RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription

Philipp Kapranov, *et al.*

Science **316**, 1484 (2007);
DOI: 10.1126/science.1138341

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

cate that these molecules also function in early growth responses to sarcomere dysfunction.

Myocardial fibrosis is characteristic of advanced HCM pathology and contributes to impaired cardiac relaxation, heart failure, arrhythmias, and sudden death (*20*, *21*). The increased expression of *Tgfb1* (transforming growth factor–β1), *Ctgf* (connective tissue growth factor) and *Postn* (periostin), potent regulators of fibrosis and collagen deposition (*22*–*24*), in prehypertrophic ventricles indicates early activation of this pathway, which raises the possibility that fibrosis is not an advanced secondary phenomenon, but a primary contributor to myocardial dysfunction.

Impaired relaxation is the fundamental physiologic abnormality in HCM (*25*). Cardiac relaxation and contraction reflects Ca^{++} cycling between the sarcoplasmic reticulum and the sarcomere in cardiomyocytes. Ca^{++} uptake into the sarcoplasmic reticulum occurs via sarcoplasmic reticulum Ca^{++} transport adenosine triphosphatase (ATPase) (SERCA2a/Atp2a2), which is regulated by phospholamban (Pln) and sarcolipin (Sln) (*26*). Transcripts encoding each of these proteins were significantly decreased in prehypertrophic hearts (fig. S11), which may directly account for the early impairment in cardiac relaxation previously observed in this model (*27*). Down-regulation of *Abcc9* [adenosine triphosphate (ATP)–binding cassette subfamily C member 9), which encodes SUR2, suggested another mechanism for Ca^{++} imbalance in prehypertrophic hearts. SUR2 is the ATPase-regulatory subunit of the inwardly rectifying cardiac K_{ATP} channel, which balances Ca^{++} homeostasis with energetic demands (*28*); *Abcc9*-null mice develop arrhythmias and myocardial calcium overload (*29*).

Notably, PMAGE also revealed significant ($P < 0.01$) differences in the expression of genes encoding 29 transcription factors between wild-type and prehypertrophic $\alpha MHC^{403/+}$ hearts (table S3). The biologic processes evoked by these molecules are likely to be considerable. By interrogating the temporal and spatial expression of these transcription factors, we can potentially dissect the networks activated in this cardiomyopathy, which, in turn, should help identify new molecular targets for therapeutic intervention.

In summary, PMAGE profiling provided reproducible, large-scale transcript identification, with sequence accuracy comparable to SAGE, and greater sensitivity for quantification of rare transcripts. We estimate that sampling ~2 million tags provides comprehensive assessment of most mRNAs (fig. S9); nevertheless, the current PMAGE platform has the capacity to read more than 4 million tags per experiment. Thus, PMAGE can be used for very deep sampling of one library or analyses of multiple libraries simultaneously by adapting polony beads that contain unique sequence identifiers. PMAGE offers several advantages over other currently available transcription profiling methods at a potentially lower cost (fig. S12). We anticipate that PMAGE studies will help further define mRNA regulatory networks that orchestrate critical cellular processes in healthy and diseased tissues.

References and Notes

1. G. A. Churchill, *Nat. Genet.* **32** (suppl.), 490 (2002).
2. V. E. Velculescu *et al.*, *Cell* **88**, 243 (1997).
3. M. J. Holland, *J. Biol. Chem.* **277**, 14363 (2002).
4. S. Draghici, P. Khatri, A. C. Eklund, Z. Szallasi, *Trends Genet.* **22**, 101 (2006).
5. S. Audic, J. M. Claverie, *Genome Res.* **7**, 986 (1997).
6. Materials and methods are available as supporting material on *Science* Online.
7. D. Dressman, H. Yan, G. Traverso, K. W. Kinzler, B. Vogelstein, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 8817 (2003).
8. J. Shendure *et al.*, *Science* **309**, 1728 (2005).
9. A. A. Geisterfer-Lowrance *et al.*, *Science* **272**, 731 (1996).
10. S. Blackshaw, J. B. Kim, B. St. Croix, K. Polyak, in *Current Protocols in Molecular Biology*, F. M. Asubel, Ed. (Greene Publishing Associates and Wiley-Interscience, New York, 2002), pp. v (loose leaf).
11. V. E. Velculescu *et al.*, *Nat. Genet.* **23**, 387 (1999).
12. S. Dinel *et al.*, *Nucleic Acids Res.* **33**, e26 (2005).
13. N. D. Hastie, J. O. Bishop, *Cell* **9**, 761 (1976).
14. J. G. Seidman, C. Seidman, *Cell* **104**, 557 (2001).
15. F. Chen *et al.*, *Cell* **110**, 713 (2002).
16. C. H. Shin *et al.*, *Cell* **110**, 725 (2002).
17. D. Srivastava, P. Cserjesi, E. N. Olson, *Science* **270**, 1995 (1995).
18. T. Maeda, D. L. Chapman, A. F. Stewart, *J. Biol. Chem.* **277**, 48889 (2002).
19. Y. Albert *et al.*, *J. Cell Biol.* **169**, 257 (2005).
20. A. M. Varnava, P. M. Elliott, N. Mahon, M. J. Davies, W. J. McKenna, *Am. J. Cardiol.* **88**, 275 (2001).
21. K. M. Harris *et al.*, *Circulation* **114**, 216 (2006).
22. P. J. Lijnen, V. V. Petrov, R. H. Fagard, *Mol. Genet. Metab.* **71**, 418 (2000).
23. M. S. Ahmed *et al.*, *J. Mol. Cell. Cardiol.* **36**, 393 (2004).
24. R. A. Norris *et al.*, *J. Cell. Biochem.* **101**, 695 (2007).
25. B. J. Maron, *JAMA* **287**, 1308 (2002).
26. M. Asahi *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 9199 (2004).
27. D. Georgakopoulos *et al.*, *Nat. Med.* **5**, 327 (1999).
28. M. Bienengraeber *et al.*, *Nat. Genet.* **36**, 382 (2004).
29. L. V. Zingman *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 13278 (2002).
30. R. T. Lee *et al.*, *J. Clin. Invest.* **81**, 431 (1988).

