

GPa,  $K'$ ,  $\alpha$  K<sup>-1</sup>) for Fe are (7.00, 124.3, 6.90, 13.00 × 10<sup>-5</sup>), for FeO are (12.95, 122.0, 4.30, 3.50 × 10<sup>-5</sup>), and for FeS are (20.1449, 17.0, 8.00, 12.00 × 10<sup>-5</sup>). The volume change due to phase transition in FeS and its phase boundary are from (48). Volumetric mixing nonideality in Fe-FeS liquids are parameterized as ( $\bar{V}$ ) =  $V_0(n-1+X)/n$ , with  $n = 7$  and 20 for Fe and FeS, respectively; other species volumes mix ideally. The thermodynamic model was validated against the low-pressure Fe-O-S liquidus (8), Fe-Fe<sub>2</sub>O<sub>3</sub> phase diagram (9), and high-pressure Fe-FeS/Fe-Fe<sub>3</sub>S eutectic experiments (21, 46). Fitting Fe-FeS/Fe-Fe<sub>3</sub>S eutectic experiments requires volume-mixing nonideality and leads to significant differences in Fe-rich liquid densities if neglected. Compositional uncertainty ~ ±1 wt %.

26. C. E. Harvie, J. P. Greenberg, J. H. Weare, *Geochim. Cosmochim. Acta* **51**, 1045 (1987).
27. A. Dziewonski, D. L. Anderson, *Phys. Earth Planet. Int.* **25**, 297 (1981).
28. The calculated core liquid densities and wave speeds within ±2% of PREM that we consider to be feasible match the outer-core properties.
29. A. Zerr, A. Diegler, R. Boehler, *Science* **281**, 243 (1998).
30. J-Array Group, *J. Geomagn. Geoelec.* **45**, 1265 (1993).
31. J. E. Vidale, H. M. Benz, *Nature* **356**, 678 (1992).
32. S. Kaneshima, G. Helffrich, *J. Geophys. Res.* **103**, 4825 (1998).
33. D. J. Stevenson, *Geophys. J. R. Astron. Soc.* **88**, 311 (1987).
34. Lateral thermal variations in the core larger than 10<sup>-4</sup> K will drive convective motions. The thermal expansivity of the core is ≈10<sup>-4</sup> K<sup>-1</sup>, so lateral  $\delta\rho/\rho > 10^{-8}$  will drive core flows.  $\delta\rho/\rho$  for our calculated core liquids is ≈10<sup>-1</sup>, eliminating any topography on the boundary between the liquids.
35. A. Morelli, A. M. Dziewonski, *Geophys. J. Int.* **112**, 178 (1993).
36. SP6 was used to calculate synthetic seismograms because it predicts the observed P4KP-PcP travel times better than other whole-Earth models (PREM, *isap91*, or AK135). The reflectivity method was used to calculate seismograms with 3500 K properties  $V_p = 8.225$  km s<sup>-1</sup>,  $\rho = 9.293$  Mg m<sup>-3</sup>, and  $l = 3.3$  km and 4300 K properties  $V_p = 8.361$  km s<sup>-1</sup>,  $\rho = 8.970$  Mg m<sup>-3</sup>, and  $l = 3.4$  km, from  $\rho_{icb} = 0.82$  Mg m<sup>-3</sup> (15).
37. K. Holland, T. J. Ahrens, *Science* **275**, 1623 (1997).
38. S. Urakawa, M. Kato, M. Kumazawa, in *High-Pressure Research in Mineral Physics*, M. Manghnani, Y. Syono, Eds. (American Geophysical Union, Washington, DC, 1987), pp. 95–111.
39. T. Okuchi, *Science* **278**, 1781 (1997).
40. V. S. Solomatov, D. J. Stevenson, *J. Geophys. Res.* **98**, 5375 (1993).
41. The rate of erosion of a density contrast across two layers is  $d\Delta\rho/dt = -8\epsilon\alpha F/(dC_p)$ . For efficiency factor  $\epsilon = 6 \times 10^{-3}$ , thermal expansivity  $\alpha = 1.32 \times 10^{-5}$  K<sup>-1</sup>, core heat flux  $F = 75$  mW m<sup>-2</sup>, convective layer thickness  $d$  2260 km, and heat capacity  $C_p = 860$  J kg<sup>-1</sup> K<sup>-1</sup>, an 850 kg m<sup>-3</sup> layer contrast would be eroded in 1120 × 10<sup>9</sup> years. Parameter values are from (49). A rotating spherical geometry modifies the analysis, but appears to reduce the likelihood of entrainment (50).
42. D. C. Rubie, C. K. Gessmann, D. J. Frost, *Nature* **429**, 58 (2004).
43. F. D. Stacey, C. H. B. Stacey, *Phys. Earth Planet. Int.* **110**, 83 (1999).
44. V. Kress, *Contrib. Mineral. Petrol.* **127**, 127 (1997).
45. T. M. Usselman, *Am. J. Sci.* **275**, 278 (1975).
46. Y. Fei, C. M. Bertka, L. W. Finger, *Science* **275**, 1621 (1997).
47. J. Li, Y. Fei, H.-K. Mao, K. Hirose, S. R. Shieh, *Earth Planet. Sci. Lett.* **193**, 509 (2001).
48. Y. Fei, C. T. Prewitt, H.-K. Mao, C. M. Bertka, *Science* **268**, 1892 (1995).
49. S. Labrosse, J.-P. Poirier, J.-L. Le Mouél, *Phys. Earth Planet. Int.* **99**, 1 (1997).
50. J. R. Lister, B. A. Buffett, *Phys. Earth Planet. Int.* **105**, 5 (1998).
51. M. Schimmel, H. Paulsen, *Geophys. J. Int.* **130**, 497 (1997).
52. J. Park, V. Levin, *Bull. Seismol. Soc. Am.* **90**, 1507 (2000).
53. We used this cross-correlation method for the

