages of the 11,931 one-to-one human–mouse ortholog transcripts that form SA pairs with nonortholog transcripts are almost the same between humans and mice. Thus, under the second possibility, humans do not have more SA pairs formed between the ortholog and nonortholog transcripts than mice.

To test whether the nonortholog transcripts form more SA pairs between themselves in humans than in mice, we randomly selected a set number of nonortholog transcripts as described above, and then clustered them and detected the SA pairs formed between the selected nonortholog transcripts. As shown in Figure 3B, with the same numbers of nonortholog transcripts, the proportions of the SA transcripts within the selected nonortholog transcripts are even slightly higher in mice than in humans. A similar pattern was observed regarding the proportions of SA genes (rather than sequences) (data not shown). Thus, under the third possibility, humans do not have more SA pairs formed between the nonortholog transcripts than mice either.

Similarly, we have identified 3537 one-to-one human–rat ortholog transcript pairs in our data set (see Methods). We found that 14 ortholog transcripts form seven SA pairs in humans, while 10 ortholog transcripts form five SA pairs in rats, and there is no significant difference between them (14/3537 vs. 10/3537; χ^2-test, $P = 0.5396$). Moreover, we have also randomly selected a set number of nonortholog transcripts to detect SA pairs formed within themselves or with the one-to-one ortholog transcripts. As shown in Supplemental Figure 2, a and b, with the same number of nonortholog transcripts, the SA proportions are not higher in humans than in rats. In addition, we identified 905 one-to-one ortholog transcripts between humans and chickens. We found no SA pairs formed between the ortholog transcripts, while the same small number of SA pairs (nine pairs) formed between the ortholog transcripts and nonortholog transcripts in both genomes. As expected, random-subsampling analysis indicates that humans do not have a higher proportion of SA pairs formed within nonortholog transcripts than chickens either (data not shown). These tests, note, additionally control for differences in the data sources in terms of relative completeness of coverage.

Taken together, the higher overall proportion of SA transcripts in humans is owing to greater availability of transcript sequences (Table 1). After controlling for transcript abundance, although the proportion of SA transcripts (in a given size of transcript set) still varies between the organisms, it is not specifically higher in humans compared with other organisms in either the whole transcript data set (Fig. 2A,B), or in a given specific subset of transcript sequences such as mRNAs (Supplemental Fig. 1a,b), protein-coding sequences (Supplemental Fig. 1c,d), or conserved or nonconserved transcripts (Fig. 3A,B; Supplemental Fig. 2a,b).

Human brain appears to be no more enriched for antisense transcription than mouse brain

Human brain is the most complex organ in the human body, and is much more complex than the brains of other mammals. It was hypothesized that an increase in noncoding regulatory capacity reflects the higher demands of human brain learning, memory, speech, and cognitive capabilities (J.S. Mattick, pers. comm.). This raises two questions: (1) Is antisense transcription more prevalent in brain tissue than other tissues? (2) Is antisense transcription more prevalent in human brain compared with the brains of other mammals? To address these two questions, we

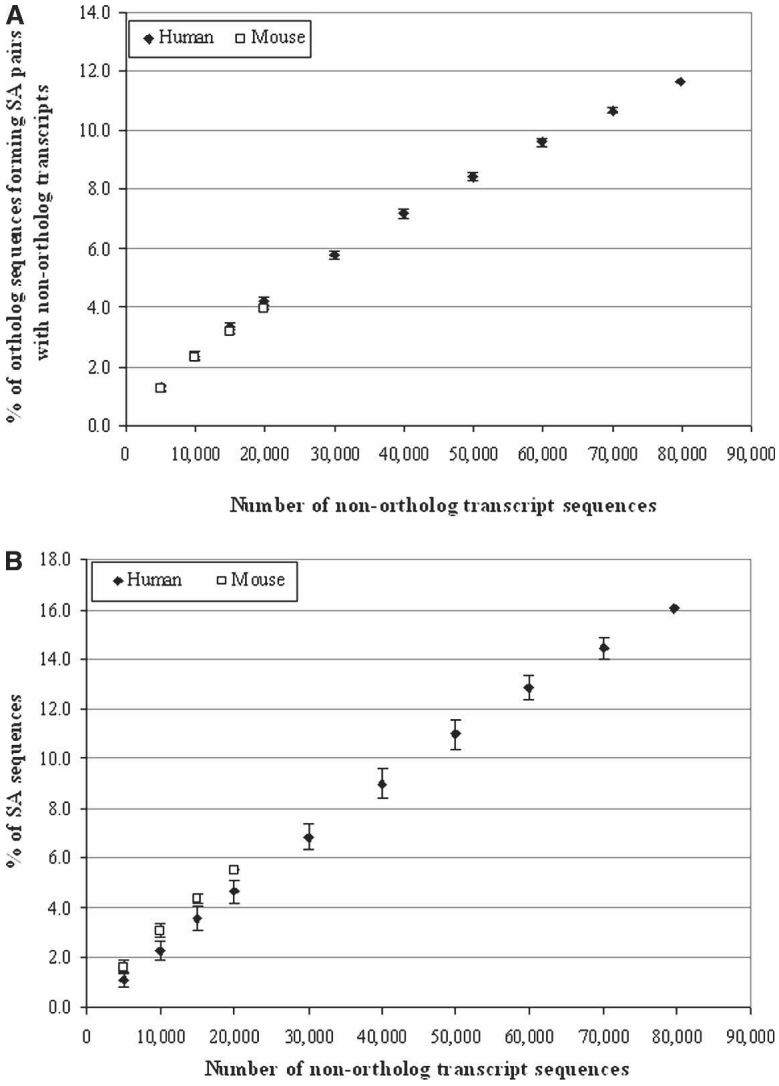


Figure 3. Relationship between the number of nonortholog transcripts and the percentage of human–mouse ortholog transcripts that form SA pairs with nonortholog transcripts, or between the number of nonortholog transcripts and the proportion of SA transcripts within the selected nonortholog transcripts in humans and mice, respectively. (*A*) We randomly selected a set number of sequences from the whole qualified nonortholog transcript sequence data set for each genome, and then combined them with the 11,931 one-to-one human–mouse ortholog transcripts and determined the SA pairs formed between ortholog and nonortholog transcripts. Thus, the percentages of the 11,931 one-to-one human–mouse ortholog transcripts that form SA pairs with nonortholog transcripts were determined in each species. (*B*) We randomly selected a set number of qualified nonortholog transcript sequences and determined the SA pairs formed between nonortholog transcripts. Thus, the percentages of SA transcripts within the selected nonortholog transcripts were determined in each species. For each point, we repeated the analysis independently 1000 times. The mean values with their standard deviation (mean ± SD) are shown.