31. We thank J. Shendure, S. Barr, S. DePalma, M. Maida, P. Teekakirikul, S. Blackshaw, Z. Arany, J. Loscalzo, and B. Vogelstein for advice and technical assistance. Funded by grants from NIH and Howard Hughes Medical Institute.

Supporting Online Material

www.sciencemag.org/cgi/content/full/316/5830/1481/DC1
Materials and Methods
Figs. S1 to S12
Tables S1 to S3
References

8 November 2006; accepted 3 May 2007
10.1126/science.1137325

RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription

Philipp Kapranov,[1] Jill Cheng,[1] Sujit Dike,[1] David A. Nix,[1] Radharani Duttagupta,[1] Aarron T. Willingham,[1] Peter F. Stadler,[2] Jana Hertel,[2] Jörg Hackermüller,[3] Ivo L. Hofacker,[4] Ian Bell,[1] Evelyn Cheung,[1] Jorg Drenkow,[1] Erica Dumais,[1] Sandeep Patel,[1] Gregg Helt,[1] Madhavan Ganesh,[1] Srinka Ghosh,[1] Antonio Piccolboni,[1] Victor Sementchenko,[1] Hari Tammana,[1] Thomas R. Gingeras[1]*

Significant fractions of eukaryotic genomes give rise to RNA, much of which is unannotated and has reduced protein-coding potential. The genomic origins and the associations of human nuclear and cytosolic polyadenylated RNAs longer than 200 nucleotides (nt) and whole-cell RNAs less than 200 nt were investigated in this genome-wide study. Subcellular addresses for nucleotides present in detected RNAs were assigned, and their potential processing into short RNAs was investigated. Taken together, these observations suggest a novel role for some unannotated RNAs as primary transcripts for the production of short RNAs. Three potentially functional classes of RNAs have been identified, two of which are syntenically conserved and correlate with the expression state of protein-coding genes. These data support a highly interleaved organization of the human transcriptome.

A large fraction of the noncoding part of a eukaryotic genome is used to make RNA that is sufficiently stable in a cell to be detected by different technological approaches (*1*–*4*). The biological significance of this pervasive transcription is unclear and controversial. One possibility is that only very short regions of such unannotated RNA are biologically relevant (*5*). In-depth characterization of RNAs as to their subcellular compartmentalization, size, modifications, and genomic origins can potentially provide clues to their functions. This study reports two general observations derived from the maps of nuclear and cytosolic polyadenylated [poly(A)$^+$] RNAs longer than 200 nucleotides (nt) (long RNAs, lRNAs) and whole-cell RNAs less than 200 nt (short RNAs, sRNAs) over the entire nonrepetitive portion of the human ge-

[1]Affymetrix Laboratory, Affymetrix, Inc., 3420 Central Expressway, Santa Clara, CA, 95051, USA. [2]University of Leipzig, Department of Computer Science, Leipzig, Germany. [3]Fraunhofer Institute for Cell Therapy and Immunology, Leipzig, Germany. [4]Institute for Theoretical Chemistry, University of Vienna, Austria.

*To whom correspondence should be addressed. E-mail: tom_gingeras@affymetrix.com

nome. First, the potential biological function of an appreciable portion of long unannotated transcripts is to serve as precursors for sRNAs. Second, these maps reveal three classes of RNAs that have specific genomic localization at gene boundaries. Biological relevance of these classes of RNAs is supported by strong correlation with the expression state of genes they associate with, as well as their syntenic conservation between human and mouse.

The complexity of steady-state RNA populations was profiled by using tiling arrays at 5-nt resolution to detect transcribed regions in the human genome (*6*, *7*). Overall, we found the extent and general properties of the lRNA portion of the human transcriptome to be similar to our earlier study (*7*). Patterns of annotated and unannotated transcription were similar among cell lines and within subcellular compartments (fig. S1, A to J, and table S2, A and B). About 64% of detected poly(A)$^+$ transcription (nucleus and cytosol) did not align with annotations (fig. S1N) (*7*). Of the 265,237 annotated exons, 80% were expressed in at least one cell line (fig. S2).

