

**The following resources related to this article are available online at [www.sciencemag.org](http://www.sciencemag.org) (this information is current as of November 26, 2009 ):**

**Updated information and services**, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/322/5909/1855>

**Supporting Online Material** can be found at:

<http://www.sciencemag.org/cgi/content/full/1163853/DC1>

A list of selected additional articles on the Science Web sites **related to this article** can be found at:

<http://www.sciencemag.org/cgi/content/full/322/5909/1855#related-content>

This article **cites 25 articles**, 8 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/322/5909/1855#otherarticles>

This article has been **cited by** 24 article(s) on the ISI Web of Science.

This article has been **cited by** 11 articles hosted by HighWire Press; see:

<http://www.sciencemag.org/cgi/content/full/322/5909/1855#otherarticles>

This article appears in the following **subject collections**:

Molecular Biology

[http://www.sciencemag.org/cgi/collection/molec\\_biol](http://www.sciencemag.org/cgi/collection/molec_biol)

Information about obtaining **reprints** of this article or about obtaining **permission to reproduce this article** in whole or in part can be found at:

<http://www.sciencemag.org/about/permissions.dtl>

# The Antisense Transcriptomes of Human Cells

Yiping He, Bert Vogelstein, Victor E. Velculescu, Nickolas Papadopoulos,\* Kenneth W. Kinzler

Transcription in mammalian cells can be assessed at a genome-wide level, but it has been difficult to reliably determine whether individual transcripts are derived from the plus or minus strands of chromosomes. This distinction can be critical for understanding the relationship between known transcripts (sense) and the complementary antisense transcripts that may regulate them. Here, we describe a technique that can be used to (i) identify the DNA strand of origin for any particular RNA transcript, and (ii) quantify the number of sense and antisense transcripts from expressed genes at a global level. We examined five different human cell types and in each case found evidence for antisense transcripts in 2900 to 6400 human genes. The distribution of antisense transcripts was distinct from that of sense transcripts, was nonrandom across the genome, and differed among cell types. Antisense transcripts thus appear to be a pervasive feature of human cells, which suggests that they are a fundamental component of gene regulation.

The DNA in each normal human cell is virtually identical. The key to cellular differentiation therefore lies in understanding the gene products—transcripts and proteins—that are derived from the genome. For more than a decade, it has been possible to measure the levels of transcripts in a cell at the whole-genome level (1). The word “transcriptome” was coined to denote this genome-wide assessment (2). However, it has been difficult to determine which of the two strands of the chromosome (plus or minus) serves as the template for transcripts in a global fashion. Sense transcripts of protein-encoding genes produce functional proteins, whereas antisense transcripts are often thought to have a regulatory role (3–7).

Several unequivocal examples of antisense transcripts, such as those corresponding to imprinted genes, have been described [reviewed in (3–7)]. However, estimates of the fraction of genes associated with antisense transcripts in mammalian cells vary from less than 2% to more than 70% of the total genes (8–18). We have developed a technique called asymmetric strand-specific analysis of gene expression (ASSAGE) that allows unambiguous assignment of the DNA strand coding for a transcript. The key to this approach is the treatment of RNA with bisulfite, which changes all cytidine residues to uridine residues. The sequence of a bisulfite-treated RNA molecule can only be matched to one of the two possible DNA template strands (fig. S1). After generating cDNA from bisulfite-treated RNA with reverse transcriptase (RT), sequencing of the reverse transcription polymerase chain reaction (RT-PCR) product can be used to establish whether a particular RNA was transcribed from the plus or minus strand. To identify the DNA

strands of origin for the entire transcriptome, we ligate cDNA fragments derived from bisulfite-treated RNA to adapters and then determine the sequence of one end of each fragment through sequencing-by-synthesis. The number and distribution of the sequenced tags provide information about the level of transcription of each gene in the analyzed cell population as well as the strand from which each transcript was derived.