have performed a comparison of antisense transcription prevalence between human and mouse regarding three different tissue/cell types including brain, liver, and embryonic stem cells. The difference in complexity in liver and embryonic stem cells between human and mouse is thought to be very limited, so that these two types of tissues/cells can be used as a control for the brain. Only the transcripts that have SAGE expression data to support their expression in a given tissue were used in the assembly of transcript clusters (i.e., genes) and further in the analysis of SA proportion in the given tissue (see Methods). We found that the overall proportion of SA genes/sequences is much higher in brain than in liver and embryonic stem cells in both species, and that in each type of tissue/cell, the overall proportion of SA genes/sequences is much higher in humans than in mice. After controlling for transcript abundance, the proportion of SA genes/sequences in a certain number of transcripts is still higher in brain than in liver and embryonic stem cells, and it is very similar between liver and embryonic stem cells in both species (Figs. 4A,B). To try to understand what underlies the finding that antisense expression is enriched in brain, we further detected in these three tissues the number of SA genes that are conserved between human and mouse (and hence likely to be functional). We identified 4951, 2465, and 2948 one-to-one human–mouse ortholog genes that are expressed in both species' brain, liver, and embryonic stem cells, respectively; of them, 3.0% (148/4951), 1.4% (35/2465), and 2.1% (63/2948) form SA pairs in both species in the relevant tissue, respectively. The fraction of conserved SA transcripts in brain (3.0%) is significantly greater than that in liver (1.4%; χ^2-test, $P < 0.0001$) and that in embryonic stem cells (2.1%; χ^2-test, $P < 0.05$). Thus, it seems that the excess of antisense transcripts in brain is not spurious and that the brain might tolerate or need a higher level of antisense expression, although further studies will be required to clarify whether this is also a reflection of differences in transcription pattern between the brain and other tissues. Importantly, however, in each type of tissue/cell, brain included, the proportion is not higher, maybe even lower, in humans compared with that in mice (Figs. 4A,B). Thus, although antisense transcription appears to be more prevalent in mammalian brain tissue compared with nonbrain tissues/cells, the human brain appears to be no more enriched for antisense transcription than the mouse brain (Figs. 4A,B). In fact, the similarity between human and mouse regarding SA proportion in each tissue or cell type (Figs. 4A,B) is the same as that in the whole data set (Figs. 2A,B).

One may also suggest that it might be necessary to control for differences in cDNA cloning methods (e.g., random EST cloning vs. full-length cDNA cloning; using total RNA vs. using poly(A)$^+$ RNA samples) in each tissue. Unfortunately, it is presently not possible to obtain enough transcript sequences that were sequenced from the same tissue and with the same method from each genome for such analysis. However, as shown above, the dramatically different proportion of mRNA sequences among

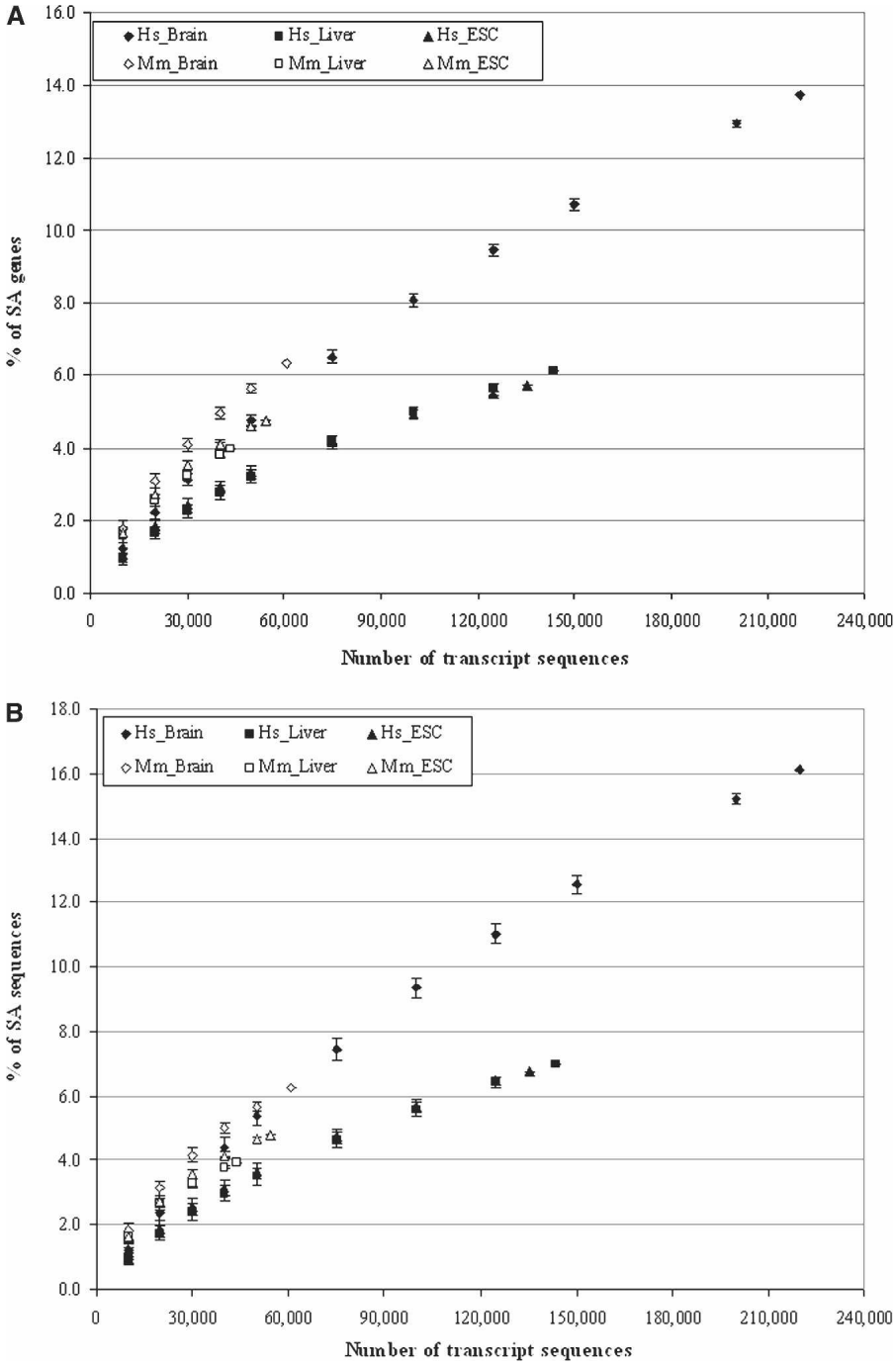


Figure 4. Relationship between the number of qualified transcripts and the estimated proportion of SA genes or sequences in human and mouse brain, liver, and embryonic stem cells. We randomly selected a set number of sequences from the whole set of qualified transcript sequences expressed in each type of tissue/cell in each genome and then estimated the SA proportion. For each number of transcripts, we repeated the analysis independently 1000 times. The mean values with their standard deviation (mean ± SD) are shown. (*A*) Relationship between the number of qualified transcripts and the estimated proportion of SA genes. (*B*) Relationship between the number of qualified transcripts and the estimated proportion of SA sequences. (Hs_) Human (*Homo sapiens*); (Mm_) mouse (*Mus musculus*); (ESC) embryonic stem cells.