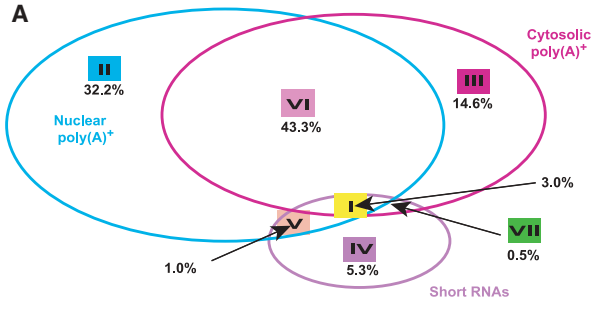
A total of 1.1% of the interrogated genome is covered by transcribed fragments (transfrags) representing sRNAs (summarized in table S1 and S2C, figs. S3 and S4). sRNA transfrags have a nonrandom association with genomic features, including EvoFold structure predictions (figs. S5 and S7 and table S1B). In addition, a tendency for some to map antisense to splice junctions was also found. sRNAs were found in intronic, intergenic, and annotated regions (fig. S4). Unannotated sRNAs were verified by Northern blots (table S3 and fig. S6) and real-time reverse transcription polymerase chain reaction (table S3), with an overall verification rate of ~70% (*7*).

Maps of sRNAs and lRNAs from different subcellular compartments can be further used to provide a virtual "genealogy" describing origins of particular classes of RNAs (Fig. 1 and table S4). A total of 12.7% of all interrogated nucleotides can be detected as composing lRNAs or sRNAs in HeLa or HepG2 cell lines. One-third of this total (33.2%) is exclusively observed in the nucleus as lRNAs and overlapping sRNAs (Fig. 1A). Another 15.1% is exclusively found as cytosolic lRNAs and overlapping sRNAs. A total of 46.3% of the sequences detected are found in both the nucleus and cytosol (Fig. 1A). Finally, 5.3% of the nucleotides were detected in sRNA transfrags exclusively.

The origin of 79.6% of all transcribed bases can be mapped back to the nucleus as lRNAs with 15.1% found only in cytosol (Fig. 1B). About 41.8% of the sequences seem to remain exclusively in the nucleus; the remainder is transported into the cytosol. Furthermore, 3.1% of the exclusively nuclear lRNA sequences and 6.6% of the nuclear sequences transported into the cytosol overlap sRNA sequences, which suggests that ~40% of the latter may be processed from long nuclear transcripts (Fig. 1B and table S4).

About one-fifth of sRNAs transfrags, 20.9% (HepG2) and 18.4% (HeLa), were identified as evolutionarily conserved. The PhastCons scores (*7*) associated with sRNAs are significantly enriched in conserved sequences (P value 2.2×10^{-16}, Wilcoxon nonparametric test) over random (fig. S8, A and B), which points to the possible biological relevance of these transcripts (Fig. 2A). There is also a statistically significant concordance ($P < 0.01$, permutation test) observed between the locations of sRNA and nuclear lRNA transfrags. A total of 13% (HepG2) and 9% (HeLa) nuclear lRNA transfrags overlap sRNAs. Conversely, 44% and 31% of sRNA transfrags overlap with nuclear lRNA transfrags. Such an association is potentially confounded by the elevated G-C composition of these regions (*7*). To explore this further, we divided the transfrags obtained from nuclear lRNA into those that do and do not overlap sRNA transfrags (Fig. 2B). Mean PhastCons scores for the lRNA transfrags that do overlap with sRNAs are significantly higher than such of the transfrags that do not (Fig. 2C and table S5). For lRNA transfrags overlapping sRNAs, a total of 23.9% (HepG2) and 26.2% (HeLa) exhibit PhastCons scores equivalent or higher than the average score observed for annotations (Fig. 2D). The conservation of the nuclear lRNA transfrags often extends beyond a sRNA transfrag it overlaps (Fig. 2B), indicating that other sequences outside of the overlapping regions may be important, reminiscent of the extended conservation seen in the miRNA precursors (*8*).

Taken together, these data suggest a possible product-precursor relation between overlapping transfrags derived from lRNAs and sRNAs, underscored by the enrichment of evolutionarily conserved sequences in genomic regions found transcribed in both lRNAs and sRNAs. Conservatively, 3.1% of HepG2 and 2.4% of HeLa nuclear lRNA transfrags may be parts of precursors of sRNAs. The full extent of transcription, which may serve as precursors of sRNA, could, however, be much larger, because lRNA transfrags that directly overlap sRNAs are almost certainly connected to other transfrags in a precursor transcript. Thus, any given lRNA transfrag can be an order of magnitude smaller than the lRNA transcript it represents.



