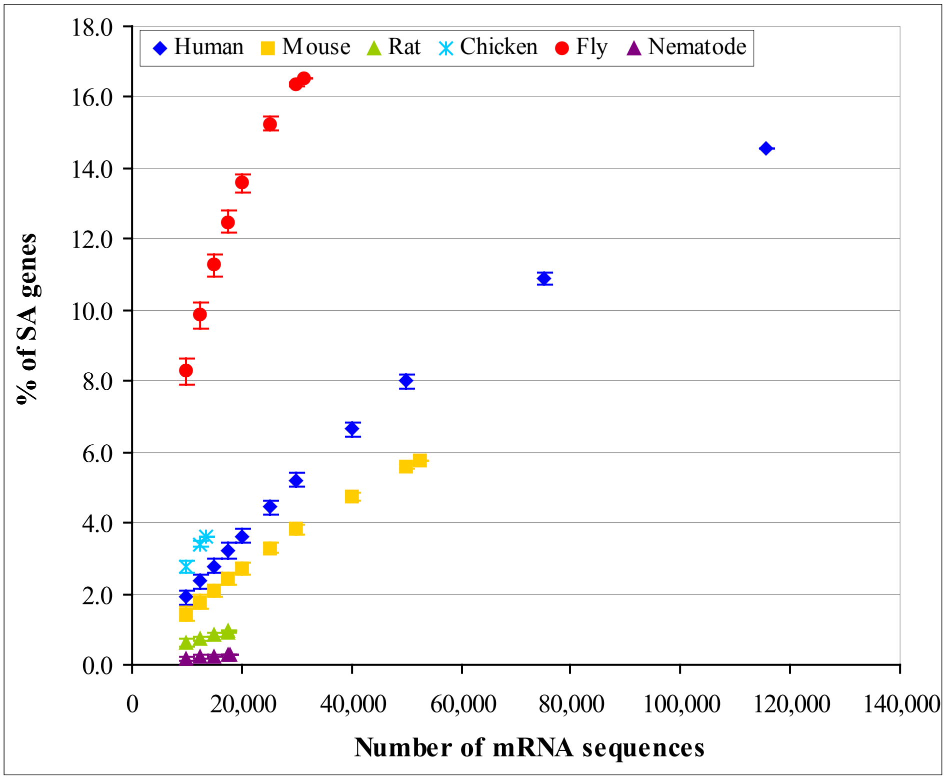
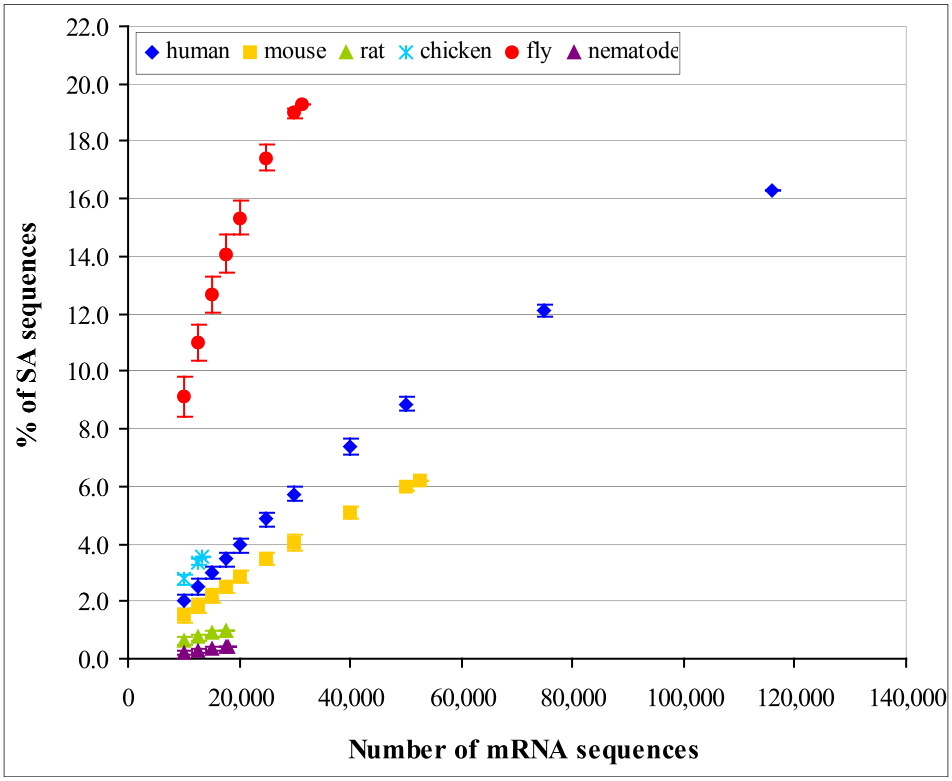
Supplementary Figure 1. Relationship between the number of qualified mRNA or protein-coding transcripts and the estimated proportions of SA genes or sequences. We randomly selected a set number of sequences from the whole qualified mRNA or protein-coding transcript sequence dataset for each genome and then estimated the SA proportions. For each number of transcripts, we repeated the analysis independently 1000 times and calculated the mean value of SA proportions. The mean values with their standard deviation (mean ± SD) are shown. (***a***) The proportion of genes involved in putative antisense transcription in a given mRNA transcript dataset. (***b***) The proportion of transcript sequences involved in putative antisense transcription in a given mRNA transcript dataset. (***c***) The proportion of genes involved in putative antisense transcription in a given protein-coding (i.e., with CDS) transcript dataset. (***d***) The proportion of transcript sequences involved in putative antisense transcription in a given protein-coding transcript dataset.