deconvolution with a 2-Hz cutoff. To window the arrivals in the linear stacks, we used the phase function for the phase-weighted stack (51), forcing it to be smooth in the neighborhood of the arrival by fitting it to a Lorentzian and tapering.

54. R. Kind, *J. Geophys.* **42**, 191 (1976).
55. We thank the Japan Society for the Promotion of Science for support; V. Kress for clarifications about the liquid thermodynamic model; G. Houseman, B. Wood, and the referees for suggestions that substantially improved the manuscript; and J. Jacobs for inspiration to learn more about the core.

Seismograms were provided by Hinet (National Research Institute for Earth Science and Disaster Prevention, Tsukuba, Japan).

**Supporting Online Material**  
[www.sciencemag.org/cgi/content/full/306/5705/2239/DC1](http://www.sciencemag.org/cgi/content/full/306/5705/2239/DC1)  
 Fig. S1  
 Table S1  
 4 June 2004; accepted 18 November 2004  
 10.1126/science.1101109

# Global Identification of Human Transcribed Sequences with Genome Tiling Arrays

Paul Bertone,<sup>1\*</sup> Viktor Stolc,<sup>1,2\*</sup> Thomas E. Royce,<sup>3</sup> Joel S. Rozowsky,<sup>3</sup> Alexander E. Urban,<sup>1</sup> Xiaowei Zhu,<sup>1</sup> John L. Rinn,<sup>3</sup> Waraporn Tongprasit,<sup>4</sup> Manoj Samanta,<sup>2</sup> Sherman Weissman,<sup>5</sup> Mark Gerstein,<sup>3†</sup> Michael Snyder<sup>1,3†</sup>

Elucidating the transcribed regions of the genome constitutes a fundamental aspect of human biology, yet this remains an outstanding problem. To comprehensively identify coding sequences, we constructed a series of high-density oligonucleotide tiling arrays representing sense and antisense strands of the entire nonrepetitive sequence of the human genome. Transcribed sequences were located across the genome via hybridization to complementary DNA samples, reverse-transcribed from polyadenylated RNA obtained from human liver tissue. In addition to identifying many known and predicted genes, we found 10,595 transcribed sequences not detected by other methods. A large fraction of these are located in intergenic regions distal from previously annotated genes and exhibit significant homology to other mammalian proteins.