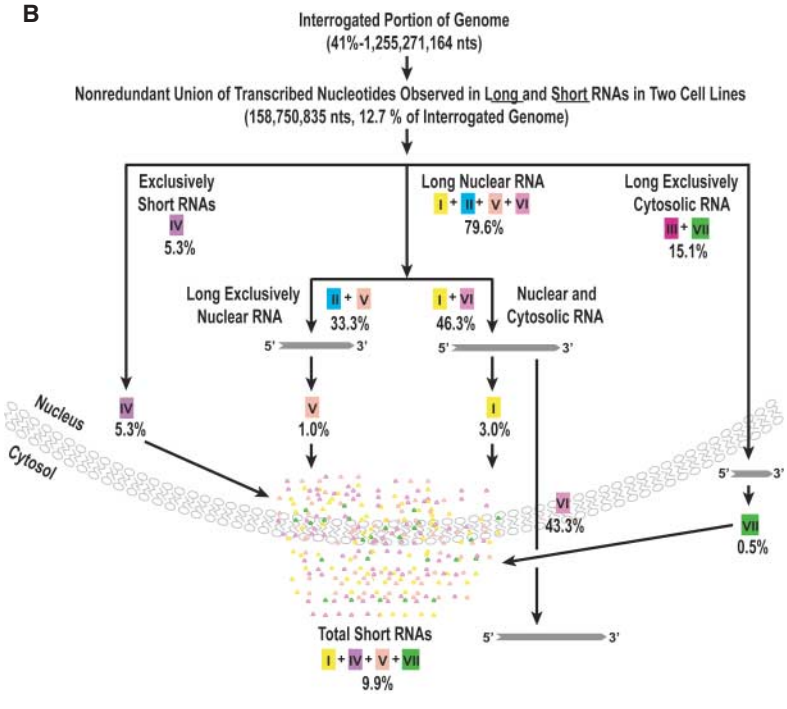


Fig. 1. Relations among the transcribed bases in the nonrepeat portions of the human genome. (**A**) Distribution of nucleotides in transfrags from long nuclear and cytosolic RNAs and sRNAs from HeLa and HepG2 (table S4). (**B**) Genealogy of nucleotides detected in lRNAs and sRNAs based on the association of the array-detected transcription shown in (A). Putative product-precursor associations between lRNAs and sRNAs are indicated by arrows.

sRNA mapping and sequence conservation analysis (fig. S8C) indicate that sRNAs cluster at the 5′ and 3′ of genes (Fig. 3). We denote these classes of sRNAs "promoter-associated sRNAs" (PASRs) and "termini-associated sRNAs" (TASRs). The occurrence of sRNAs centers around 5′ or 3′ termini and is statistically significant compared with G-C-matched random regions (Fig. 3, B and C, and table S1). Northern hybridization analysis revealed that PASRs and TASRs can vary in length (22 to 200 nt), with one prominent class of PASRs with lengths of ~ 26, 38, and 50 nt (Fig. 3, B and C, and figs. S9 and S12 and tables S6 and S7). PASRs were expressed at levels similar to those of the protein-coding genes they overlap (*7*).

Several characteristics of both PASRs and TASRs support the biological significance of these sRNAs. As explained below, gene expression correlates with the density of PASRs, and PASRs associate with other lRNAs at the 5′ boundaries of genes. Also, expressed PASRs are syntenic with mouse.

The correlation of gene expression with the density of PASRs (Fig. 3D) is similar to a trend seen for antisense TASRs (fig. S10, A and B). Overall, 44.6 and 43.8% of genes found to be expressed in cytosol or nucleus have PASR association. Another 11.8 and 18.1% of genes had signal only in the first exon in cytosolic or nuclear RNAs, respectively. Almost half of those are observed to have PASRs (fig. S10, C and D). Conversely, for ~80% of silent genes (<10% of exons detected), no PASRs were observed.

A third class of RNAs is the long transcripts that overlap 5′ boundaries of protein-coding genes but do not include most of the other exons. This is exemplified by genes that show signal only in their first exons (Fig. 3A). To characterize these promoter-associated lRNAs (PALRs), we performed 5′ and 3′ RACE analysis (rapid amplification of cDNA ends) followed by hybridization to tiling arrays (fig. S11). These experiments revealed that transcripts overlapping the promoter and the first exon and intron regions, ranging in length from hundreds of base pairs to more than 1 kb, are made and map to the same genomic regions as PASRs.

We constructed sRNA maps in two syntenic regions of the human and mouse genomes (*IL4R* cytokine cluster and four Hox loci) using mouse STO and R1 and human HepG2 and HeLa cell lines (*7*). Both species-specific and conserved PASRs and TASRs were found, with ~39% of PASR sequences and 35% of TASR sequences mapping into syntenically conserved regions (Fig. 4A). Genomic regions shared by the PASR (HMSY19) at the 5′ boundaries of the Hox D9 and the TASR (HMSY5) in the 3′ termini of HoxD10 genes of both species are illustrated in Fig 4. The sizes of PASRs and TASRs are similar for mouse and human cell lines (fig. S13 and table S8).