Supplementary Figure 2. Relationship between the number of non-ortholog transcripts and the percentage of human-rat ortholog transcripts that form SA pairs with non-ortholog transcripts, or between the number of non-ortholog transcripts and the proportion of SA transcripts within the selected non-ortholog transcripts in humans and rats respectively. (***a***) We randomly selected a set number of sequences from the whole qualified non-ortholog transcript sequence dataset for each genome, and then combined with the 3537 one-to-one human-rat ortholog transcripts and determined the SA pairs formed between ortholog and non-ortholog transcripts. Thus, the percentages of the 3537 one-to-one human-rat ortholog transcripts that form SA pairs with non-ortholog transcripts were determined in each species. (***b***) We randomly selected a set number of qualified non-ortholog transcript sequences and determined the SA pairs formed between non-ortholog transcripts. Thus, the percentages of SA transcripts within the selected non-ortholog 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. A similar pattern was observed regarding the proportions of SA genes (rather than sequences; data not shown).

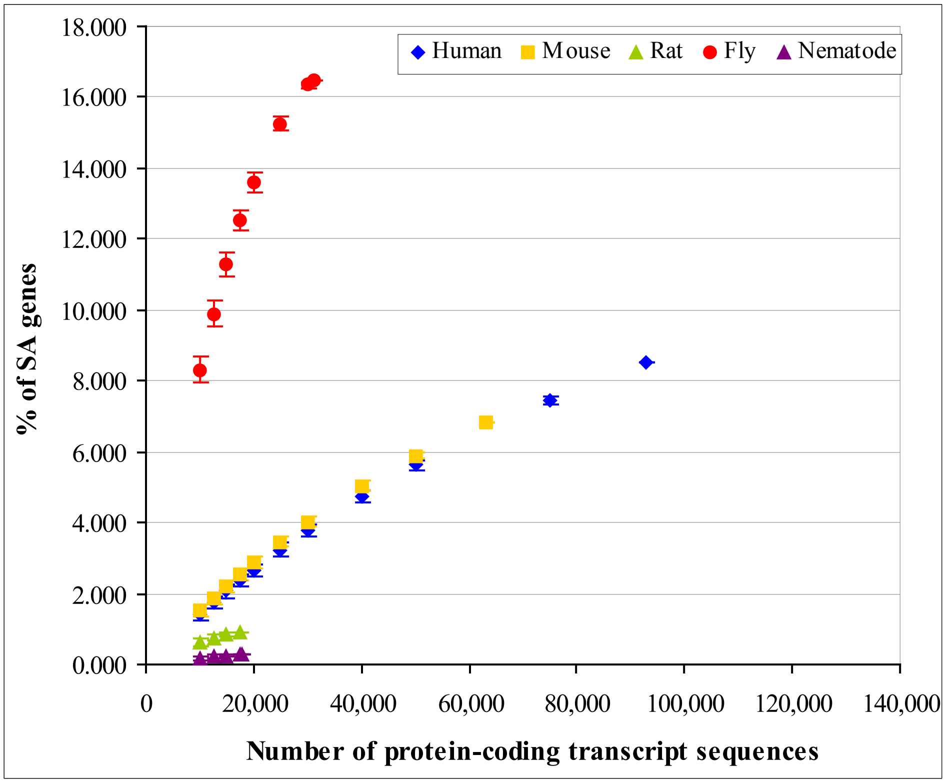
Supplementary Figure 1a.