the genomes (Table 1), which is largely caused by the difference in cDNA cloning methods (e.g., random EST cloning vs. full-length cDNA cloning) preferentially used in different species, does not significantly bias our findings (Supplemental Fig. 1a,b). Thus, the difference in cDNA cloning methods is unlikely to significantly bias our findings. In addition, according to our procedure (see Methods), mRNA sequences with CDS would be selected as qualified transcript sequences no matter whether they have poly(A) tails/signals or not, and, in fact, >95% of the whole selected protein-coding sequences are mRNAs with CDS; therefore, difference in RNA-sample selection [total RNA vs. poly(A)$^+$ RNA] has already been controlled for in the analysis of protein-coding sequences among the genomes, and our results imply that the difference in RNA-sample selection is unlikely to significantly bias our findings either (Supplemental Fig. 1c,d).

Discussion

Taken together, our results suggest that it is the number of qualified transcript sequences, not the heterogeneities in sequence sources, that has a major effect on the estimated proportions of SA genes and sequences. While it is likely that many more and a greater diversity of transcripts are required to detect the vast majority of antisense transcripts in each species, we find no evidence that humans are in any way unusual among the vertebrates regarding antisense transcription. Why the fly is especially rich in antisense transcription while the nematode is especially poor is far from clear. Genomic compaction seems not to be the primary reason for the enrichment of antisense in fly, because the nematode genome is even more compact than the fly genome (Table 1; see also Levine and Tjian 2003). One might speculate that the low proportion in worm may in part be related to the abundance of operons in the worm genome (Blumenthal et al. 2002), because genes in operons are typically on the same strand, and the individual transcripts are made by splicing the polycistronic RNA, and thus neighboring genes in operons are less likely to be SA regulated *in cis*. Nonetheless, although only 0.34% (8/2326) of the genes incorporated into operons belong to putative SA genes, lower than that (0.53%; 76/14,406) in the whole gene set, the difference is not significant ($P > 0.05$). In addition, because only 15%–16% of the worm genes are incorporated into operons (Blumenthal 1998; Blumenthal et al. 2002; this study), after removing these genes, the overall SA gene proportion only slightly increased from

0.53% to 0.56%. Therefore, further explanation needs to be explored. One might speculate that the dearth in worm is associated with the highly programmed set of cell divisions seen in this taxa.

Might antisense regulation be more prevalent in humans?

As our study identifies putative SA pairs, rather than experimentally confirmed ones, we cannot exclude the possibility that antisense regulation might be more prevalent in humans than in other vertebrate genomes. There are two important parameters. First, it might be the case that in humans, for a given number of predicted putative SA pairs, a higher proportion of these actually function in antisense regulation. If so, even if humans have a similar proportion of putative SA transcripts to other vertebrates, antisense regulation would be more prevalent in humans. Second, humans may have a larger and more diverse transcriptome than other vertebrates. Were this the case, we would predict a higher absolute number of putative SA pairs for the genome with the large transcriptome, even if the proportion of predicted SA pairs that are functional is no different between the taxa. Owing to the combinatorial nature of transcription regulation, possible differences in the absolute number of antisense transcripts could produce a dramatic expansion in regulatory complexity (Levine and Tjian 2003). Is either of these possibilities reasonable?

Might human putative SA transcripts function more frequently in antisense regulation?

We first ask whether human SA genes are more frequently subjected to A-to-I editing. *Cis*-encoded antisense RNAs may function in gene regulation by forming long and perfect double-strand RNAs (dsRNAs) with target sense transcripts in the nucleus (Kumar and Carmichael 1998; Vanhee-Brossollet and Vaquero 1998; Carmichael 2003; Chen et al. 2004, 2005a; Lavorgna et al. 2004; Wang and Carmichael 2004). These, in turn, might be A-to-I edited by adenosine deaminases that act on dsRNA (ADARs) (Bass 2002; Tonkin and Bass 2003). Recent genome-wide computational surveys revealed that 1%~10% of human genes might be subjected to A-to-I editing (Athanasiadis et al. 2004; Blow et al. 2004; Kim et al. 2004a; Levanon et al. 2004; Eisenberg et al. 2005), which is at least an order of magnitude more frequent than in the mouse, rat, chicken, or fly genomes (Kim et al. 2004a; Eisenberg et al. 2005). If the majority of the genes subjected to A-to-I editing belong to SA genes, these findings would suggest that functional antisense transcription is much more prevalent in humans (J.S. Mattick, pers. comm.). However, in analysis of human transcript sequences that have been identified as undergoing RNA editing (Kim et al. 2004a), we see no significant enrichment in putative SA transcripts. Among the 2674 full-length human cDNAs in which Kim et al. (2004a) observed A-to-I editing, 2104 were collected in our human transcript data set; of them, only 30.6% (645) belong to SA transcripts, close to the overall proportion (27.0%) of SA transcripts in the whole human transcript data set. Moreover, the majority of the editing sites are unlikely to be in the putative SA (exonic) overlapping regions, as the sites are predominantly in intronic and intergenic RNAs (Athanasiadis et al. 2004; Blow et al. 2004). In addition, while most of the human A-to-I editing sites reside in primate-specific *Alu* elements (Athanasiadis et al. 2004; Blow et al. 2004; Kim et al. 2004a; Levanon et al. 2004; Eisenberg et al. 2005), Athanasiadis et al. (2004) found that *Alu*-mediated RNA duplexes targeted by RNA editing are formed mainly intramolecularly. Indeed, a recent study (Neeman et al. 2005) demonstrated that the RNA editing level in sense–antisense overlapping areas (apart from the *Alu* regions within them) is negligible in both human and mouse genomes. If so, it is not clear how and why the potential SA dsRNAs avoid the fate of A-to-I editing. One possibility is that sense–antisense pairing might actually occur within the cell, but the resulting RNA duplexes, edited or unedited, are either retained in the nucleus (Carmichael 2003; Wang and Carmichael 2004) or degraded, and are thus not represented in expressed sequence data (Neeman et al. 2005).

