

Widespread occurrence of antisense transcription in the human genome

Rodrigo Yelin^{1*}, Dvir Dahary^{1*}, Rotem Sorek^{1*}, Erez Y. Levanon^{1*}, Orly Goldstein^{1*}, Avi Shoshan^{2*}, Alex Diber¹, Sharon Biton¹, Yael Tamir¹, Rami Khosravi¹, Sergey Nemzer¹, Elhanan Pinner¹, Shira Walach¹, Jeanne Bernstein¹, Kinneret Savitsky¹, and Galit Rotman^{1†}

Published online 17 March 2003; doi:10.1038/nbt808

An increasing number of eukaryotic genes are being found to have naturally occurring antisense transcripts. Here we study the extent of antisense transcription in the human genome by analyzing the public databases of expressed sequences using a set of computational tools designed to identify sense-antisense transcriptional units on opposite DNA strands of the same genomic locus. The resulting data set of 2,667 sense-antisense pairs was evaluated by microarrays containing strand-specific oligonucleotide probes derived from the region of overlap. Verification of specific cases by northern blot analysis with strand-specific riboprobes proved transcription from both DNA strands. We conclude that $\geq 60\%$ of this data set, or $\sim 1,600$ predicted sense-antisense transcriptional units, are transcribed from both DNA strands. This indicates that the occurrence of antisense transcription, usually regarded as infrequent, is a very common phenomenon in the human genome. Therefore, antisense modulation of gene expression in human cells may be a common regulatory mechanism.

Numerous examples of naturally occurring antisense transcripts have been documented in prokaryotes and viruses, where they are found to regulate gene expression by affecting mRNA transcription, processing, and translation¹. A growing number of endogenous antisense RNA transcripts have also been reported during the last several years in a variety of eukaryotic organisms^{2–5}. Antisense transcripts often code for proteins involved in diverse biological functions. Noncoding antisense transcripts have also been identified. Their role appears to be mainly regulatory⁶ and their transcription is often associated with genomic imprinting⁷. Although the effects of eukaryotic antisense RNAs on the corresponding sense RNAs have not been clearly established, a number of documented examples indicate that they may exert control at various levels of gene expression, such as transcription, mRNA processing, splicing, stability, transport, and translation^{2,8,9}.

Whatever the mechanism by which antisense RNAs alter sense expression, it is clear that the presence of double-stranded (ds) RNA (corresponding to annealed sense and antisense sequences) is a potent trigger of posttranscriptional gene regulation¹⁰. Eukaryotic cells contain specialized enzymatic machineries for processing dsRNA, such as dsRNA-specific nucleases¹¹ and dsRNA adenosine deaminase¹². Of major relevance is the recent discovery that dsRNA can trigger posttranscriptional gene silencing through a phenomenon called RNA interference¹³. This evolutionarily conserved process involves the excision of small interfering RNAs (siRNAs) from dsRNA precursors by a multidomain ribonuclease III protein named 'Dicer'.

Here we set out to study the extent of overlapping transcription in the human genome. Using a set of computational tools designed for identification and assembly of sense-antisense transcripts, we

analyzed human expressed sequences in public databases and identified 2,667 genomic loci with evidence of transcriptional units on both DNA strands. Approximately one-tenth of this hypothetical data set of sense-antisense pairs was evaluated using microarrays containing strand-specific oligonucleotide probes derived from the region of overlap. Subsequent verification of specific cases by northern blot analysis using strand-specific RNA probes, confirmed overlapping expression.

Results

In silico identification of sense-antisense gene pairs. To identify transcripts that derive from both strands of the same genomic locus, we used the output of LEADS, a software platform that cleans expressed sequences, omitting repeats, vectors, and highly abundant genes such as immunoglobulins, and then aligns them to the genome, taking into account alternative splicing (described in Supplementary Note online). Overlapping expressed sequences are assembled and combined into 'clusters' that represent genes or partial genes. Analysis of the August 2001 draft human genome sequence and the human expressed sequences (82,289 mRNAs and 3,733,145 expressed-sequence tags (ESTs)) from GenBank (version 125) with this software yielded 61,048 clusters, excluding singletons and doubletons (i.e., clusters with only one or two ESTs). Of these, 20,301 clusters contained at least one mRNA sequence—a result that in general correlates with UniGene build no. 148, which contains 20,876 such clusters. These 20,301 clusters contained 2.4 million ESTs. The remaining 40,747 clusters contained 0.34 million ESTs. The rest of the EST sequences were either discarded in the cleaning process, or found in singleton or doubleton clusters that were not analyzed.

¹Compugen Ltd., 72 Pinchas Rosen St., Tel Aviv 69512, Israel. ²Compugen Inc., 7 Centre Drive, Jamesburg, NJ 08831, USA.

*These authors contributed equally to this work. †Corresponding author (galitr@compugen.co.il).

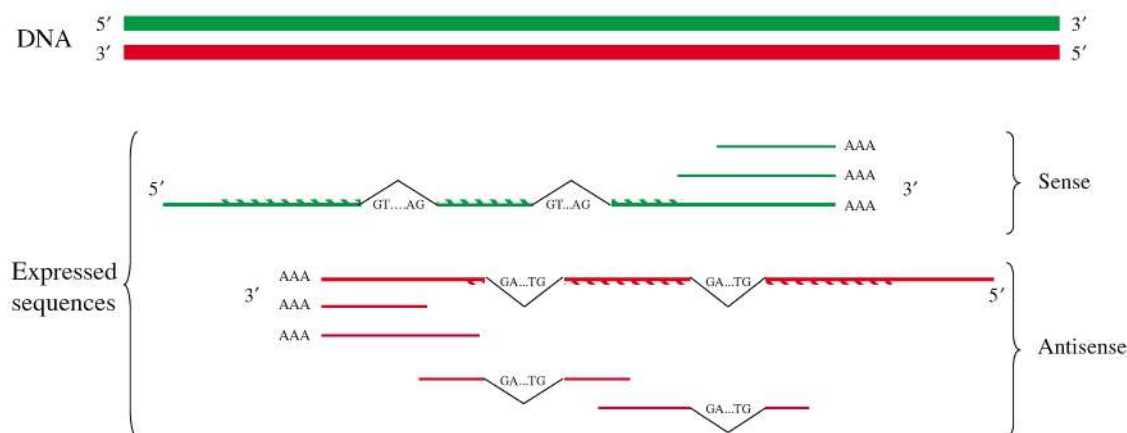


Figure 1. Schematic representation of the sources of information that can be used to deduce the DNA strand from which an expressed sequence is derived. Thin lines depict expressed sequences (ESTs or mRNAs); thick lines denote genomic DNA strands. Sequences belonging to the sense and the antisense strands are colored green and red, respectively. Splice sites begin with GT and end with AG in >98% of all known introns, thereby pointing to the strand of origin. Poly(A) tails of expressed sequences are marked. These tails, on the 3' end of the expressed sequence, disclose its strand of origin. Annotation of the coding sequence (represented here as dashed lines over the mRNA sequence) in full-length mRNAs is directional, and thus can serve as another reliable source of information.