We used ASSAGE to study transcription in normal human peripheral blood mononuclear cells (PBMCs). Several quality controls were performed to evaluate the library of tags derived

from this RNA source. First, we calculated the bisulfite conversion efficiency from the sequences of the tags and found that 95% of the C residues in the original RNA had been converted to U residues (19). Second, we determined whether the bisulfite treatment altered the distribution of tags by preparing libraries without bisulfite treatment. We found a good correlation between the number of sense tags in a gene derived from ASSAGE data and the number of tags derived from sequencing of DNA synthesized from the same RNA used for ASSAGE without bisulfite treatment from the same cells ( $R^2 = 0.59$ ). We also found a correlation between the relative expression levels determined by ASSAGE and those assessed by hybridization to microarrays [ $R^2 = 0.45$  (19)].

From the PBMC tag library, 4 million experimental tags could be unambiguously assigned to a specific genomic position in the converted genome (table S1). Of the 4 million tags, 47.5% had the sequence of the plus strand (that is, the template of these transcripts had been the minus strand), and 52.5% had the sequence of the minus strand. This is consistent with the expected equal distribution of sense transcripts from the two strands (20). As shown in table S1, 90.3% of the 4 million tags could be assigned to known genes; the remaining tags were in unannotated regions of the genome. The fraction of unannotated tags (9.7%) is consistent with data from other sources indicating the likely existence of actively transcribed genes in human cells that have not yet been discovered or annotated (6, 21–24). Of the

**Table 1.** Classification of genes with respect to antisense tags. We classified only those genes whose sum of distinct sense and antisense tags was 5 or more. S genes contained only sense tags or had a sense/antisense tag ratio of 5 or more; AS genes contained only antisense tags or had a sense/antisense tag ratio of 0.2 or less; SAS genes contained both sense and antisense tags and had a sense/antisense ratio between 0.2 and 5. Samples were derived from the following sources: PBMC, peripheral blood mononuclear cells isolated from a healthy volunteer; Jurkat, a T cell leukemia line; HCT116, a colorectal cancer cell line; MiaPaCa2, a pancreatic cancer line; MRC5, a fibroblast cell line derived from normal lung.

	Cell type									
	PBMC		Jurkat		HCT116		MiaPaCa2		MRC5	
	No. of genes	Fraction	No. of genes	Fraction	No. of genes	Fraction	No. of genes	Fraction	No. of genes	Fraction
<i>All genes</i>										
S genes	10,586	81.60%	9,928	89.60%	11,176	88.00%	9,500	89.50%	10,165	89.30%
AS genes	329	2.50%	240	2.20%	203	1.60%	155	1.50%	212	1.9%
SAS genes	2,061	15.9%	908	8.2%	1,327	10.4%	959	9%	1,002	8.8%
Total	12,976		11,076		12,706		10,614		11,379	
<i>Coding genes</i>										
S genes	10,375	81.30%	9,778	89.50%	10,770	87.60%	9,348	89.40%	10,029	89.20%
AS genes	325	2.50%	239	2.20%	201	1.60%	154	1.50%	210	2%
SAS genes	2,055	16.1%	907	8.3%	1,325	10.8%	959	9.2%	1,000	8.9%
Total	12,755		10,924		12,296		10,461		11,239	
<i>Noncoding genes</i>										
S genes	211	95.50%	150	98.70%	406	99.00%	152	99.30%	136	97.10%
AS genes	4	1.80%	1	0.70%	2	0.50%	1	0.70%	2	1.4%
SAS genes	6	2.7%	1	0.70%	2	0.50%	0	0%	2	1.4%
Total	221		152		410		153		140	

Ludwig Center for Cancer Genetics and Therapeutics and Howard Hughes Medical Institute, Johns Hopkins Kimmel Cancer Center, Baltimore, MD 21231, USA.

\*To whom correspondence should be addressed. E-mail: npapado1@jhmi.edu

informative tags in annotated regions, 11% were antisense and 89% were sense (table S1).