The prevailing gene structures encountered in many organisms consist primarily of coding sequences with few and short intervening regions, and thus their characterization is largely straightforward. In contrast, mammalian genes often contain many short exons interspersed with very large introns, making the identification of coding sequences difficult; a comprehensive and accurate map of human coding sequences therefore does not exist. Functional assays are expected to be essential for the identification of coding segments and verification of predicted genes.

In principle, genome tiling microarrays offer the opportunity to comprehensively

investigate the RNA coding regions of any species using an unbiased approach. Recently, various microarray technologies have been applied to assess genome-wide transcription in bacterial and plant genomes (1–3), as well as transcription over human chromosomes 21 and 22 (4, 5). Each of these methods identified many previously unannotated features, noting a high degree of novel transcription beyond that expected by existing gene annotation data. These studies clearly demonstrated the merit of the microarray approach to the problem of large-scale transcript mapping; however, until now the large size of mammalian genomes has precluded the construction of a genome-wide high-resolution tiling array.

Using maskless photolithographic DNA synthesis technology (6, 7), we constructed 134 high-density oligonucleotide microarrays to represent ~1.5 Gb of nonrepetitive genomic DNA from each strand of the human genome (8, 9). A total of 51,874,388 36-nucleotide (nt) probes, positioned every 46 nt on average, were selected to interrogate sense and antisense strands of the genome and synthesized at a feature density of ~390,000 probes per array [fig. S1 (10)]. To measure transcriptional activity, we hybridized the

<sup>1</sup>Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT 06520–8103, USA. <sup>2</sup>Center for Nanotechnology, NASA Ames Research Center, Moffett Field, CA 94035, USA. <sup>3</sup>Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520–8114, USA. <sup>4</sup>Eloret Corporation, Sunnyvale, CA 94087, USA. <sup>5</sup>Department of Genetics, Yale University School of Medicine, New Haven, CT 06520–8005, USA.

\*These authors contributed equally to this work.  
 †To whom correspondence should be addressed.  
 E-mail: michael.snyder@yale.edu (M.S.), mark.gerstein@yale.edu (M.G.)

arrays to fluorescence-labeled cDNA reverse-transcribed from triple-selected polyadenylated [poly(A)<sup>+</sup>] liver tissue RNA pooled from several individuals (10).