Sequences from genes that are transcribed from opposite strands of the same genomic locus and have overlapping expressed regions are aligned by LEADS into one cluster. We therefore designed an 'Antisensor' algorithm capable of detecting clusters with sequences from opposite strands (described in detail in Supplementary Note online). The 'Antisensor' identifies the correct DNA strand for each expressed sequence in such clusters, thereby creating two separate clusters that represent overlapping genes on opposite DNA strands. The algorithm uses various sources of information (Fig. 1): (i) availability of mRNA sequences in a cluster, (ii) annotation of sequence orientation, (iii) splice junction consensus sequences, and (iv) poly(A) tail sequence. The algorithm also considers artifacts in the public ESTs database, such as genomic and intronic contamination, and misannotation of orientation.

Applying this algorithm to the 61,048 LEADS clusters resulted in a total of 2,667 clusters that contained sequences aligned to both DNA strands of the same genomic locus, and were thus predicted to represent sense-antisense pairs (listed in Supplementary Table 1 online). Each of these clusters was separated by the 'Antisensor' into two new clusters (a 'cluster pair'), each of which represents a single gene. The length of the overlaps between sense-antisense pairs in our data set varies from 20 (the minimum set by the program) to 3,393 nucleotides, with an average of 372 and a median of 247.

Characterization of the sense-antisense data set. To assess the robustness of this data set, we checked for the two most informative characteristics of a strand from which expressed sequences are transcribed: the presence of splice junctions and that of mRNAs in each cluster of the pair. Splice donor and acceptor sites are GT-AG for

>98% of introns, or GC-AG for most of the remaining introns¹⁴. Thus, expressed sequences that span an intron disclose their strand of origin quite reliably by the presence of these consensus sites at the intron's borders. The sequence of a known mRNA (i.e., a sequence annotated as 'mRNA' in GenBank) obviously indicates the DNA strand by which it is encoded, and any ambiguity about its sequence orientation is rectified by the presence of an annotated coding sequence (Fig. 1).

Table 1 summarizes the composition of the sense-antisense cluster pairs in our data set with regard to those two parameters. The vast majority (86%) contains known mRNAs in at least one of the clusters in the pair, and in 71% of the cases the expressed sequences in both clusters span at least one 'consensus' intron (flanked by consensus donor and acceptor splice sites). These results reflect the weight given to these two parameters in the selection of sense-antisense pairs by the 'Antisensor'.

The genomic organization of sense-antisense cluster pairs was further analyzed in a subset of 744 pairs that contain known mRNAs in both clusters, corresponding to 1,488 genes. Of these, 1,078 (72.5%) contained the overlap region only in their external exons, either the first exon (31% of the genes) or the last exon (41.5% of them). External 5' and 3' exons contain mainly the untranslated regions (UTRs) of mRNAs, which are known to be involved in regulation of their localization, stability, and translation¹⁵. Thus, our finding that sense-antisense overlaps appear to be confined primarily to external exons supports the notion that sense-antisense transcription may be involved in regulation of gene expression.

The possible molecular function, cellular localization, and involvement in physiological process of each of the genes in the subset that contains known mRNAs were predicted using an automatic approach (Gene Ontology Engine; ref. 16). No significant correlation was detected in terms of function, localization, or process between the two members of sense-antisense pairs in this cohort. In addition, the distribution of the genes in this group regarding these parameters was not significantly different from that of the rest of human genes (data not shown).

Table 1. Characterization of candidate sense-antisense cluster pairs^a

	No cluster with intron(s)	One cluster with intron(s)	Two clusters with intron(s)	Total
No cluster with mRNA	48	132	197	377 (14%)
One cluster with mRNA	17	490	1,039	1,546 (58%)
Two clusters with mRNA	1	85	658	744 (28%)
Total	66 (2.5%)	707 (26%)	1,894 (71%)	2,667 (100%)

^aThe table divides the predicted 2,667 pairs according to the following two parameters: number of clusters in the pair (0, 1, or 2) that contain a known mRNA; and number of clusters in the pair (0, 1, or 2) that contain at least one expressed sequence spanning an intron.