Second, antisense transcripts have unusually short introns, which may reflect selection of antisense transcripts for rapid gene regulation (Chen et al. 2005b,c), because transcription is a slow process (Castillo-Davis et al. 2002) and thus long introns would interfere with the need for rapid expression and regulation (Gerhart et al. 1998; Altuvia and Wagner 2000). Here then we ask whether human putative antisense transcripts are unusual in this regard. If this were the case, it would suggest that human antisense transcripts might function more frequently in antisense regulation. Because the numbers of putative SA genes are too small in rats, chickens, and nematodes, we performed this analysis only in humans, mice, and flies. As shown in Figure 5, in all three genomes, antisense genes have significantly shorter introns than do any other category of genes (S, AL, SL and NBD; see Methods for their classification). Notably, the difference of average intron lengths between antisense and other genes is even more evident in mice and flies than in humans (Fig. 5). Moreover, as expected, in both human and mouse genomes, the 347 Human–Mouse (HM)-conserved SA pairs in which at least one member has an ortholog in both species (see Methods) have significantly shorter introns than do nonconserved antisense genes and other genes (Fig. 5). Thus, it might be a common feature in all the genomes that antisense transcripts have unusually short introns.

Third, our previous study (Chen et al. 2005a) indicated that human SA pairs tend to be coexpressed (i.e., simultaneously expressed in the same tissue/cell) and/or inversely expressed (i.e., whereby a high level of one transcript, relative to the titer of the other, in a given tissue at a given time, is matched by a relative low titer of the same transcript in the same tissue at different time), which suggests that antisense regulation might be a common and important mechanism of gene regulation in humans. Then, is such a feature common in other genomes? We investigated coexpression and inverse expression patterns of putative SA pairs in both human and mouse genomes (see Methods). Among the 3097 human SA pairs, 37.8% (1172) and 31.1% (964) are coexpressed and inversely expressed, respectively. Similar proportions of mouse SA pairs are observed to be coexpressed (40.1%; 443 out of 1106 pairs) and to be inversely expressed (33.6%; 372 out of 1106 pairs). In both genomes, the proportions of SA pairs to be coexpressed and/or inversely expressed are significantly more frequent ($P < 10^{-4}$) than expected by chance (see Methods). These results provide no support for the hypothesis that putative SA pairs are more commonly functional in humans.

Taken together, our results show no evidence that SA transcripts in humans are functionally unusual compared with those in other genomes. Thus, potentially functional SA transcripts might be prevalent in almost all the metazoan genomes.

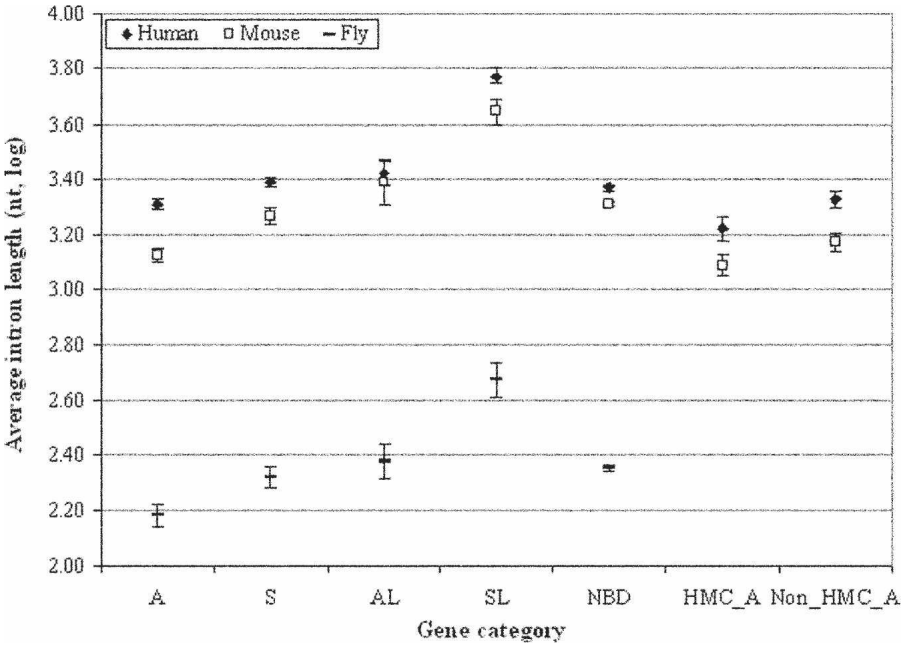


Figure 5. Comparison of the average intron lengths between antisense and other categories of genes in the human, mouse, and fly genomes. The mean values with their 95% confidence intervals of the logarithm (log) values of the intron lengths are shown. We use "independent samples *t*-test" to analyze the logarithm values of the intron lengths to determine significance (*P*-value) in difference of intron lengths between antisense and other genes. In all three genomes, antisense (*A*) genes have significantly shorter introns ($P < 10^{-4}$ in the mouse and fly genomes; $P < 0.01$ in the human genome) than do any other category of genes (S, AL, SL, and NBD; see Methods for their classification). Notably, the difference of average intron lengths between antisense and other genes is even more significant in mice and flies than in humans. Moreover, as expected, in both human and mouse genomes, the average intron length of antisense genes in the 347 HM-conserved SA pairs (HMC_A) is significantly shorter than that of the remaining antisense genes (Non_HMC_A), and than those of the other categories of genes. Similar patterns were observed in the analysis of the total length of intron sequences (data not shown).

Do humans have a richer transcriptome than other mammals?

The subsampling simulations suggest that the diversity of the transcriptomes of mouse and human are comparable. This being so, we need then to ask whether humans have more transcripts, as, if they do, they would then be likely to have a greater number of SA pairs. The fact that humans have more transcript sequences deposited in public databases compared with other mammals is not especially informative, as this must in some part be owing to the fact that the human transcriptome is the best studied one. Indeed, recent studies suggest that the human, mouse, and rat genomes encode similar numbers of genes (Lander et al. 2001; Venter et al. 2001; Waterston et al. 2002; Gibbs et al. 2004; International Human Genome Sequencing Consortium 2004) and have similar numbers of alternative transcripts (Brett et al. 2002; Harrington et al. 2004); thus a priori we should expect similarly sized transcriptomes. However, any such conclusion must be somewhat provisional as it remains unknown just how big each transcriptome is, and recent studies suggest that we have missed many transcripts in each genome. Even in the well-studied human genome, tiling-array studies indicate that there are approximately twice as many nucleotides in poly(A)$^+$ transcripts as are currently annotated in transcript databases (Kapranov et al. 2002; Bertone et al. 2004; Cawley et al. 2004; Kampa et al. 2004; Schadt et al. 2004; Cheng et al. 2005; Johnson et al. 2005; Kapranov et al. 2005). Similarly, a large effort to isolate mouse full-length cDNAs (Okazaki et al. 2002) found that >65% of the cloned full-length cDNAs (39,694 out of 60,770) are novel.