We next assessed the expression of each gene by counting the total number of tags matching a gene or by counting tags with identical sequence matching a gene only once (distinct tags). On average, there were three total tags for each distinct tag, but this number varied widely and reflected the level of expression of the corresponding transcript. With respect to antisense transcription, genes could be divided into three main classes. S genes were defined as those with a  $\geq 5:1$  ratio of distinct sense tags to distinct antisense tags; AS genes were defined as those with a  $\geq 5:1$  ratio of distinct antisense tags to distinct

sense tags. The SAS class included the remaining genes, all of which contained both sense and antisense tags. In PBMCs, we identified 329 (2.5%) AS genes, 2061 (15.9%) SAS genes, and 10,586 (81.6%) S genes among the 12,976 Ensembl genes in which at least five distinct tags were observed (Table 1 and table S2). There were 6457 genes in which at least two distinct antisense tags were found.

When normalized by length, there was an obvious concentration of antisense tags in exons relative to the entire genome or to introns ( $P < 0.0001$ ; Fig. 1). Within promoter regions, there was a concentration of antisense tags near the transcription initiation site of the sense transcripts,

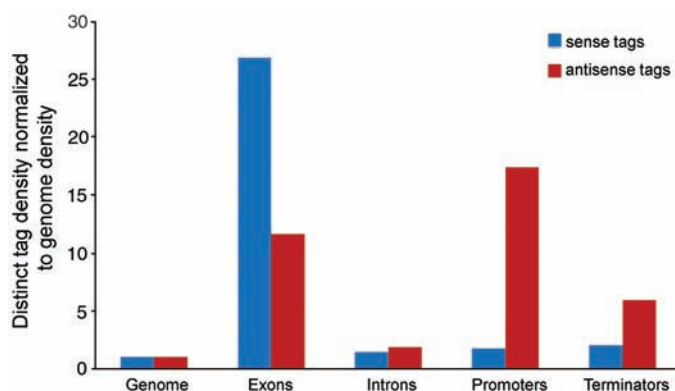
which gradually tapered off upstream ( $P < 0.01$ ; Fig. 1 and fig. S2). We also found clear differences between the relative distributions of sense and antisense tags, with a higher proportion of antisense tags than sense tags within promoter and terminator regions of genes ( $P < 0.0001$ ; Fig. 1). Examples of the distribution of sense and antisense tags derived from S and AS genes are shown in Fig. 2 and fig. S3. The predicted AS transcripts could be confirmed by ASSAGE using gene-specific primers (fig. S4).

To determine whether the patterns described above were particular to PBMCs, we used ASSAGE to study four additional human cell types. In all cases, the patterns observed—including the proportions of S, AS, and SAS genes—were similar to those in PBMCs (Table 1 and table S1). However, the identity of the S, AS, and SAS genes varied among the cell lines, which suggests that the expression of antisense tags may be regulated in a cell- or tissue-specific manner (fig. S5 and tables S2 and S3). These differences were not related to interexperimental variation, as repeat experiments performed with independently generated ASSAGE libraries from the same RNA sample were highly correlated (fig. S6 and table S2) and differential expression could be confirmed by strand-specific PCR from RNA (fig. S7). In every sample, there was a concentration of both sense and antisense tags within exons (relative to the whole genome or to intronic regions) and a preferential concentration in promoter and terminator regions ( $P < 0.01$ ; figs. S2 and S8).