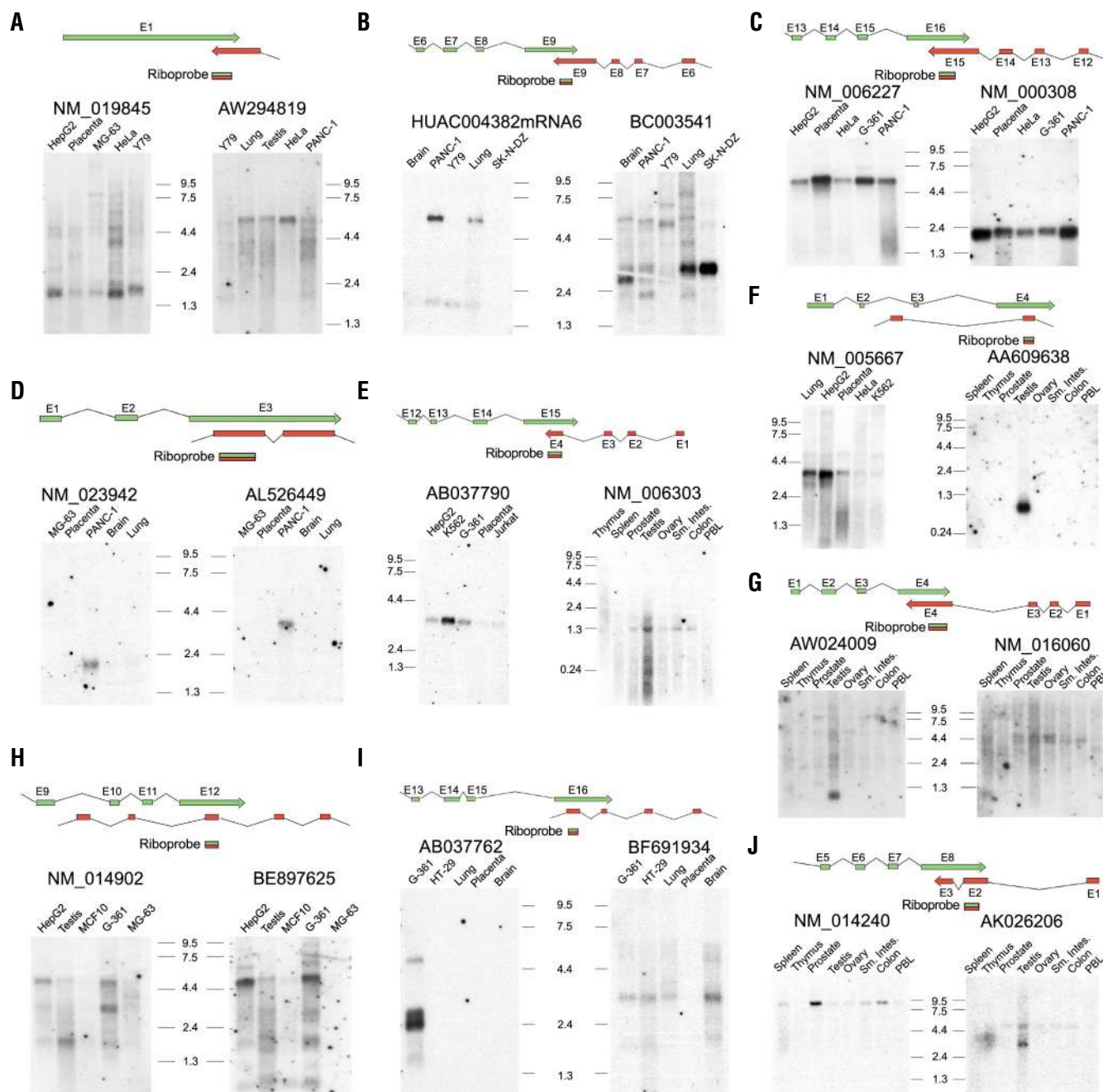


Figure 2. Validation of sense-antisense pairs by northern blot analysis. Ten pairs were selected, eight (A–H) that showed positive signals for both sense and antisense genes on the microarray analysis, and two (I–J) that showed positive signals for only one of the genes, whereas the counterpart is a known RefSeq mRNA. The diagrams describe the genomic organization of the relevant region for each of the sense-antisense genes. Exons are numbered only when an mRNA sequence is available for the cluster. Also shown is the region of overlap from which the strand-specific riboprobes were derived. The GenBank accession number of the overlapping sequence is given for each gene. Below are the autoradiograms after hybridization with the appropriate riboprobe. Northern blots contained poly(A) mRNAs from the different human tissues or cell lines. Equal loading was evaluated by ethidium bromide staining. Size markers (in kb) are indicated. PBL, peripheral blood lymphocytes.

Microarray-based experimental evaluation. For experimental validation of the predicted sense-antisense pairs in our original data set (of 2,667 pairs), we carried out a microarray-based analysis using oligonucleotide probes that hybridize to the target in a strand-specific manner. For this purpose we designed two complementary 60-mer oligonucleotide probes derived from the predicted overlap regions of the sense-antisense pairs. Single 60-mer oligonucleotides were previously shown to offer reliability and sensitivity for detecting specific transcripts¹⁷.

To select the sense-antisense pairs to be evaluated by microarrays, we first chose only those with an overlap >60 bases (2,464 pairs satisfied this restriction). The overlap region of each antisense pair was then checked for the presence of 60-mer oligonucleotides that matched a set of standards, such as minimal sequence similarity elsewhere in the human genome, uniform G+C content and T_m (melting temperature), and absence of palindromic sequences (see ‘Oligonucleotide Design’ in Supplementary Note online), to maxi-