While there is no evidence that humans have a greater number of different transcripts, might they yet have more antisense transcripts? As the majority of antisense transcripts are noncoding RNAs (Cawley et al. 2004; Chen et al. 2004), then, are noncoding RNAs much more abundant in humans than in mice? The best current evidence suggests not. Okazaki et al. (2002) found that >47% (15,815) of a nonredundant set of 33,409 full-length mouse cDNAs are apparently noncoding. This is comparable with 50% (5481) of the 10,897 clusters of human full-length cDNA sequences that appear to be noncoding (Ota et al. 2004). Perhaps, however, are human ncRNAs more enriched in putative SA transcripts? As shown in Table 1, 23.7% of human ncRNA genes appear to be putative SA transcripts, close to that (22.0%) of mouse ncRNA genes.

Taken together, we see no reason to suppose that the human transcriptome is larger than that of mice, with regard to either protein-coding or noncoding transcripts. Therefore, there is no evidence that humans have a much greater number of SA pairs than any other mammal. It is reasonable to suggest that with extensive large-scale cDNA sequencing effort, a similar high abundance of SA transcripts would be detected in nonhuman genomes. Indeed, based on 104,876 FANTOM2 mouse cDNA and public mRNA sequences (Okazaki et al. 2002), Kiyosawa et al. (2003) observed that 15.1% of the assembled mouse transcriptional units (TUs) might form SA pairs; two years later, based on 158,807 FANTOM3 mouse cDNA and public mRNA sequences (Carninci et al. 2005), Katayama et al. (2005) very recently reported that 28.7% of the assembled mouse TUs (with overlapping cDNA evidence) might form SA pairs.

Conclusions

In summary, although antisense regulation could contribute to organismic complexity and we did observe that the abundance of antisense transcripts varies between multicellular animals, there is no prima facie evidence that sense–antisense regulation *in cis* is any more common in humans than in other vertebrates, or is always more common in vertebrates than in invertebrates. Moreover, while brain tissue appears to be enriched for SA transcripts in both mouse and human, the human brain shows no greater enrichment than the mouse brain. We therefore propose that the difference in complexity between mouse and human is unlikely to be owing to different rates of sense–antisense regulation, assuming, as seems parsimonious, that the human transcriptome is no larger than that of mouse. The apparent abundance of antisense transcription in flies and the dearth of it in worm are unexplained.

Methods

Identification of transcript clusters in the six genomes

We used the same protocol described in our previous study (Chen et al. 2004) to identify transcript clusters (i.e., genes) in the human (*Homo sapiens*; an updated version) (Sun et al. 2005), mouse (*Mus musculus*) (Sun et al. 2005), rat (*Rattus norvegicus*), chicken (*Gallus gallus*), fruit fly (*Drosophila melanogaster*), and nematode (*Caenorhabditis elegans*) genomes based on recent versions of databases. In brief, transcript clusters were created based on the mRNA and EST sequences downloaded from UniGene (Schuler et al. 1996) database (human Build #175; mouse Build #141; rat Build #139; chicken Build #26; fly Build #35; nematode Build #20) alignments to the relevant genome (human Build 35.1; mouse Build 33.1; rat Build 3.1; chicken Build 1.1; fly Build 4.0; nematode Wormbase Release WS138). CAP3 (Huang and Madan 1999) and BLAT (Kent 2002) were used for transcript assembling and genome mapping, respectively (Chen et al. 2004; Sun et al. 2005). The transcript sequences and alignments were filtered stringently to ensure the correct orientation: (1) Only the transcripts whose correct orientation could be determined were selected for the study. mRNA sequences had to have at least an annotated protein-coding region (CDS), a poly(A) tail (namely, containing a stretch of at least 10 As at the 3′-end of a sequence), or a poly(A) signal (namely, containing one of the six polyadenylation sites, AATAAA, ATTAAA, AATTAA, AATAAT, CATAAA, and AGTAAA [see Caron et al. 2001], within the last 50 bp of the 3′-end of a sequence). ESTs and any other sequences had to have a poly(A) tail and/or a poly(A) signal if having CDS, or both a poly(A) tail and a poly(A) signal if without CDS. (2) All transcript sequences having suspicious splice sites (e.g., CT-AC, CT-GC, and GT-AT, which are reverse complements of the typical splice donor and acceptor sites GT-AG, GC-AG, and AT-AC, respectively) were discarded. The conditions for genome alignment are: Identity ≥96%, Coverage ≥70%, and Alignment ≥97%. The transcript sequences representing highly abundant and tandem duplicate genes such as immunoglobulins and T-cell receptors were excluded. All transcript sequences aligned to the same genomic locus were assembled into one transcript cluster. After assembly, all clusters that contained only one sequence that did not span an intron were excluded.

Classification of bidirectional transcript cluster pairs

As in our previous study (Chen et al. 2004, 2005a,b,c; Sun et al. 2005), the transcript clusters were classified according to the transcribed pattern in the genomes. Clusters containing at least one pair of transcript sequences transcribed from opposite strands of the same genomic locus were called "bidirectional (BD) clusters," while the remaining clusters containing only one-directional transcripts were called "nonbidirectional (NBD) clusters." We further separated each BD cluster into two new clusters (a cluster pair) based on their overlapping patterns: Sense (S) and antisense (A) clusters form putative sense–antisense (SA) pairs with exon overlaps (identity ≥94%), while the sense-like (SL) and antisense-like (AL) clusters form non-exon-overlapping bidirectional (NOB) pairs without exon overlaps.

In our previous studies (Chen et al. 2004, 2005a,b,c), we defined the S and A or SL and AL genes in each BD gene pair mainly based on a conventional concept (e.g., Lipman 1997) that the S (or SL) gene should exist in more tissues and/or be expressed at a higher level, and thus would have been detected more frequently (i.e., having more transcript sequences deposited in the expressed sequence databases) than its A (or AL) partner. Nevertheless, there is another (even more) common notion that almost all sense genes are protein-coding genes, whereas antisense genes might be coding or noncoding RNA (Kumar and Carmichael 1998; Vanhee-Brossollet and Vaquero 1998; Kiyosawa et al. 2003). The fact that >90% of the defined S (SL) genes in our previous study (Chen et al. 2004) are protein-coding genes (i.e., with annotated CDS regions) is in accord with this notion. However, in a few pairs, the defined S (or SL) lacks CDS, while the corresponding A (or AL) partner has CDS. Thus, recently we revised the previous rules as follows (Sun et al. 2005): (1) For the SA (or NOB) pairs in which one member has CDS while the other lacks CDS, define the one with CDS as the S (or SL) and the other as the A (or AL). (2) For the remaining SA (NOB) pairs, the previous rules (Chen et al. 2004) are applied: (i) define the one containing more transcript sequences as the S or SL cluster, the other as the A or AL cluster; (ii) if the sequence numbers were the same, define the one with more mRNA sequences as the S or SL cluster, the other as the A or AL cluster; (iii) if their mRNA sequence numbers were still the same, define the one with intron-spanning sequence(s) as the S or SL cluster, while the other one without such intron-spanning sequence(s) would be the A or AL cluster. If none of the above conditions are satisfied, define the one mapped to the sense strand of chromosome as the S or SL cluster and the other as the A or AL cluster. After such separation, five categories of unique gene clusters were obtained: S, A, SL, AL, and NBD. A total of 27,333 human, 19,100 mouse, 11,332 rat, 7390 chicken, 10,542 fly, and 14,406 nematode unique genes were identified, each of which represents a single protein- or RNA-coding gene, of which 22.7% (6194) human, 11.6% (2212) mouse, 4.8% (548) rat, 4.8% (356) chicken, 17.2% (1814) fly, and 0.5% (76) nematode unique genes form 3097, 1106, 274, 178, 907, and 38 putative SA pairs, respectively. The full list of the putative SA gene pairs in each genome with information of genomic loci and their representative transcripts is available in Supplemental Tables 2–7.