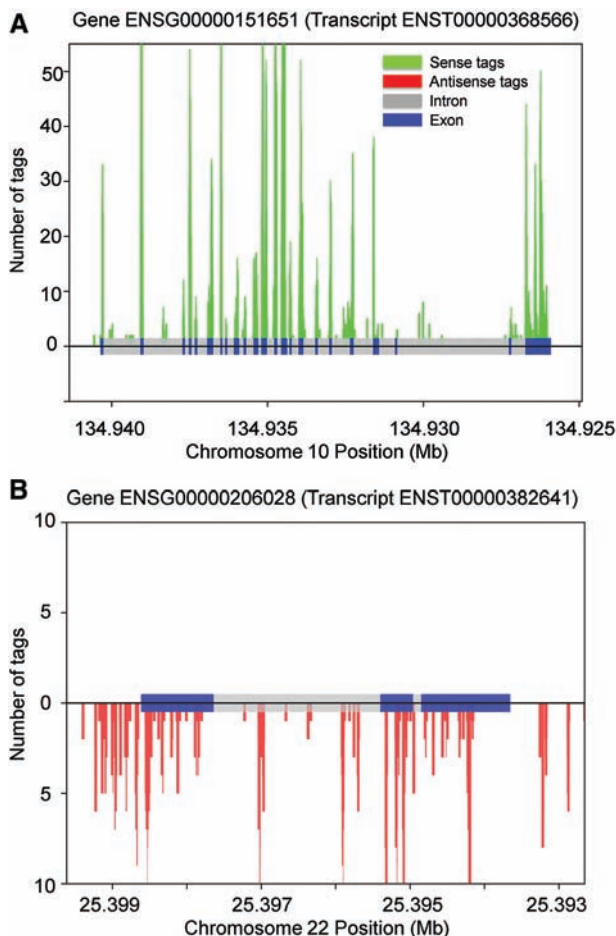
To determine whether splicing of antisense transcripts occurred, we constructed new libraries from Jurkat and MRC5 cells and determined the sequences of both ends of each cDNA fragment (“paired-end sequencing”). As expected, transcripts levels assessed with this paired-end ASSAGE and the original ASSAGE were highly correlated (fig. S9). The size-selected transcript fragments used to construct these libraries were, on average, ~175 base pairs in length. A cDNA fragment whose ends were located at genomic positions more than 3 times this distance (>600 base pairs apart) would be expected to represent spliced transcripts. By this criterion, more than 20% of sense-strand cDNA fragments were spliced (fig. S10). In contrast, only ~1% of antisense fragments exhibited this spliced pattern. Sequencing of five putative spliced antisense transcripts confirmed the splicing, and comparison with genomic DNA revealed the splice site consensus sequences at the expected locations (figs. S11 and S12).

Our results raise many questions about the genesis and metabolism of antisense transcripts. It has been hypothesized that antisense transcripts are widely and promiscuously expressed, perhaps because of weak promoters distributed throughout the genome [reviewed in (25, 26)]. Our data argue against this hypothesis in human cells: Promiscuous expression would lead to a uniform distribution of antisense tags across the genome, whereas the observed distribution

**Fig. 1.** ASSAGE tag densities in PBMCs. The densities of distinct sense and antisense tags in the indicated regions were normalized to the overall genome tag density. The promoter and terminator regions were defined as the 1 kb of sequence that was upstream or downstream, respectively, of the transcript start and end sites.



**Fig. 2.** Tag distribution in the indicated S (A) and AS (B) genes in PBMCs.



was nonrandom, localized to genes and within particular regions of genes, much like sense transcripts (Fig. 1 and figs. S2 and S8). This distribution is consistent with a model wherein many antisense transcripts initiate and terminate near the terminators and promoters, respectively, of the sense transcripts. Some of the apparent antisense transcripts from a gene on the plus strand could actually be sense transcripts originating from unterminated transcription of a downstream gene on the minus strand (or vice versa). However, this idea is not generally supported because there was a poor correlation between antisense tag density within a gene and the density of sense tags from the closest downstream gene (fig. S13). One explanation for the higher density of antisense tags in transcribed regions is that transcription of the sense transcripts from correct initiation sites would reduce nucleosome density throughout the entire transcribed region, thereby increasing DNA accessibility and hence the likelihood of nonspecific transcription (26). This is unlikely, given that genes with high sense tag densities did not generally have high antisense densities. There is substantial evidence that sense transcripts can be negatively regulated by antisense transcripts (3–7). Such regulation can occur either by transcriptional interference or through posttranscriptional mechanisms involving splicing or RNA-induced silenc-

ing complexes (RISCs). Our data support the possibility that antisense-mediated regulation affects a large number of genes.