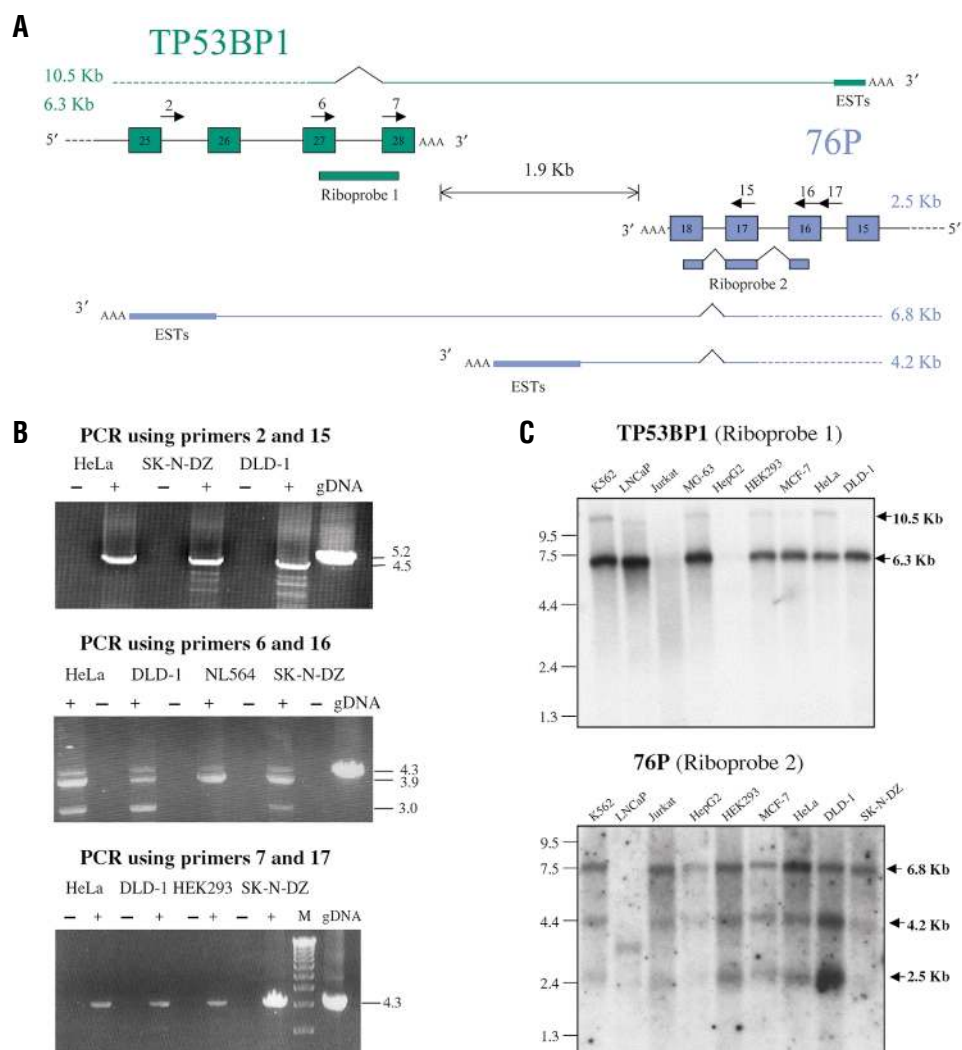


Figure 3. Analysis of the overlap region between the transcriptional units of *TP53BP1* and *76P*. (A) Schematic representation of the known mRNA of each gene (with the respective exons), and their overlapping ESTs (depicted by thick short lines, which represent several ESTs). The thin colored lines depict the predicted extensions of the respective 3' UTRs through alternative polyadenylation, which were confirmed by the experimental analysis shown in (B) and (C). The expected sizes of the known and the putative transcripts containing these alternative 3' UTRs are given. Also shown is the location of the primers (depicted as arrows) used in the RT-PCR analysis, and the region from which the riboprobes were derived for northern blot analysis. (B) PCR analysis with the primers shown in (A), using genomic DNA (gDNA) or RT products from the indicated cell lines as template. The presence (+) or absence (-) of enzyme in the RT reaction is indicated. The sizes of the PCR products that were obtained are indicated in kilobases. The sizes of the expected and observed PCR products are summarized in Table 2. The transcriptional unit from which the RT-PCR products were derived (*TP53BP1* or *76P*) was confirmed by the spanning of the respective splicing event, as detected upon sequencing. (C) Northern blot analysis with poly(A) mRNA from different cell lines, using the strand-specific riboprobes described in A. The size of transcripts detected fit the prediction of the 3' UTR extensions, as described in (A).

mize the hybridization specificity. Optimal oligonucleotide probes were found for 1,211 sense-antisense pairs. From these, a random sample of 264 pairs was selected for analysis by microarrays (listed in Supplementary Table 2 online), roughly one-tenth of the original data set of 2,667 sense-antisense pairs. In this sample, the proportion of each of the nine subgroups depicted in Table 1 is similar to that of the original data set, indicating a good representation of the various subgroups.

Microarrays were constructed by spotting each of the aforementioned oligonucleotide probes onto treated glass slides in quadruplicate. The two counterpart oligonucleotide probes, derived from the overlap region of a sense-antisense pair, were spotted next to each

other to ensure similar hybridization conditions. For positive controls we selected probes for four ubiquitously expressed genes (see Supplementary Note online). For negative controls we used two random oligonucleotides. These computer-created arbitrary sequences display no substantial alignment to the human genome (upon BLAST search using default parameters) but have the same characteristics as the other oligonucleotide probes. In addition, 22 probes for 11 previously documented sense-antisense pairs were also spotted on these microarrays.

The microarrays were hybridized with poly(A)⁺ RNAs obtained from 19 human cell lines representing a variety of tissues and four normal human tissues (see Experimental Protocol). Each poly(A)⁺ RNA was reverse-transcribed by priming with oligo(dT) and random nonamers in the presence of fluorescently labeled dNTPs. A pooled sample containing an equal mix of the RNAs from all cell lines was also transcribed and used as a reference target. The resulting fluorescently labeled cDNAs were combined and hybridized to the oligonucleotide microarrays. The experiments were done in duplicate with a fluorescent reversal of the cyanin-3 (Cy3)- and Cy5-labeled cDNA and pool. We first tested different conditions of hybridization and chose stringent conditions so as to minimize the appearance of false-positive signals, although this could compromise the detection of low-abundance transcripts. (The raw data from the microarray experiments is presented in http://www.labonweb.com/anti-sense/Raw_Data.)

The raw data were normalized at several levels: within each slide, between reciprocal slides, and globally among slides (see Supplementary Note online). Nonspecific levels of hybridization were estimated from the negative controls. The threshold

for a significant positive signal resulting from authentic hybridization was set at 4 standard deviations of the mean normalized signals of the negative controls. The processed data are presented as normalized signal intensity and as normalized signal ratios (Supplementary Table 3 online). Positive signals were obtained for both sense and antisense transcripts in 65 cluster pairs. In another 47 cases, we detected significant hybridization signals for antisense sequences whose counterpart sense transcripts are known RefSeq (<http://www.ncbi.nlm.nih.gov/LocusLink/RSfaq.html>) mRNAs that did not give clear hybridization signals on our microarrays. Thus, in a total of 112 cases, or 42.5% of the 264 sense-antisense pairs represented on the microarrays, we could detect antisense transcription.

Table 2. Analysis of the overlap between *TP53BP1* and *76P* by PCR^a

Primers	TP53BP1		76P		gDNA	
	Exp	Obs	Exp	Obs	Exp	Obs
2 + 15	/	/	4.5	4.5	5.2	5.2
6 + 16	3.9	3.9	3.0	3.0	4.3	4.3
7 + 17	4.3	4.3	/	/	4.3	4.3

^aSizes (in kb) of expected and observed PCR products (in Fig. 3B) using different primer pairs for *TP53BP1* and *76P*, on genomic DNA (gDNA) or RT products as template.

The sensitivity of our experimental approach, in other words our ability to detect a given transcript, is determined by the stringency of the microarray hybridization conditions and the tissue specificity and expression levels of the mRNA. This sensitivity can be estimated from the positive signals that we obtained for 65% of the oligos representing known RefSeq mRNAs on the microarrays. This level of detection is comparable to that obtained in other studies, such as the 58% of known exons verified using microarray analysis¹⁸. To detect a sense-antisense pair we need to detect two genes, each of which has a 0.65 probability of detection. Thus, our level of detection (sensitivity) for a sense-antisense pair is $0.65 \times 0.65 = 0.42$. Of the 264 predicted sense-antisense pairs that were analyzed with microarrays, the 65 cases that showed significant signals for both sense and antisense transcripts represent the 0.42 fraction that we expect to detect from the real sense-antisense pairs on the microarray. Thus, there were actually 154 (60/0.42) such pairs on the microarray, and therefore at least 60% (154 of 264) of the clusters we analyzed represented true cases of overlapping antisense transcription. Extrapolating this figure to our predicted antisense data set of 2,667 clusters, we conclude that there are $\geq 1,600$ sense-antisense transcriptional units in the human genome.