We have found that ~10% of detected transcription is present in sRNA sequences (ranging from 22 to 200 nt in length). The distribution of these sRNAs is not uniform across the genome, because sRNAs are more frequent among genes than in intergenic regions. Further-

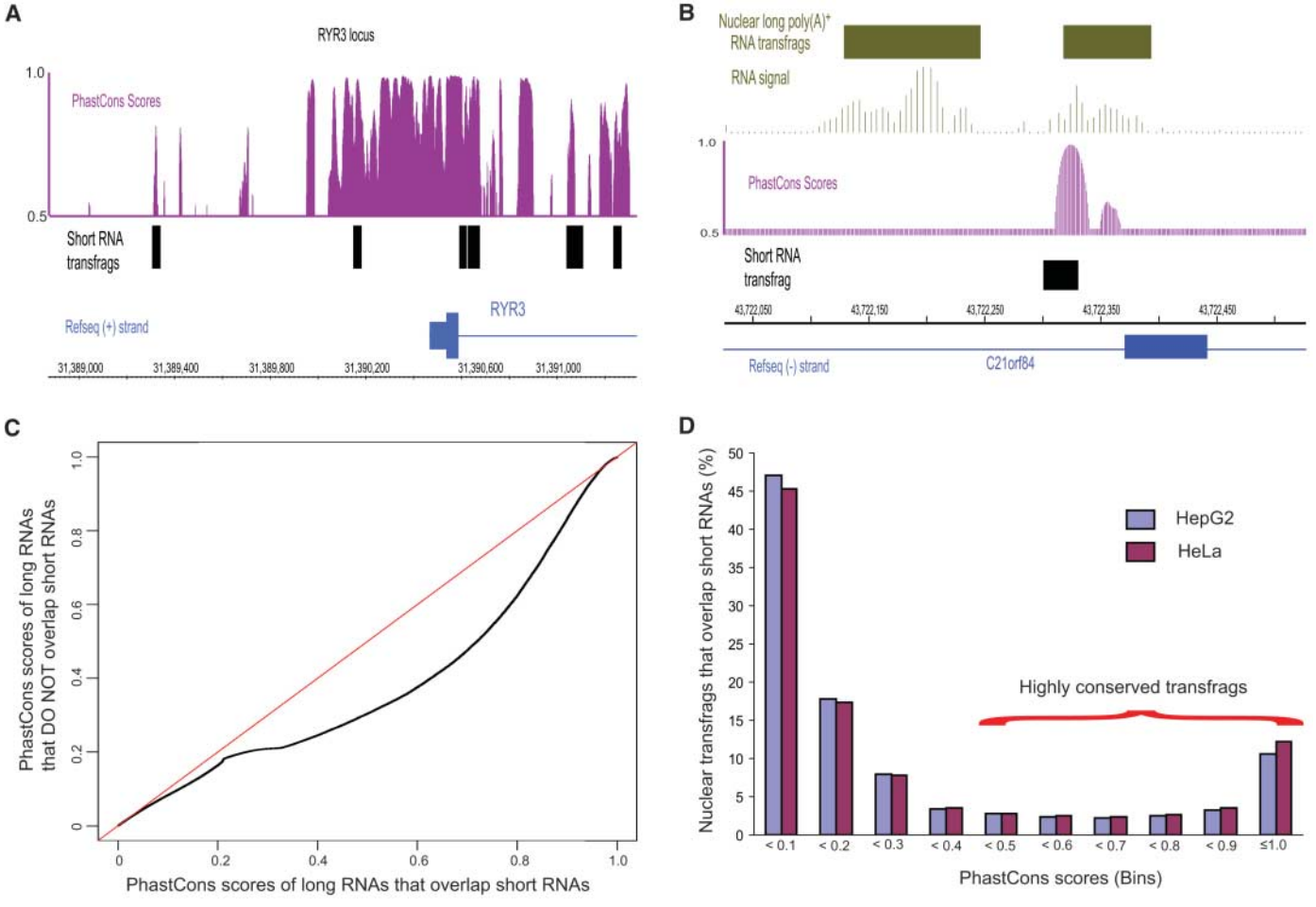


Fig. 2. Sequence conservation analysis of short and long nuclear RNA (*7*). (**A**) Conserved sRNAs surrounding the first exon of *RYR3* gene. (**B**) Long nuclear transfrag that overlaps sRNA is more conserved than adjacent lRNA transfrag that does not. (**C**) Quantile-quantile plot of PhastCons scores of long nuclear transfrags that do (*x* axis) and do not (*y* axis) overlap sRNAs. For any given point on the curve, an equal proportion of each "quantile distribution" occurs at this juncture. (**D**) Distribution of PhastCons scores of long nuclear transfrags that overlap sRNAs, binned on the basis of PhastCons scores (*x* axis), versus percentage of transfrags in each bin (*y* axis). Highly conserved transfrags (scores > 0.4) are indicated.

more, sRNA transfrags overlap a collection of lRNA transfrags that are significantly enriched in conserved sequences. Taken together, these data suggest that these lRNA transfrags potentially represent parts of nuclear primary transcripts that encode conserved functional sRNAs.