#### References and Notes

1. P. O. Brown, D. Botstein, *Nat. Genet.* **21**, 33 (1999).
2. V. E. Velculescu *et al.*, *Cell* **88**, 243 (1997).
3. M. Lapidot, Y. Pilpel, *EMBO Rep.* **7**, 1216 (2006).
4. A. Mazo, J. W. Hodgson, S. Petruk, Y. Sedkov, H. W. Brock, *J. Cell Sci.* **120**, 2755 (2007).
5. J. A. Timmons, L. Good, *Biochem. Soc. Trans.* **34**, 1148 (2006).
6. P. Kapranov, A. T. Willingham, T. R. Gingeras, *Nat. Rev. Genet.* **8**, 413 (2007).
7. O. Yazgan, J. E. Krebs, *Biochem. Cell Biol.* **85**, 484 (2007).
8. J. Chen *et al.*, *Nucleic Acids Res.* **32**, 4812 (2004).
9. M. E. Fahey, T. F. Moore, D. G. Higgins, *Comp. Funct. Genomics* **3**, 244 (2002).
10. B. Lehner, G. Williams, R. D. Campbell, C. M. Sanderson, *Trends Genet.* **18**, 63 (2002).
11. J. Shendure, G. M. Church, *Genome Biol.* **3**, RESEARCH0044 (2002).
12. R. Yelin *et al.*, *Nat. Biotechnol.* **21**, 379 (2003).
13. H. Kiyosawa, I. Yamanaka, N. Osato, S. Kondo, Y. Hayashizaki, *Genome Res.* **13**, 1324 (2003).
14. D. J. Lipman, *Nucleic Acids Res.* **25**, 3580 (1997).
15. G. G. Carmichael, *Nat. Biotechnol.* **21**, 371 (2003).
16. RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) and FANTOM Consortium, *Science* **309**, 1564 (2005).
17. Y. Okazaki *et al.*, *Nature* **420**, 563 (2002).
18. D. Kampa *et al.*, *Genome Res.* **14**, 331 (2004).
19. See supporting material on Science Online.
20. A. J. Simpson, S. J. de Souza, A. A. Camargo, R. R. Brentani, *Comp. Funct. Genomics* **2**, 169 (2001).

21. B. A. Peters *et al.*, *Genome Res.* **17**, 287 (2007).
22. M. Sultan *et al.*, *Science* **321**, 956 (2008); published online 3 July 2008 (10.1126/science.1160342).
23. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, *Nat. Methods* **5**, 621 (2008).
24. J. Q. Wu *et al.*, *Genome Biol.* **9**, R3 (2008).
25. J. M. Johnson, S. Edwards, D. Shoemaker, E. E. Schadt, *Trends Genet.* **21**, 93 (2005).
26. K. Struhl, *Nat. Struct. Mol. Biol.* **14**, 103 (2007).
27. We thank W. Yu for assistance with microarrays. Supported by the Virginia and D. K. Ludwig Fund for Cancer Research and NIH grants CA57345, CA43460, CA62924, and CA121113. Under a licensing agreement between Johns Hopkins University and Genzyme, technologies related to SAGE were licensed to Genzyme for commercial purposes, and B.V., V.E.V., and K.W.K. are entitled to a share of the royalties received by the university from the sales of the licensed technologies. The university and researchers (B.V. and K.W.K.) own Genzyme stock, which is subject to certain restrictions under university policy. The terms of these arrangements are being managed by the university in accordance with its conflict-of-interest policies. There are existing patents for SAGE that have been licensed as disclosed above, and similar patents are likely to be filed for ASSAGE.

#### Supporting Online Material

www.sciencemag.org/cgi/content/full/1163853/DC1  
Materials and Methods

Figs. S1 to S13

Tables S1 to S3

References

26 July 2008; accepted 10 November 2008

Published online 4 December 2008;

10.1126/science.1163853

Include this information when citing this paper.