RNA-based strand-specific verification. From the sense-antisense pairs showing positive signals on the microarrays, we chose ten cases for further verification by northern blot analysis. Eight of the selected pairs had shown positive signals for both the sense and antisense genes in the microarray analysis, and two pairs were chosen because only one of the genes passed the validation by microarrays and its counterpart oligo was fully contained within a known RefSeq mRNA. As seen in Figure 2A–J, the presence of transcripts with complementary sequences was confirmed for all ten gene pairs through hybridization with strand-specific riboprobes derived from the predicted region of overlap.

Two additional sense-antisense pairs were selected for detailed study of the extent of overlap. Analysis using bioinformatic tools pointed to the presence of neighboring ESTs in the antisense orientation in both cases—suggesting alternative polyadenylation that results in long 3' UTRs and an extensive overlap between the two transcriptional units. The known mRNAs of the first pair, *TP53BP1* (the gene encoding 53BP1, p53-binding protein-1, ref. 19) and *76P* (the gene encoding 76p, a member of the γ -tubulin-associated protein family, ref. 20), do not overlap but are derived from opposite strands of the same genomic locus, within 1.9 kilobases from each other's 3' end (Fig. 3A). The presence of ESTs in the antisense orientation in both genes suggested the existence of long alternative 3' UTRs, creating an overlap of several kilobases between the two genes (Fig. 3A). This assumption was verified by a RT-PCR-based approach (Fig. 3B). The newly predicted transcripts were further confirmed by northern blot analysis using strand-specific riboprobes (Fig. 3C). A similar situation was observed between the transcriptional units of *CCNE2* (the gene encoding cyclin E2, refs. 21, 22) and *FLJ20530* (RefSeq mRNA no. NM_017864.1, encoding a

hypothetical protein) (Fig. 4A). The apparent overlap between the two known mRNAs is only 75 base pairs long. However, the presence of ESTs in the antisense direction to *CCNE2* suggested alternative polyadenylation, resulting in a much longer 3' UTR of the *FLJ20530*, and an overlap of several kilobases between the two genes. This was again confirmed by RT-PCR and northern blot analysis (Fig. 4B,C). These results indicate that the true overlap between sense and antisense genes may be much longer than that expected from analysis of mRNAs deposited in GenBank.

Discussion

The results of this study, which integrates computational tools with experimental methods, suggest the presence of $\geq 1,600$ sense-antisense gene pairs in the human genome. This finding indicates that the occurrence of loci transcribed from both DNA strands in the human genome is a common phenomenon. We believe that naturally occurring antisense transcripts may be even more prevalent than estimated from this study for the following reasons: (i) Our computational tools enable detection of sense-antisense genes only if both sense and antisense expressed sequences were deposited in the public databases, and the transcripts overlap within exonic regions. Antisense genes for which expressed sequences are not yet available, or that are too short to show the overlap, will be missed by our detection tools. (ii) Our methods led us to the identification of *cis*-antisense transcripts in the human transcriptome, in other words, transcripts that are transcribed from opposite DNA strands of the same locus. *Trans*-encoded antisense RNAs (those transcribed from another locus) cannot be detected by our tools. (iii) Antisense transcripts that do not span introns, or do not contain clear poly(A) tails, are more likely to be rejected by our current tools.

In addition, it appears that the overlap between the sense-antisense transcriptional units is often more extensive than can be estimated from the sequences in the public databases. This is due to underrepresentation of long 3' UTRs, particularly in the cDNA databases, and therefore the overlap between genes is frequently established by complementary ESTs, even when mRNAs are present in the clusters. This observation is supported by our detailed analysis of two sense-antisense pairs, *TP53BP1-76P* and *CCNE2-FLJ20530*.

Three recent papers, one published during the course of our work and two while it was being reviewed, report the computational discovery of transcripts with the potential for sense-antisense pairing^{23–25}. In two of these reports (refs. 23 and 24) the study was limited to BLAST analysis of publicly available, full-length human mRNAs (RefSeq database containing $\sim 12,000$ human mRNAs). Using this approach, Lehner *et al.* (ref. 23) reported the *in silico* identification of 87 *cis*-antisense human transcripts, whereas Fahey *et al.* (ref. 24) identified 56 such pairs. By extrapolation, both groups estimated ~ 800 pairs of genes with sense-antisense transcription in the human genome, assuming 40,000 genes. Lehner *et al.* refrained from using EST databases because of the uncertainties regarding the correct orientation of ESTs. Fahey *et al.* tried to use EST databases but found them unreliable for the detection of antisense transcription, again

Table 3. Analysis of the overlap between *CCNE2* and *FLJ20530* by PCR^a

Primers	<i>CCNE2</i>		<i>FLJ20530</i>		gDNA	
	Exp	Obs	Exp	Obs	Exp	Obs
1 + 2	1.6	1.6	2.0	2.0	2.0	2.0
1 + 3	/	/	2.1	2.1	5.9	5.9

^aSizes (in kb) of expected and observed PCR products (in Fig. 4B) using two primer pairs for *CCNE2* and *FLJ20530*, on genomic DNA (gDNA) or RT products as template.