Several other observations are also derived from these mapping data. First, an appreciable fraction of protein-coding genes have expression only in the first exon and intron. This suggests that transcription may have two different states that are characterized by the lengths of transcripts made from the transcriptional start site of gene locus. Second, the fate of the transcripts derived from a particular detected transcribed region could be predicted on the basis of their retention in the nucleus, transport into cytosol, or processing into sRNAs. Overall, these RNA maps provide a virtual genealogy of RNAs (Fig. 1B). Third, PASRs often align within the boundaries of some of the PALRs. The genomic loci and boundaries of PASRs appear to be well conserved in two human cell lines and, in some cases, between mouse and human cells; this may indicate that there could be common processing signals used to create them. Fourth, the ends of almost half of human protein-coding genes were found to be bracketed by PASRs and TASRs. Given that large regions (i.e., >1 kb) are contained in the sequences covered by the sRNAs, the functional roles of these sRNAs may involve broad domains consistent with involvement in chromatin alterations.

Other recent studies also report the presence of multiple transcripts at the 5′ boundaries of genes (*9*), including unstable lRNAs postulated to be involved in regulation of gene expression (*10*, *11*). Thus, these results suggest a model of genome organization where protein-coding genes are at the center of a complex network of overlapping sense and antisense lRNA transcription, with interleaved sRNAs often marking their boundaries and correlating with their expression state (fig. S14). Our studies also highlight a possible important biological function for a portion of unannotated nuclear transcription as possible precursors for sRNAs. Such interleaved transcription produces a variety of non–protein coding sRNA and lRNA species that offer cis- and trans-regulatory potential (*12*–*14*).