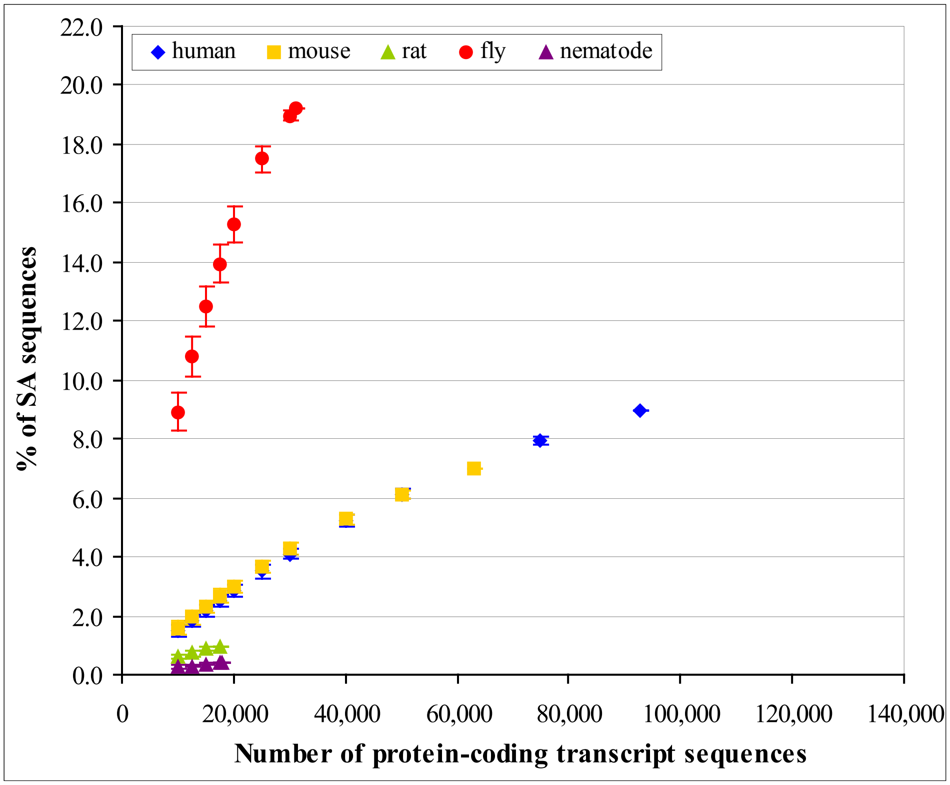
Supplementary Figure 1b.



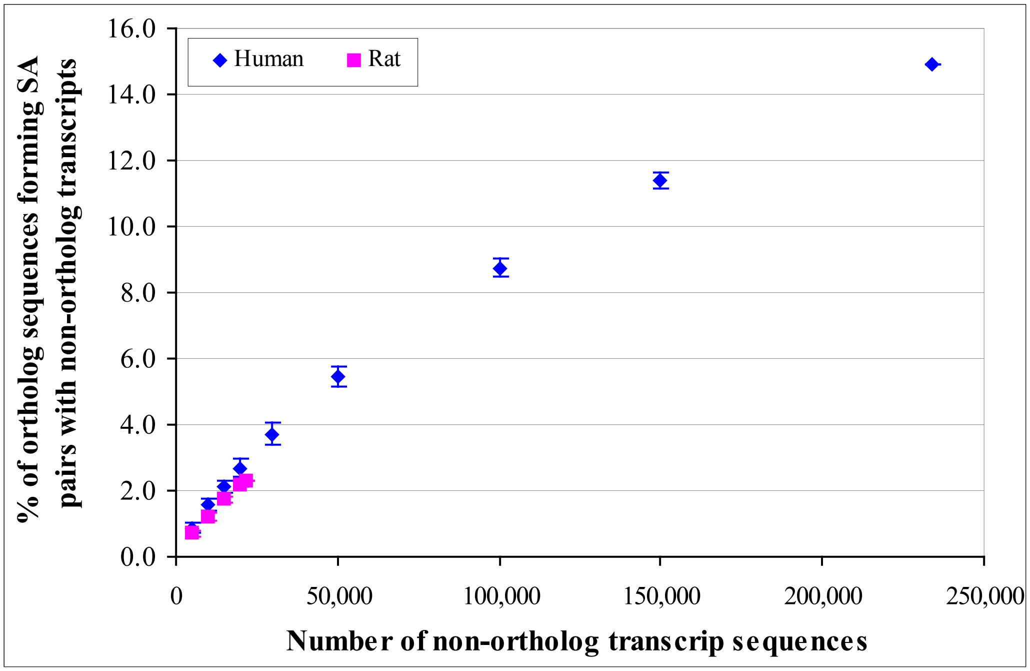
Supplementary Figure 1c.