Analysis of evolutionary conservation of putative SA pairs in the human, mouse, rat, and chicken genomes

As described previously (Sun et al. 2005), we examined ortholog pairs between mouse and human that were reciprocal best "hits" (matches) between the two genomes. We combined the ortholog pairs from the Mouse Genome Informatics Web Site (ftp://ftp.informatics.jax.org/pub/reports/HMD_HumanSequence.rpt; December 2004) and Ensembl MartView (http://www.ensembl.org/Multi/martview; December 2004). By comparing sequence IDs in our mouse and human gene sets with those in the combined ortholog data set, we obtained 11,931 one-to-one human–mouse ortholog pairs in our data sets. Among them, 2681 genes belong to sense–antisense transcripts in the human genome, and so do 1210 genes in the mouse genome. Of these, 347 putative SA pairs in which at least one member has an ortholog in both the human and mouse genomes are conserved in putative SA form in both genomes, and were called HM-conserved putative SA pairs (Sun et al. 2005). Owing to the facts that (1) the number of putative SA pairs in the mouse genome (even in the human genome) is significantly underestimated because of the limitation of qualified transcript sequences, and (2) many antisense transcripts are ncRNAs that are not included in the human–mouse ortholog databases, the number of HM-conserved putative SA pairs might be seriously underestimated. Note that, in the 347 HM-conserved putative SA pairs, it is not necessary that both members be ortholog transcripts. As described in the text, we selected 11,931 one-to-one human–mouse ortholog transcript (unique) sequence pairs, and found that 142 of them form 71 SA pairs in both species.

Similarly, we identified all the human–rat ortholog gene pairs in our data set based on the human–rat ortholog pairs from Ensembl MartView (http://www.ensembl.org/Multi/martview; March 2006). In some cases, one human gene might have several different ortholog genes in rats, and vice versa. To simplify the analysis, we excluded all the one-to-multiple or multiple-to-multiple ortholog pairs (including all the transcripts that belong to these ortholog genes) from the analysis. After such treatment, we identified 3537 one-to-one human–rat ortholog gene pairs in our data set. To avoid the potential bias on the analysis of SA proportion due to the difference in transcript number between the paired human and rat ortholog gene clusters, we further selected a single ortholog transcript sequence with the longest size from each ortholog gene cluster to represent that ortholog gene. Thus, we obtained 3537 one-to-one human–rat ortholog transcript sequence pairs, so that humans and rats have the same number of unique ortholog transcripts. We found that 14 ortholog transcripts form seven SA pairs in human, while 10 ortholog transcripts form five SA pairs in rats.

In addition, we identified 905 one-to-one ortholog transcripts between humans and chickens from Ensembl MartView (http://www.ensembl.org/Multi/martview; March 2005). We found no SA pairs formed between the ortholog transcripts, while the same small number of SA pairs (nine pairs) formed between the ortholog transcripts and nonortholog transcripts in both genomes.

Investigation of coexpression and inverse expression patterns of putative SA pairs in the human and mouse genomes

As described in our previous study (Chen et al. 2005a), we evaluated the coexpression and inverse expression of SA pairs at the whole genome level based on their expression profiles obtained from SAGE (serial analysis of gene expression) data (Velculescu et al. 1995). In our recent study (Sun et al. 2005), we have made some modifications on our established procedures (Chen et al. 2005a). We downloaded SAGE expression data (NlaIII SAGE libraries) from the NCBI GEO platform (http://www.ncbi.nlm.nih.gov/projects/geo; December 2004), including 245 human SAGE libraries and 76 mouse SAGE libraries. For both human and mouse, we constructed 16 tissue/cell-type SAGE library (including brain, liver, and embryonic stem cells that are available for both human and mouse) combinations to determine coexpression of gene pairs, and constructed 50 comparison cases, each of which is a pair of two states (two different unique SAGE libraries) at different developmental, differentiation, physiologic, or pathological stages/conditions of the same tissue, to determine inverse expression of gene pairs (Sun et al. 2005). Tag counts were converted to counts per million (cpm), and the expression data were cross-linked to our genes by extracting the 3′-most NlaIII SAGE tag for each transcript in the genes (i.e., transcript clusters). Only tags that matched to a single gene were taken into account. All SAGE tags mapped to the same gene were then combined, and the sum of their counts per million in a tissue/cell represented the expression level of that gene in that tissue/cell. To eliminate the potential sequence errors in low-count SAGE tags (Chen et al. 2005a), we kept only the genes that have an expression level of at least 3 cpm across all the 16 tissues (i.e., the sum of expression levels in the 16 tissues).

To evaluate the coexpression of an SA pair, we adopted an index of coexpression between two genes *a* and *b* ($ICE_{a,b}$) defined by Lercher et al. (2002) that is the number of tissues with common positive expression, weighted by the geometric mean of the two breadths. Note that, unlike the conventional "Pearson correlation coefficient (*r*)," coexpression in this context refers not to the extent to which levels of transcripts are correlated, but rather to the coupled presence or absence of the transcripts across different tissues or cells (Chen et al. 2005a). $ICE_{a,b}$ ranges from 0 (no coexpression) to 1 (perfect coexpression). We found that it is higher than the 99% confidence intervals (i.e., $P < 0.01$) of the average $ICE_{a,b}$ values of all the possible gene pairs when $ICE_{a,b} \geq 0.6$ in humans or ≥ 0.5 in mice. Thus, we define two genes (*a* and *b*; e.g., the sense and antisense in a putative SA pair) to be coexpressed if the $ICE_{a,b} \geq 0.6$ in humans or ≥ 0.5 in mice (Sun et al. 2005).