## Label-Free Biomedical Imaging with High Sensitivity by Stimulated Raman Scattering Microscopy

Christian W. Freudiger,<sup>1,2\*</sup> Wei Min,<sup>1\*</sup> Brian G. Saar,<sup>1</sup> Sijia Lu,<sup>1</sup> Gary R. Holtom,<sup>1</sup> Chengwei He,<sup>3</sup> Jason C. Tsai,<sup>4</sup> Jing X. Kang,<sup>3</sup> X. Sunney Xie<sup>1†</sup>

Label-free chemical contrast is highly desirable in biomedical imaging. Spontaneous Raman microscopy provides specific vibrational signatures of chemical bonds, but is often hindered by low sensitivity. Here we report a three-dimensional multiphoton vibrational imaging technique based on stimulated Raman scattering (SRS). The sensitivity of SRS imaging is significantly greater than that of spontaneous Raman microscopy, which is achieved by implementing high-frequency (megahertz) phase-sensitive detection. SRS microscopy has a major advantage over previous coherent Raman techniques in that it offers background-free and readily interpretable chemical contrast. We show a variety of biomedical applications, such as differentiating distributions of omega-3 fatty acids and saturated lipids in living cells, imaging of brain and skin tissues based on intrinsic lipid contrast, and monitoring drug delivery through the epidermis.

Vibrational microscopies based on infrared absorption and Raman scattering (1, 2) have been used as label-free contrast mechanisms due to characteristic frequencies of various chemical bonds. However, infrared microscopy has limited spatial resolution because of long infrared wavelengths. Spontaneous Raman scattering microscopy, while having higher spatial resolution due to shorter excitation wavelengths, is insensitive and thus often has limited imaging speed. Coherent anti-Stokes Raman scattering (CARS) microscopy offers higher sensitivity than spontaneous Raman microscopy (3, 4). However, a CARS spectrum is different from its corresponding spontaneous Raman spectrum due to a nonresonant background, which complicates spectral assignment, causes difficulties in image interpretation, and limits detection sensitivity.

Here we explore stimulated Raman scattering (SRS) as an imaging contrast mechanism. SRS is analogous (5, 6) to the well-known phenomenon of stimulated emission (7) and was first observed

in 1962 (8). Since then it has been used in many spectroscopic studies (9–12). In spontaneous Raman scattering, one laser beam at a frequency  $\omega_p$  illuminates the sample and the signal is generated at the Stokes and anti-Stokes frequencies  $\omega_S$  and  $\omega_{AS}$ , respectively, due to inelastic scattering. In SRS, however, two laser beams at  $\omega_p$  and  $\omega_S$  coincide on the sample (Fig. 1A). When the difference frequency,  $\Delta\omega = \omega_p - \omega_S$ , also called the Raman shift, matches a particular molecular vibrational frequency  $\Omega$ , amplification of the Raman signal is achieved by virtue of stimulated excitation. Consequently, the intensity of the Stokes beam,  $I_S$ , experiences a gain,  $\Delta I_S$  (stimulated Raman gain, SRG), and the intensity of the pump beam,  $I_p$ , experiences a loss,  $\Delta I_p$  (stimulated Raman loss, SRL), as shown in Fig. 1B. In contrast, when  $\Delta\omega$  does not match any vibrational resonance, SRL and SRG cannot occur. Therefore, unlike CARS, SRL and SRG do not exhibit a nonresonant background (11).

The intensity of SRG or SRL is described by  $\Delta I_S \propto N \times \sigma_{\text{Raman}} \times I_p \times I_S$  and  $\Delta I_p \propto -N \times$

<sup>1</sup>Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA. <sup>2</sup>Department of Physics, Harvard University, Cambridge, MA 02138, USA.

<sup>3</sup>Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA.

<sup>4</sup>Pfizer Global Medical, 685 3rd Avenue, MS 1325, New York, NY 10017, USA.

\*These authors contributed equally to this work.

†To whom correspondence should be addressed. E-mail: xie@chemistry.harvard.edu