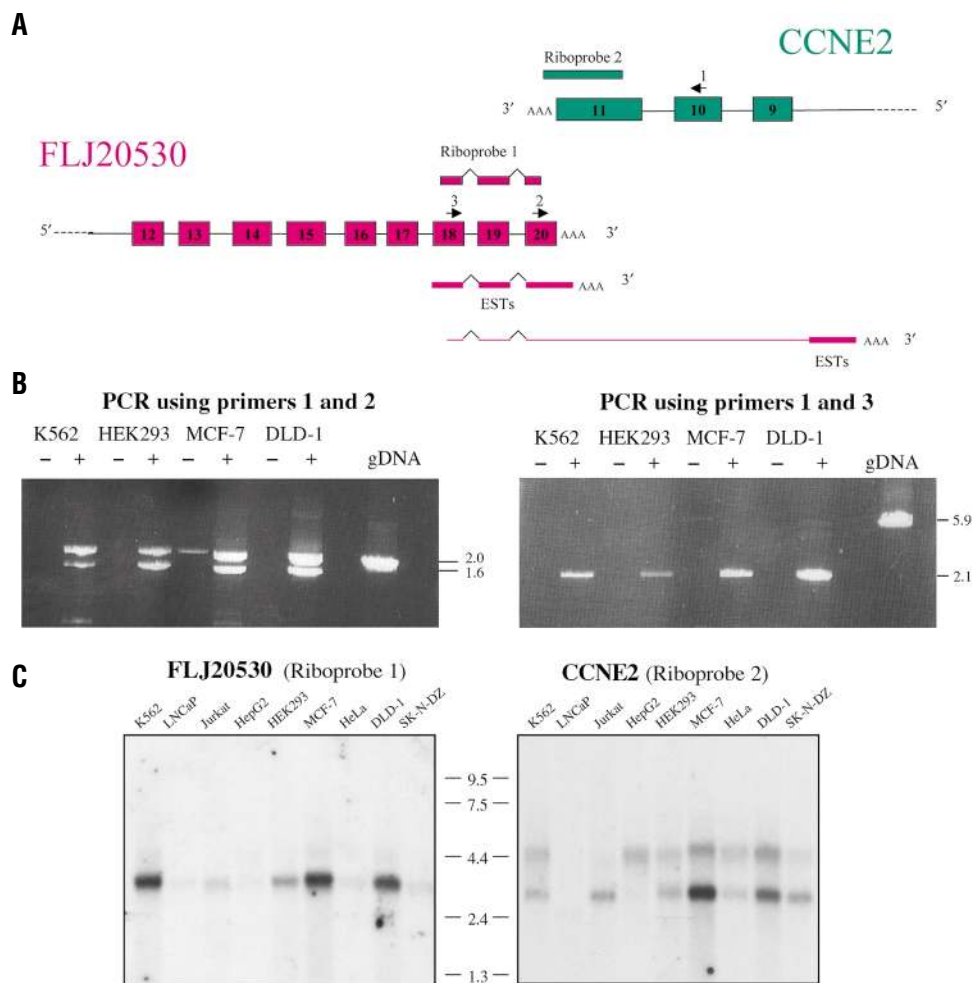


Figure 4. Analysis of the overlap region between the transcriptional units of *CCNE2* and *FLJ20530*. (A) Schematic representation showing the known mRNAs (with the respective exons) and the ESTs (short thick lines) that suggest an extension of the 3' UTR of *FLJ20530* (depicted by the thin pink line), which was confirmed by RT-PCR as shown in (B). Also shown are the location of the primers (depicted as arrows) used in the RT-PCR analysis, and the region from which the riboprobes were derived. (B) PCR analysis using the primers shown in (A), essentially as described in Figure 3B. The sizes of the expected and observed PCR products are summarized in Table 3. (C) Northern blot analysis on poly(A) mRNAs from different cell lines, using the strand-specific riboprobes described in (A).

because of the problematic assignment of ESTs orientation. Our design of the appropriate computational tools facilitates the use of the vast public data on millions of human ESTs and partial cDNAs, permitting their assembly, clustering, and alignment to the correct DNA strand. Thus, our approach provides a more realistic estimate of the extent of antisense transcription. Furthermore, we validate our *in silico* findings with experimental evidence.

The recent study by Shendure and Church (ref. 25) analyzed databases of expressed sequences using publicly available bioinformatic tools. To avoid problems related to the correct orientation of ESTs, they used only sequences from high-quality directionally cloned EST libraries. They identified 144 candidate overlapping transcriptional units in the human genome. They also carried out experimental validation of a subset of their predictions by orientation-specific RT-PCR, and found evidence of antisense transcription in 33 out of 39 cases tested. The number of sense-antisense loci predicted by our computational tools, and experimentally validated using oligonucleotide microarrays, exceeds those of Shendure and Church by roughly an order of magnitude. Furthermore, their experimental

analysis is limited to one tissue, whereas we analyzed >20 human tissues by both microarray and northern blot analysis, and carried out further in-depth characterization of two specific examples. A comparison between our data set of sense-antisense pairs and those predicted in the studies just described showed that >85% of their pairs appear in our data set. Manual examination of their pairs that were missing in our data set showed most of these to be artifacts. A detailed description of this comparison is given in the Supplementary Note online).

Our estimate of $\geq 1,600$ sense-antisense pairs (i.e., 3,200 genes) suggests that >8% of the estimated 40,000 human genes have an antisense partner. As mentioned by Shendure and Church²⁵, further support for this estimate comes from the work of Shoemaker *et al.*¹⁸, who used 60-mer oligonucleotides to validate a large number of predicted novel exons by microarray analysis. As negative controls, they used the reverse-complement probes for ~78,000 exons annotated as 'confirmed'. Using two sources of mRNA, they observed hybridization to ~5% of these supposedly 'negative controls'. In light of our results, we believe that some of these 'false positives' could actually stem from antisense transcription.

Widespread regulation of gene expression by antisense transcription offers an explanation for the puzzling observation of ortholog-specific conservation of 3' UTRs in >30% of vertebrate mRNAs²⁶. Lipman proposed that such long conserved blocks of unique noncoding sequences, previously

shown to be associated with mRNA stability, are involved in forming long perfect duplexes with antisense transcripts²⁷. Of relevance is our observation that ~70% of the sense-antisense genes overlap in their 5' or 3' external exons, giving further support to the notion that sense-antisense overlap could be involved in gene regulation, as external exons contain the untranslated regulatory regions of mRNAs.

Knowledge of human antisense transcription can be of help in the study of RNA interference, an important biological phenomenon that is based on the presence of dsRNA and leads to degradation of complementary mRNAs and to gene silencing. Such information may also be important for the selection of synthetic antisense oligonucleotides in functional studies and drug design²⁸. The broad occurrence of complementary transcripts in the human transcriptome has clear implications for the interpretation of experimental data resulting from the use of dsDNA probes, such as northern blot analysis and cDNA microarrays. Clearly, both experimental approaches would benefit from the exclusive use of strand-specific probes.

Recent reports in the literature indicate that antisense transcription in other eukaryotes may be more common than had been

thought. Kumar *et al.* (ref. 29) reported that out of 137 newly discovered open reading frames in yeast, 79 are predicted to lie opposite previously annotated genes. Antisense transcription in plants has also been shown in a growing number of cases⁵.

Taken together, the above findings and our results suggest that large numbers of antisense transcripts are found throughout the eukaryotic world and may play a role in antisense-mediated gene regulation, as is the case in prokaryotes. Our data set forms a basis for future studies on the modulation of gene expression by antisense transcription in eukaryotes.