To measure inverse-expression pattern in a more quantitative way compared with that described previously (Chen et al. 2005a), we defined (Sun et al. 2005) a new index of inverse expression between two genes *a* and *b* ($IIE_{a,b}$) that is the number of comparison cases in which the two partners exhibit an inverse expression pattern between two states (i.e., a member is expressed at a higher level at state 1 but a lower level at state 2 compared with its partner; and vice versa) and a significantly greater change of the relative expression ratio of gene *a* to gene *b* between two states than expected by chance (i.e., exceeding the 99% confidence interval of the mean changes of all the randomly formed gene pairs), weighted by the geometric mean of the two presence breadths. A given gene with positive expression in at least one of the two states of a comparison case would be recognized as being presented in that case. The presence breadth for each gene is the number of cases in which the gene is presented. $IIE_{a,b}$ ranges from 0 (no inverse expression) to 1 (perfect inverse expression). Similarly, we define two genes (*a* and *b*; e.g., the sense and antisense in a putative SA pair) to be inversely expressed if the $IIE_{a,b}$ is higher than the 99% confidence intervals (i.e., $P < 0.01$) of the average $IIE_{a,b}$ values of all the randomly formed gene pairs (Sun et al. 2005).

To examine whether coexpression and inverse expression of human and mouse SA pairs are more frequent than expected by chance, we generated control data sets (i.e., "non-expression-level-dependent randomly-replaced" [NEDRR] pseudo SA pair sets) (see Chen et al. 2005a) by replacing each gene in the natural SA set with a randomly picked gene from non-SA genes regardless of its expression level. We compared the coexpression or inverse expression rate of the natural SA set with those from 100,000 pseudo SA sets.

The detailed list of the 3097 human and 1106 mouse putative SA gene pairs with information of evolutionary conservation, coexpression, and inverse expression is available in Supplemental Tables 2 and 3, respectively.

Analysis of the proportion of SA genes/sequences in human and mouse brain, liver, and embryonic stem cells

The 3′-most NlaIII SAGE tag was extracted from each qualified transcript in human (371,528 transcripts in total) (Table 1) and in mouse (110,076 transcripts in total) (Table 1), respectively. After removing those tags that matched to more than one gene, the remaining extracted tags were aligned to the real experimental SAGE tags collected in each of the three tissue/cell-type SAGE library combinations (brain, liver, and embryonic stem cells) in human and mouse, respectively. There are a total of 219,752, 143,616, 135,203, 60,684, 43,570, and 54,183 qualified transcripts that have SAGE expression data to support their expression in human brain, human liver, human embryonic stem cell, mouse brain, mouse liver, and mouse embryonic stem cell, respectively. The transcripts that have SAGE tags detected in a given tissue/cell-type were used to assemble transcript clusters (i.e., genes) as described above for the given tissue or cell type. A total of 21,142, 10,253, 9533, 11,632, 7712, and 10,284 transcript

clusters (i.e., genes) were assembled in human brain, human liver, human embryonic stem cell, mouse brain, mouse liver, and mouse embryonic stem cell, respectively, of which 2902, 628, 544, 738, 308, and 488 form SA pairs in the given tissue/cell type, respectively. In addition, we detected 4951, 2465, and 2948 one-to-one human–mouse ortholog genes that are expressed in both species' brain, liver, and embryonic stem cells, respectively; of them, 3.0% (148/4951), 1.4% (35/2465), and 2.1% (63/2948) form SA pairs in both species in the relevant tissue, respectively.

Analysis of the average intron lengths of five gene categories in the human, mouse, and fly genomes

As described above (Chen et al. 2005b,c), to avoid non-intron-spanning EST transcripts that might skew the result of intron-length analysis, we only included intron-spanning genes for the study. 1757 A, 2929 S, 864 AL, 1630 SL, and 14,851 NBD genes were collected in humans; 642 A, 1046 S, 304 AL, 632 SL, and 14,325 NBD genes were collected in mice; 743 A, 865 S, 429 AL, 503 SL, and 6919 NBD genes were collected in flies.

Acknowledgments

We thank Janet D. Rowley for her support on this study. We also thank John S. Mattick for constructive discussion on the relationship between the prevalence of antisense transcription and organismic complexity, W. James Kent and Xiaoqiu Huang for their help in genome BLAT analysis and CAP3 assembly, and three anonymous referees for constructive comments. This work was supported by the G. Harold and Leila Y. Mathers Charitable Foundation (J.C.), Cancer Research Foundation Young Investigator Award (J.C.), NIH grant CA84405 (Janet D. Rowley), and the Spastic Paralysis Foundation of the Illinois, Eastern Iowa Branch of Kiwanis International (Janet D. Rowley). L.D.H. was supported by the UK Biotechnology and Biological Sciences Research Council. G.G.C. was supported by NIH grant GM066816.

References

Altuvia, S. and Wagner, E.G. 2000. Switching on and off with RNA. *Proc. Natl. Acad. Sci.* **97:** 9824–9826.

Ambros, V. 2004. The functions of animal microRNAs. *Nature* **431:** 350–355.

Athanasiadis, A., Rich, A., and Maas, S. 2004. Widespread A-to-I RNA editing of *Alu*-containing mRNAs in the human transcriptome. *PLoS Biol.* **2:** e391.

Bartel, D.P. 2004. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **116:** 281–297.

Bass, B.L. 2002. RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.* **71:** 817–846.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., et al. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306:** 2242–2246.

Blow, M., Futreal, P.A., Wooster, R., and Stratton, M.R. 2004. A survey of RNA editing in human brain. *Genome Res.* **14:** 2379–2387.

Blumenthal, T. 1998. Gene clusters and polycistronic transcription in eukaryotes. *Bioessays* **20:** 480–487.

Blumenthal, T., Evans, D., Link, C.D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W.L., Duke, K., Kiraly, M., et al. 2002. A global analysis of *Caenorhabditis elegans* operons. *Nature* **417:** 851–854.

Brett, D., Pospisil, H., Valcarcel, J., Reich, J., and Bork, P. 2002. Alternative splicing and genome complexity. *Nat. Genet.* **30:** 29–30.

Carmichael, G.G. 2003. Antisense starts making more sense. *Nat. Biotechnol.* **21:** 371–372.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309:** 1559–1563.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A., et al. 2001. The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. *Science* **291:** 1289–1292.

Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E.V., and Kondrashov, F.A. 2002. Selection for short introns in highly expressed genes. *Nat. Genet.* **31:** 415–418.

Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P., Sekinger, E.A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A.J., et al. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116:** 499–509.

Chen, J., Sun, M., Kent, W.J., Huang, X., Xie, H., Wang, W., Zhou, G., Shi, R.Z., and Rowley, J.D. 2004. Over 20% of human transcripts might form sense–antisense pairs. *Nucleic Acids Res.* **32:** 4812–4820.

Chen, J., Sun, M., Hurst, L.D., Carmichael, G.G., and Rowley, J.D. 2005a. Genome-wide analysis of coordinate expression and evolution of human *cis*-encoded sense–antisense transcripts. *Trends Genet.* **21:** 326–329.