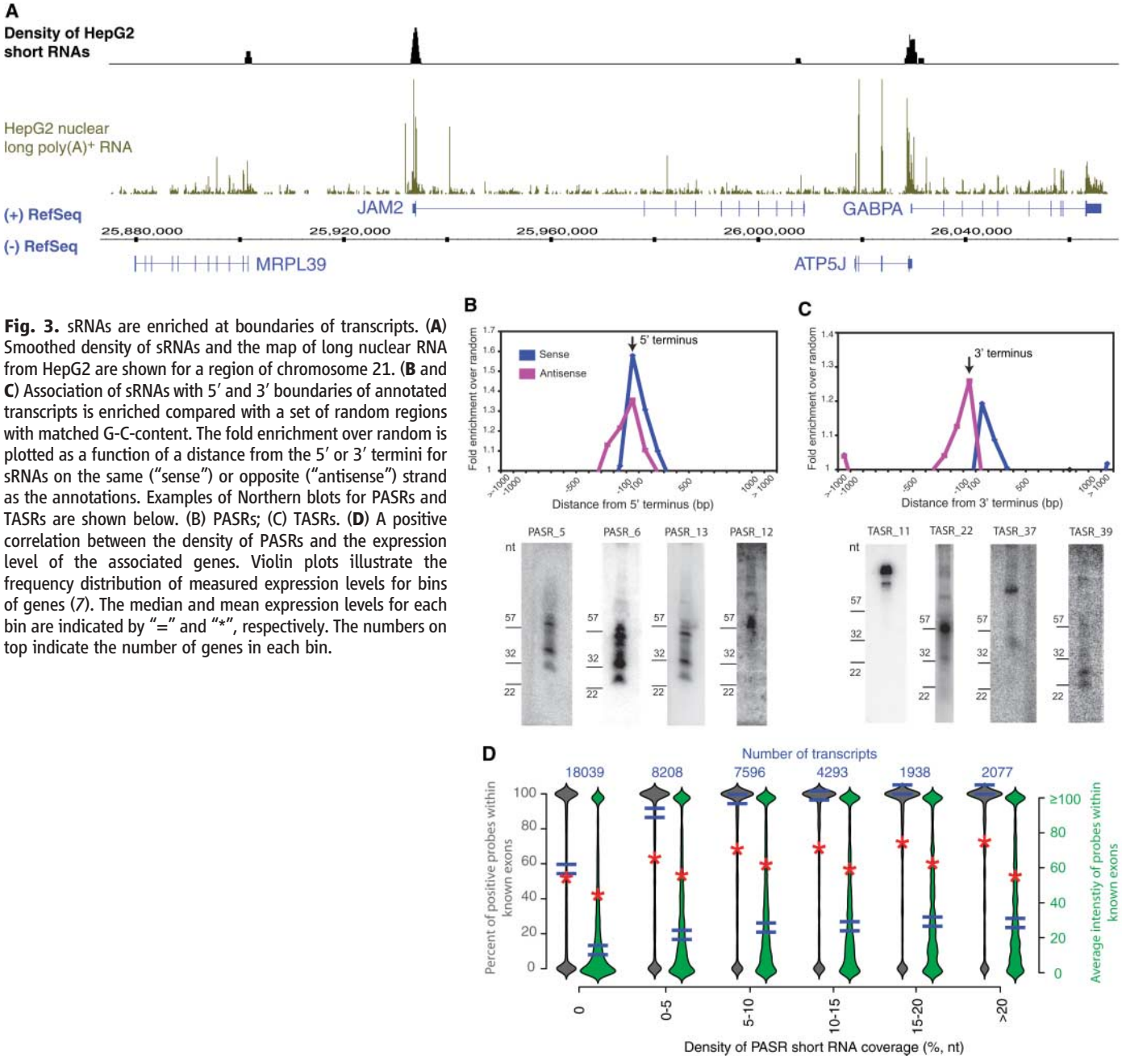


Fig. 3. sRNAs are enriched at boundaries of transcripts. (**A**) Smoothed density of sRNAs and the map of long nuclear RNA from HepG2 are shown for a region of chromosome 21. (**B** and **C**) Association of sRNAs with 5′ and 3′ boundaries of annotated transcripts is enriched compared with a set of random regions with matched G-C-content. The fold enrichment over random is plotted as a function of a distance from the 5′ or 3′ termini for sRNAs on the same ("sense") or opposite ("antisense") strand as the annotations. Examples of Northern blots for PASRs and TASRs are shown below. (B) PASRs; (C) TASRs. (**D**) A positive correlation between the density of PASRs and the expression level of the associated genes. Violin plots illustrate the frequency distribution of measured expression levels for bins of genes (*7*). The median and mean expression levels for each bin are indicated by "=" and "*", respectively. The numbers on top indicate the number of genes in each bin.

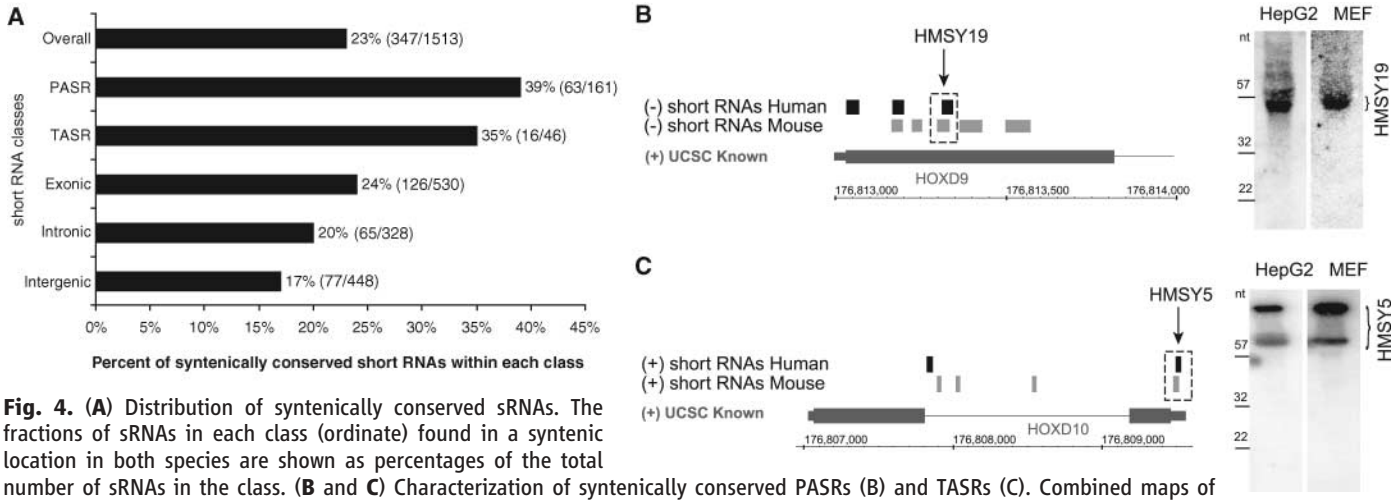


Fig. 4. (**A**) Distribution of syntenically conserved sRNAs. The fractions of sRNAs in each class (ordinate) found in a syntenic location in both species are shown as percentages of the total number of sRNAs in the class. (**B** and **C**) Characterization of syntenically conserved PASRs (B) and TASRs (C). Combined maps of syntenic sRNAs from HeLa and HepG2 human cell lines (black) and R1mES and MEF mouse cell lines (gray) are shown. Syntenic PASR HMSY19 and TASR HMSY5 are shown on either top (+) or bottom (–) strands. Northern blots show HMSY19 and HMSY5 in both species with comparable sizes.