Experimental protocol

Human cell lines. The following cell lines were used: MCF-7 (breast adenocarcinoma, American Type Culture Collection (ATCC) no. HTB-22), HeLa (cervical adenocarcinoma, ATCC no. CCL-2), HEK-293 (embryonal kidney cells, ATCC no. CRL-1573), Jurkat (acute T-cell leukemia, ATCC no. TIB-152), K562 (chronic myelogenous leukemia, ATCC no. CCL-243), HepG2 (liver carcinoma, ATCC no. HB-8065), T24 (urinary bladder carcinoma, ATCC no. HTB-4), SK-N-DZ (neuroblastoma, ATCC no. CRL-2149), NK-92 (non-Hodgkin's lymphoma, ATCC no. CRL-2407), MG-63 (osteosarcoma, ATCC no. CRL-1427), DU 145 (prostatic carcinoma, ATCC no. HTB-81), G-361 (melanoma, ATCC no. CRL-1424), PANC-1 (pancreatic carcinoma, ATCC no. CRL-1469), ES-2 (ovary clear-cell carcinoma, ATCC no. CRL-1978), Y79 (retinoblastoma, ATCC no. HTB-18), HT-29 (colorectal adenocarcinoma, ATCC no. HTB-38), DLD-1 (colorectal adenocarcinoma, ATCC no. CCL-221), H1299 (large-cell lung carcinoma, ATCC no. CRL-5803), NL564 (Epstein-Barr virus (EBV)-transformed human lymphoblasts), SNU1 (gastric carcinoma, ATCC no. CRL-5971), LNCaP (prostatic adenocarcinoma), and MCF10 (benign breast tissue cells).

RNA purification. Total RNA was extracted from the aforementioned human cell lines using TriReagent (Molecular Research Center, Cincinnati, OH). Poly(A)⁺ mRNA was purified using two cycles of the Dynabeads mRNA Purification Kit (DynaL Biotech ASA, Oslo, Norway). The removal of traces of ribosomal RNA was confirmed by agarose gel electrophoresis. Poly(A)⁺ mRNA from human testis, placenta, lung, and brain was purchased from BioChain Institute, Inc. (Hayward, CA).

RT-PCR. Reverse transcription (RT) was carried out in a final volume of 20 μ l using 2 μ g of total RNA and 2.5 units of Superscript II Reverse Transcriptase (Invitrogen, Paisley, UK), in the buffer supplied by the manufacturer, supplemented with 10 pmol of oligo(dT)₁₅ (Promega, Madison, WI) and 30 units of RNasin (Promega). PCR was carried out using 1 μ l of the RT reaction, in the presence of 2 mM dNTPs, 25 pmol of primers, and 2.5 units of DNA Pol mix of the Expand Long Template PCR System (Roche, Mannheim, Germany), in the reaction buffer supplied by the manufacturer. RT reactions carried out in the absence of enzyme were used as negative controls for genomic DNA contamination.

The following primers were used for the RT-PCR analysis shown in Figure 3: Primer 2, 5'-ggtaggaggttcttctgtgaacc-3'; primer 6, 5'-ccagaatctgaaggtactcttg-3'; primer 7, 5'-tcacaagagtggtgatccagtg-3'; primer 17, 5'-ccttaag-gagctcacaatctagttgg-3'; primer 16, 5'-cgctggtcagtcagaacctag-3'; primer 15, 5'-cactctgttcaagattctctcc-3'.

Primers used for the analysis shown in Figure 4 are as follow: Primer 1, 5'-ctagtccagtgagctgaagac-3'; primer 2, 5'-tacagctggcagcgagagaag-3'; primer 3, 5'-cctactacgactacatattggatg-3'.

Riboprobe synthesis. For each sense-antisense pair analyzed in Figure 2, the overlapping riboprobes were designed to include the 60-mer oligonucleotides used in the microarray analysis. Riboprobes were obtained by RT-PCR on selected RTs. The sequence of the primers used to obtain the various

RT-PCR products are given in Supplementary Table 4 online. RT-PCR products were electrophoresed on agarose gels, extracted with Gel Extraction Kit (Qiagen, Hilden, Germany), and ligated to the T7 adapter using the Lig'nScribe Kit (Ambion, Austin, TX). The P2 primer (in the T7 adapter) and a forward-specific primer were used to amplify the sense-specific riboprobe, and a reverse-specific primer to amplify the antisense-specific riboprobe. These PCR products were then separated on agarose gels, and extracted as described earlier. Each one was subsequently sequenced with the appropriate primers for verification. For labeling, 50–100 ng of each PCR product was used per 500 bp riboprobe, in a reaction containing α -[³²P]UTP and T7 polymerase, using the SP6/T7 Transcription Kit (Roche, Mannheim, Germany), as suggested by the manufacturer. The riboprobe was separated from the nonincorporated ribonucleoside triphosphates using Mini Quick Spin RNA Columns (Roche, Indianapolis, IN).

The following primers were used to obtain riboprobes for the northern blot analysis shown in Figure 3C: Riboprobe 1, primer R1, 5'-ccagaatctgaaggtactcttg-3'; primer F1, 5'-cacaatctccagtagcagg-3'. Riboprobe 2, primer R2, 5'-gctagaattgccaatctgttag-3'; primer F2, 5'-cgctggtcagtcagaacctag-3'.

The following primers were used to obtain riboprobes for the northern blot analysis shown in Figure 4C: Riboprobe 1, primer F1, 5'-cctactacgactacatattggatg-3'; primer R1, 5'-gctctagacagatattgtaattctcttcc-3'. Riboprobe 2, primer F2, 5'-gacagagttgaatgcaagcaatcc-3'; primer R2, 5'-gctgtggctcttcc-taactgg-3'.

Northern blot analysis. Poly(A)⁺ RNA (1 μ g) of human cells or tissues was fractionated by electrophoresis on 1.0% (wt/vol) agarose gels containing formaldehyde and blotted onto Nytran Super Charge membranes (Schleicher & Schuell, Dassel, Germany). Equal loading was verified by ethidium bromide staining of the gels. In addition, we used commercial northern blots containing poly(A)⁺ RNA samples from a variety of tissues (Human Multiple Tissue Northern blots; BD Biosciences, Palo Alto, CA). Northern blots were prehybridized at 68 °C for 30 min in Ultrahyb Buffer (Ambion, Austin, TX). ³²P-labeled riboprobe (at 10⁶ c.p.m./ml) was added and the hybridization was carried out for 16 h at 68 °C. The membranes were rinsed twice with 2 \times SSC, 0.1% (wt/vol) SDS at room temperature, washed twice for 5 min each with 2 \times SSC, 0.1% SDS at 68 °C, and a last wash for 30 min with 0.1 \times SSC, 0.1% SDS at 68 °C. Autoradiograms were obtained by exposing the membranes to BioMax MS film (Kodak, Rochester, NY) for various times, according to the intensity of the signals.