———. 2005b. Human antisense genes have unusually short introns: Evidence for selection for rapid transcription. *Trends Genet.* **21:** 203–207.

Chen, J., Sun, M., Rowley, J.D., and Hurst, L.D. 2005c. The small introns of antisense genes are better explained by selection for rapid transcription than by "genomic design." *Genetics* **171:** 2151–2155.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308:** 1149–1154.

Eisenberg, E., Nemzer, S., Kinar, Y., Sorek, R., Rechavi, G., and Levanon, E.Y. 2005. Is abundant A-to-I RNA editing primate-specific? *Trends Genet.* **21:** 77–81.

Gerhart, E., Wagner, H., and Brantl, S. 1998. Kissing and RNA stability in antisense control of plasmid replication. *Trends Biochem. Sci.* **23:** 451–454.

Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493–521.

Harrington, E.D., Boue, S., Valcarcel, J., Reich, J.G., and Bork, P. 2004. In Reply to "Estimating rates of alternative splicing in mammals and invertebrates." *Nat. Genet.* **36:** 916–917.

Huang, X. and Madan, A. 1999. CAP3: A DNA sequence assembly program. *Genome Res.* **9:** 868–877.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.

Johnson, J.M., Edwards, S., Shoemaker, D., and Schadt, E.E. 2005. Dark matter in the genome: Evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.* **21:** 93–102.

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., et al. 2004. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14:** 331–342.

Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P., and Gingeras, T.R. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296:** 916–919.

Kapranov, P., Drenkow, J., Cheng, J., Long, J., Helt, G., Dike, S., and Gingeras, T.R. 2005. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.* **15:** 987–997.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J., et al. 2005. Antisense transcription in the mammalian transcriptome. *Science* **309:** 1564–1566.

Kent, W.J. 2002. BLAT—The BLAST-like alignment tool. *Genome Res.* **12:** 656–664.

Kim, D.D., Kim, T.T., Walsh, T., Kobayashi, Y., Matise, T.C., Buyske, S., and Gabriel, A. 2004a. Widespread RNA editing of embedded *Alu* elements in the human transcriptome. *Genome Res.* **14:** 1719–1725.

Kim, H., Klein, R., Majewski, J., and Ott, J. 2004b. Estimating rates of alternative splicing in mammals and invertebrates. *Nat. Genet.* **36:** 915–916; author reply 916–917.

Kiyosawa, H., Yamanaka, I., Osato, N., Kondo, S., and Hayashizaki, Y. 2003. Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res.* **13:** 1324–1334.

Kramer, C., Loros, J.J., Dunlap, J.C., and Crosthwaite, S.K. 2003. Role for antisense RNA in regulating circadian clock function in *Neurospora crassa*. *Nature* **421:** 948–952.

Kumar, M. and Carmichael, G.G. 1998. Antisense RNA: Function and fate of duplex RNA in cells of higher eukaryotes. *Microbiol. Mol. Biol.*

Rev. **62:** 1415–1434.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Lavorgna, G., Dahary, D., Lehner, B., Sorek, R., Sanderson, C.M., and Casari, G. 2004. In search of antisense. *Trends Biochem. Sci.* **29:** 88–94.
Lercher, M.J., Urrutia, A.O., and Hurst, L.D. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.* **31:** 180–183.
Levanon, E.Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z.Y., Shoshan, A., Pollock, S.R., Sztybel, D., et al. 2004. Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat. Biotechnol.* **22:** 1001–1005.
Levine, M. and Tjian, R. 2003. Transcription regulation and animal diversity. *Nature* **424:** 147–151.
Lewis, B.P., Burge, C.B., and Bartel, D.P. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15–20.
Lipman, D.J. 1997. Making (anti)sense of non-coding sequence conservation. *Nucleic Acids Res.* **25:** 3580–3583.
Mattick, J.S. 2001. Non-coding RNAs: The architects of eukaryotic complexity. *EMBO Rep.* **2:** 986–991.
———. 2004. RNA regulation: A new genetics? *Nat. Rev. Genet.* **5:** 316–323.
———. 2005. What makes a human? *Scientist* **19:** 32–33.
McShea, D.W. 1996. Metazoan complexity and evolution: Is there a trend? *Evolution Int. J. Org. Evolution* **50:** 477–492.
Merino, E., Balbas, P., Puente, J.L., and Bolivar, F. 1994. Antisense overlapping open reading frames in genes from bacteria to humans. *Nucleic Acids Res.* **22:** 1903–1908.
Misra, S., Crosby, M.A., Mungall, C.J., Matthews, B.B., Campbell, K.S., Hradecky, P., Huang, Y., Kaminker, J.S., Millburn, G.H., Prochnik, S.E., et al. 2002. Annotation of the *Drosophila melanogaster* euchromatic genome: A systematic review. *Genome Biol.* **3:** research0083.
Modrek, B. and Lee, C. 2002. A genomic view of alternative splicing. *Nat. Genet.* **30:** 13–19.
Neeman, Y., Dahary, D., Levanon, E.Y., Sorek, R., and Eisenberg, E. 2005. Is there any sense in antisense editing? *Trends Genet.* **21:** 544–547.
Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563–573.
Osato, N., Yamada, H., Satoh, K., Ooka, H., Yamamoto, M., Suzuki, K., Kawai, J., Carninci, P., Ohtomo, Y., Murakami, K., et al. 2003. Antisense transcripts with rice full-length cDNAs. *Genome Biol.* **5:** R5.
Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., Nagai, K., et al. 2004. Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet.* **36:** 40–45.
Schadt, E.E., Edwards, S.W., GuhaThakurta, D., Holder, D., Ying, L., Svetnik, V., Leonardson, A., Hart, K.W., Russell, A., Li, G., et al. 2004. A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biol.* **5:** R73.
Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., et al. 1996. A gene map of the human genome. *Science* **274:** 540–546.
Sun, M., Hurst, L.D., Carmichael, G.G., and Chen, J. 2005. Evidence for a preferential targeting of 3′-UTRs by *cis*-encoded natural antisense transcripts. *Nucleic Acids Res.* **33:** 5533–5543.
Tonkin, L.A. and Bass, B.L. 2003. Mutations in RNAi rescue aberrant chemotaxis of ADAR mutants. *Science* **302:** 1725.
Vanhee-Brossollet, C. and Vaquero, C. 1998. Do natural antisense transcripts make sense in eukaryotes? *Gene* **211:** 1–9.
Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene expression. *Science* **270:** 484–487.
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.
Wang, Q. and Carmichael, G.G. 2004. Effects of length and location on the cellular response to double-stranded RNA. *Microbiol. Mol. Biol. Rev.* **68:** 432–452.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Received June 18, 2005; accepted in revised form April 24, 2006.