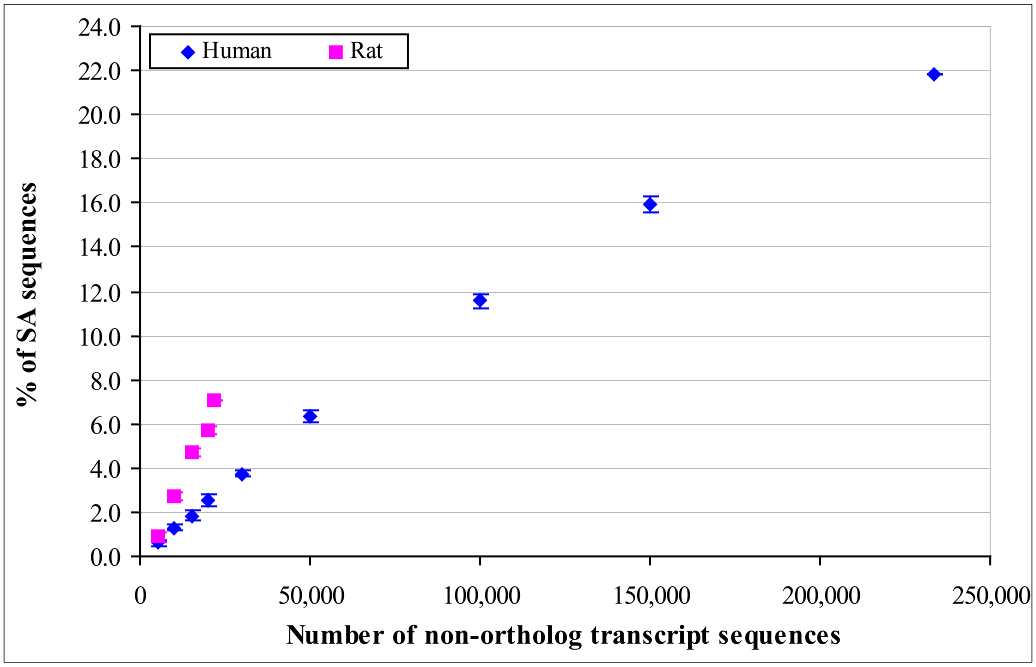
Supplementary Figure 1d.



Supplementary Figure 2a.



Supplementary Figure 2b.



Supplementary Table 1. Estimated proportions of SA genes and sequences based on different numbers of qualified transcript sequences in the six individual genomes[a].