References and Notes

1. J. Cheng *et al.*, *Science* **308**, 1149 (2005).
2. ENCODE-Project-Consortium, in preparation.
3. P. Kapranov *et al.*, *Science* **296**, 916 (2002).
4. A. T. Willingham, T. R. Gingeras, *Cell* **125**, 1215 (2006).
5. J. Ponjavic, C. P. Ponting, G. Lunter, *Genome Res.* **17**, 556 (2007).
6. Data available at http://transcriptome.affymetrix.com/hs_whole_genome; www.ncbi.nlm.nih.gov/geo/; http://genome.ucsc.edu/.
7. Materials and methods are available as supporting material on *Science* Online.
8. U. Ohler, S. Yekta, L. P. Lim, D. P. Bartel, C. B. Burge, *RNA* **10**, 1309 (2004).
9. P. Carninci *et al.*, *Nat. Genet.* **38**, 626 (2006).
10. C. A. Davis, M. Ares Jr., *Proc. Natl. Acad. Sci. U.S.A.* **103**, 3262 (2006).
11. I. Martianov, A. Ramadass, A. Serra Barros, N. Chow, A. Akoulitchev, *Nature* **445**, 666 (2007).
12. R. J. Britten, E. H. Davidson, *Science* **165**, 349 (1969).
13. F. Jacob, J. Monod, *J. Mol. Biol.* **3**, 318 (1961).
14. J. S. Mattick, *Curr. Opin. Genet. Dev.* **4**, 823 (1994).

15. We thank M. Mittmann, D. Le, and E. Schell for design of tiling arrays; K. Kole, D. Barone, and C. Chen for their help with direct RNA labeling; G. Hannon, K. Fejes-Toth, D. Gerhard, and K. Nussbacher for technical discussion and assistance in manuscript preparation; Mt. Sinai Hospital and A. Nagy, R. Nagy, W. Abramow-Newerly, J. Rossant, and J. Roder for procurement of the R1mES cell line; and M. Brown and D. Menke at Stanford for help in preparation of mouse embryo fibroblasts. This project has been funded in part with funds from the National Cancer Institute, NIH, under contract no. N01-CO-12400 and from the National Human Genome Research Institute, NIH, under grant no. U01HG003147, and by Affymetrix, Inc. The data discussed in this publication have been deposited in National Center for Biotechnology Information's Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo/) and are accessible through GEO Series accession number GSE-7576.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1138341/DC1
Materials and Methods
Figs. S1 to S14
Tables S1 to S8
References

4 December 2006; accepted 24 April 2007
Published online 17 May 2007;
10.1126/science.1138341
Include this information when citing this paper.

A Common Allele on Chromosome 9 Associated with Coronary Heart Disease

Ruth McPherson,[1]*† Alexander Pertsemlidis,[2]* Nihan Kavaslar,[1] Alexandre Stewart,[1] Robert Roberts,[1] David R. Cox,[3] David A. Hinds,[3] Len A. Pennacchio,[4,5] Anne Tybjaerg-Hansen,[6] Aaron R. Folsom,[7] Eric Boerwinkle,[8] Helen H. Hobbs,[2,9] Jonathan C. Cohen[2,10]†

Coronary heart disease (CHD) is a major cause of death in Western countries. We used genome-wide association scanning to identify a 58-kilobase interval on chromosome 9p21 that was consistently associated with CHD in six independent samples (more than 23,000 participants) from four Caucasian populations. This interval, which is located near the *CDKN2A* and *CDKN2B* genes, contains no annotated genes and is not associated with established CHD risk factors such as plasma lipoproteins, hypertension, or diabetes. Homozygotes for the risk allele make up 20 to 25% of Caucasians and have a ~30 to 40% increased risk of CHD.

Coronary heart disease (CHD) is the single greatest cause of death worldwide (*1*, *2*). Although CHD is highly heritable, the DNA sequence variations that confer cardiovascular risk remain largely unknown. To identify sequence variants associated with CHD, we undertook a genome-wide association study using 100,000 single-nucleotide polymorphisms (SNPs). To minimize false positive associations without unduly sacrificing statistical power, we designed the study to comprise three sequential case-control comparisons performed at a nominal significance threshold of $P < 0.025$ (Fig. 1). For the initial genome-wide scan, cases and controls were Caucasian men and women from Ottawa, Canada who participated in the Ottawa Heart Study (OHS). Cases had severe, premature CHD with a documented onset before the age of 60 years and culminating in coronary artery revascularization (table S1). To limit confounding by factors that strongly predispose to premature CHD, we excluded individuals with diabetes or plasma cholesterol levels consistent with mono-

[1]Division of Cardiology, University of Ottawa Heart Institute, Ottawa K1Y4W7, Canada. [2]Donald W. Reynolds Cardiovascular Clinical Research Center and the Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA. [3]Perlegen Sciences, Mountain View, CA 94043, USA. [4]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [5]U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA. [6]Department of Clinical Biochemistry, Rigshospitalet, Copenhagen University Hospital, Copenhagen DK-2100, Denmark. [7]Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, MN 55454, USA. [8]Human Genetics Center and Institute for Molecular Medicine, University of Texas Health Science Center, Houston, TX 77030, USA. [9]Howard Hughes Medical Institute at the University of Texas Southwestern Medical Center, Dallas, TX 75390, USA. [10]Center for Human Nutrition at the University of Texas Southwestern Medical Center, Dallas, TX 75390, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: jonathan.cohen@utsouthwestern.edu (J.C.C.); rmcpherson@ottawaheart.ca (R.M.)