URL. The LEADS-Antisensor is accessible for academic research through the following website: <http://www.labonweb.com/antisense>.

Raw data from microarray experiments can be found at: http://www.labonweb.com/antisense/Raw_Data.

Note: Supplementary information is available on the Nature Biotechnology website.

Acknowledgments

We want to thank Alon Wasserman and Brian Meloon for their critical help with the Antisensor design and analysis of microarray results; Mervi Heiskanen, Ruth Goldin, and Paul Nisson for their valuable assistance in the setup of the microarray methodology; Han Xie for the GO engine analysis; Dan Szybel for technical help; Raveh Gill-More, Gady Cojocar, Naftali Kaminski, and Gidi Rechavi for critical reading of the manuscript and helpful discussions; Sarah Pollock for her innovative ideas, and Salomon Langer for his inspirational support.

Competing interests statement

The authors declare competing financial interests: see the Nature Biotechnology website (<http://www.nature.com/naturebiotechnology>) for details.

Received 28 August 2002; accepted 7 January 2003

1. Wagner, E.G. & Simons, R.W. Antisense RNA control in bacteria, phages, and plasmids. *Annu. Rev. Microbiol.* **48**, 713–742 (1994).
2. Knee, R. & Murphy, P.R. Regulation of gene expression by natural antisense RNA transcripts. *Neurochem. Int.* **31**, 379–392 (1997).
3. Kumar, M. & Carmichael, G.G. Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes. *Microbiol. Mol. Biol. Rev.* **62**, 1415–1434 (1998).
4. Vanhee-Brossollet, C. & Vaquero, C. Do natural antisense transcripts make sense in eukaryotes? *Gene* **211**, 1–9 (1998).
5. Terryn, N. & Rouze, P. The sense of naturally transcribed antisense RNAs in plants. *Trends Plant Sci.* **5**, 394–396 (2000).
6. Kelly, R.L. & Kuroda, M.I. Noncoding RNA genes in dosage compensation and imprinting. *Cell* **103**, 9–12 (2000).
7. Sleutels, F., Barlow, D.P. & Lyle, R. The uniqueness of the imprinting mechanism. *Curr. Opin. Genet. Dev.* **10**, 229–233 (2000).
8. Li, A.W. & Murphy, P.R. Expression of alternatively spliced FGF-2 antisense RNA transcripts in the central nervous system: regulation of FGF-2 mRNA translation. *Mol. Cell. Endocrinol.* **162**, 69–78 (2000).
9. Hastings, M.L., Ingle, H.A., Lazar, M.A. & Munroe, S.H. Post-transcriptional regulation of thyroid hormone receptor expression by *cis*-acting sequences and a naturally occurring antisense RNA. *J. Biol. Chem.* **275**, 11507–11513 (2000).
10. Bass, B.L. Double-stranded RNA as a template for gene silencing. *Cell* **101**, 235–238 (2000).
11. Wu, H., MacLeod, A.R., Lima, W.F. & Crooke, S.T. Identification and partial purification of human double strand RNase activity. A novel terminating mechanism for oligoribonucleotide antisense drugs. *J. Biol. Chem.* **273**, 2532–2542 (1998).
12. Bass, B.L. RNA editing and hypermutation by adenosine deamination. *Trends Biochem. Sci.* **22**, 157–162 (1997).
13. Zamore, P.D. Ancient pathways programmed by small RNAs. *Science* **296**, 1265–1269 (2002).
14. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
15. Pesole, G. *et al.* Structural and functional features of eukaryotic mRNA untranslated regions. *Gene* **276**, 73–81 (2001).
16. Xie, H. *et al.* Large-scale protein annotation through gene ontology. *Genome Res.* **12**, 785–794 (2002).
17. Hughes, T.R. *et al.* Expression profiling using microarrays fabricated by an ink-jet oligonucleotides synthesizer. *Nat. Biotechnol.* **19**, 342–347 (2001).
18. Shoemaker, D.D. *et al.* Experimental annotation of the human genome using microarray technology. *Nature* **409**, 922–927 (2001).
19. Iwabuchi, K., Bartel, P.L., Li, B., Marraccino, R. & Fields, S. Two cellular proteins that bind to wild-type but not mutant p53. *Proc. Natl. Acad. Sci. USA* **91**, 6098–6102 (1994).
20. Fava, F. *et al.* Human 76p: a new member of the γ -tubulin-associated protein family. *J. Cell Biol.* **147**, 857–868 (1999).
21. Lauper, N. *et al.* Cyclin E2: a novel CDK2 partner in the late G1 and S phases of the mammalian cell cycle. *Oncogene* **17**, 2637–2643 (1998).
22. Zariwala, M., Liu, J. & Xiong, Y. Cyclin E2, a novel human G1 cyclin and activating partner of CDK2 and CDK3, is induced by viral oncoproteins. *Oncogene* **17**, 2787–2798 (1998).
23. Lehner, B., William, G., Campbell, R.D. & Sanderson, C.M. Antisense transcripts in the human genome. *Trends Genet.* **18**, 63–65 (2002).
24. Fahey, M.E., Moore, T.F. & Higgins, D.G. Overlapping antisense transcription in the human genome. *Comp. Funct. Genom.* **3**, 244–253 (2002).
25. Shendure, J. & Church, G.M. Computational discovery of sense-antisense transcription in the human and mouse genomes. *Genome Biol.* **3**, 1–14 (2002).
26. Duret, L., Dorkeld, F. & Gautier, C. Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression. *Nucleic Acids Res.* **21**, 2315–2322 (1993).
27. Lipman, D.J. Making (anti)sense of non-coding sequence conservation. *Nucleic Acids Res.* **25**, 3580–3583 (1997).
28. Delihias, N., Rokita, S.E. & Zheng, P. Natural antisense RNA/target RNA interactions: possible models for antisense oligonucleotide drug design. *Nat. Biotechnol.* **15**, 751–753 (1997).
29. Kumar, A. *et al.* An integrated approach for finding overlooked genes in yeast. *Nat. Biotechnol.* **20**, 58–63 (2002).