Sequence number		Human	Mouse	Rat	Chicken	Fly	Nematode
Estimated SA gene proportion	10,000	1.06 ± 0.17	1.51 ± 0.20	1.47 ± 0.19	2.14 ± 0.22	6.98 ± 0.40	0.32 ± 0.07
	15,000	1.57 ± 0.18	2.22 ± 0.19	2.12 ± 0.19	3.17 ± 0.19	9.76 ± 0.36	0.43 ± 0.05
	20,000	2.05 ± 0.18	2.90 ± 0.20	2.74 ± 0.17	4.22 ± 0.13	12.09 ± 0.31	0.52 ± 0.01
	20,194						**0.53**
	22,845				**4.82**		
	25,000	2.52 ± 0.18	3.54 ± 0.19	3.33 ± 0.15		13.97 ± 0.26	
	30,000	2.98 ± 0.18	4.15 ± 0.20	3.91 ± 0.12		15.44 ± 0.20	
	35,000	3.43 ± 0.19	4.74 ± 0.19	4.44 ± 0.09		16.51 ± 0.13	
	38,909			**4.84**			
	39,612					**17.21**	
	40,000	3.87 ± 0.19	5.30 ± 0.19				
	50,000	4.71 ± 0.19	6.39 ± 0.18				
	75,000	6.70 ± 0.20	8.79 ± 0.15				
	100,000	8.54 ± 0.20	10.83 ± 0.09				
	110,076		**11.58**				
	200,000	14.81 ± 0.18					
	300,000	19.71 ± 0.13					
	371,528	**22.66**					
Estimated SA sequence proportion	10000	1.09 ± 0.18	1.52 ± 0.20	1.76 ± 0.81	2.06 ± 0.25	6.82 ± 0.69	0.47 ± 0.11
	15000	1.65 ± 0.20	2.25 ± 0.25	2.85 ± 1.01	3.01 ± 0.26	9.75 ± 0.72	0.69 ± 0.08
	20000	2.17 ± 0.22	2.96 ± 0.22	3.64 ± 1.03	3.92 ± 0.16	12.38 ± 0.65	0.89 ± 0.02
	20194						**0.90**
	22845				**4.43**		
	25000	2.70 ± 0.24	3.64 ± 0.23	4.38 ± 1.01		14.52 ± 0.61	
	30000	3.22 ± 0.26	4.34 ± 0.29	5.25 ± 0.86		16.23 ± 0.48	
	35000	3.72 ± 0.28	4.99 ± 0.31	5.98 ± 0.58		17.59 ± 0.36	
	38909			**6.46**			
	39612					**18.61**	
	40000	4.23 ± 0.31	5.60 ± 0.31				
	50000	5.26 ± 0.33	6.80 ± 0.37				
	75000	7.74 ± 0.33	9.42 ± 0.40				
	100000	10.02 ± 0.35	11.53 ± 0.23				
	110076		**12.24**				
	200000	17.63 ± 0.39					
	300000	23.56 ± 0.28					
	371528	**27.04**					

[a]We randomly selected a set number of sequences from the whole qualified transcript sequence dataset for each genome and then followed the same procedure (see Methods) to estimate the proportion of SA genes and of SA sequences. For each number of transcripts, we repeated the analysis independently 1000 times and calculated the mean value of SA proportions. The mean values with their standard deviation (mean ± SD) are shown. The data in bold refer to the whole sequence set and the relevant SA proportions for the corresponding genome.

SUPPLEMENTARY MATERIAL

Identification of transcript clusters in the six genomes

We employed 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 insure 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 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 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 complement 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 'non-bidirectional (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 over 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 above conditions were 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 online Supplementary 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 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 dataset, we obtained 11,931 one-to-one human-mouse ortholog pairs in our datasets. 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). Due to the facts that a) the number of putative SA pairs in the mouse genome (even in the human genome) are significantly underestimated because of the limitation of qualified transcript sequences, and b) 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 are 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 dataset 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 dataset. 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 7 SA pairs in human while 10 ortholog transcripts form 5 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 (9 pairs) formed between the ortholog transcripts and non-ortholog transcripts in both genomes.

Investigation of co-expression 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 co-expression 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 (*Nla*III SAGE libraries) from the NCBI GEO platform (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) combination to determine co-expression 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 *Nla*III 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 cpm 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 co-expression of a SA pair, we adopted an index of co-expression between two genes *a* and *b* ($\mathrm{ICE}_{a,b}$) defined by Lercher *et al.* (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)*', co-expression 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). $\mathrm{ICE}_{a,b}$ ranges from 0 (no co-expression) to 1 (perfect co-expression). We found that it is higher than the 99% confidence intervals (i.e., $P < 0.01$) of the average $\mathrm{ICE}_{a,b}$ values of all the possible gene pairs when $\mathrm{ICE}_{a,b} \geq 0.6$ 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 co-expressed if the $\mathrm{ICE}_{a,b} \geq 0.6$ in humans or $\geq$ 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* ($\mathrm{IIE}_{a,b}$) 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. $\mathrm{IIE}_{a,b}$ 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 $\mathrm{IIE}_{a,b}$ is higher than the 99% confidence intervals (i.e., $P < 0.01$) of the average $\mathrm{IIE}_{a,b}$ values of all the randomly formed gene pairs (Sun et al. 2005).

To examine whether co-expression 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 co-expression 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, co-expression, and inverse expression is available in online Supplementary Table 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 *Nla*III 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 combination (brain, liver and embryonic stem cells) in human and mouse respectively. There are a total 219,752, 143,616, 135,203, 60,684, 43,570 and 54,183 qualified transcripts 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 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 before (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.