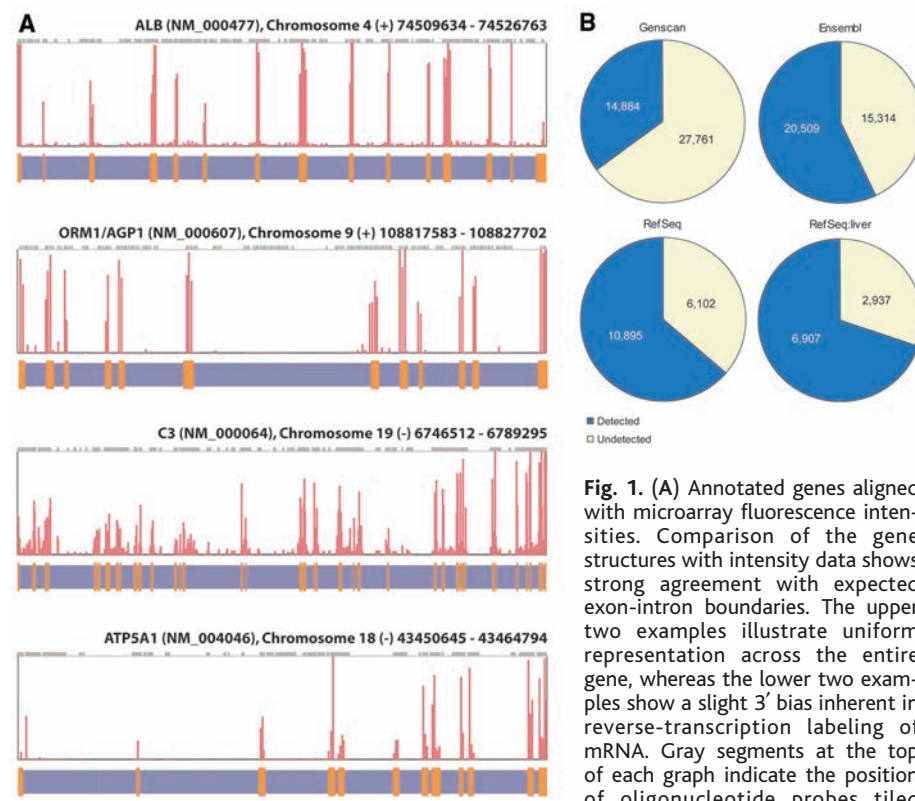
We first performed a pilot study to test the reproducibility of the platform. Multiple arrays were probed with cDNA samples derived from identical and independent labeling reactions, producing technical replicates having *r*<sup>2</sup> correlations between 0.90 and 0.95 (11), indicating that the experiments are highly reproducible. To further reduce the effect of potential variation across individual cDNA samples, we used pooled reverse transcription products of 20 separate labeling reactions to probe the genome tiling arrays.

To correlate fluorescence intensity values with meaningful chromosomal features, we aligned the oligonucleotide probe coordinates with current gene annotation data, using the RefSeq (12) and Ensembl (13, 14) databases. Alignment of the fluorescence intensities to the chromosomal coordinates of many known genes shows strong agreement between hybridization signals and annotated exons (Fig. 1A). To systematically determine the number of annotated genes detected with our approach, we devised a simple statistical method for scoring the observed transcriptional activity of annotated genes (15). This measurement essentially compares the fluorescence intensity of each probe within a gene against the median probe intensity across the entire microarray to determine whether they are significantly different. We scored 16,997 annotated genes from RefSeq, 35,823 genes from Ensembl, and 42,645 genes predicted by Genscan (16). Based on our criteria, transcription was detected from 64% (10,895), 57% (20,509), and 35% (14,884) of genes in each data set, respectively (Fig. 1B). These results agree with the expectation that fewer genes should be experimentally detected from annotation data sets that include putative genes predicted by homology or ab initio methods, as opposed to a curated collection of characterized genes. Nonetheless, our results provide the first genome-wide experimental confirmation that many of the predicted genes are transcribed, suggesting that they are functional. A subset of 9844 RefSeq genes with corresponding UniGene (17) annotations that indicate transcription in liver tissue was also examined; 70% (6907) of these were detected using our approach (Table 1).

In addition to detecting known and predicted genes, a primary goal of this study was to identify novel transcribed regions. Transcribed regions outside of previously annotated exons are expected to correspond primarily to (i) unannotated exons from alternatively spliced messages, (ii) under-

represented 3'- and 5'-untranslated regions, (iii) non-protein-coding RNA transcripts, and (iv) novel transcripts coding for functional proteins. We considered aggregate transcription units consisting of at least five consecutive probes exhibiting fluorescence intensities in the top 90th intensity percentile, and the genomic coordinates of which lay within a 250-nt window (Fig. 2A). These were compiled from throughout the genome and their locations compared relative to those of annotated gene components (Fig. 2B). A total of 13,889 transcription units,

ranging in size from 209 to 3438 nt, were identified in the genome by these criteria; ~400 are expected under the null hypothesis of zero transcription. One-third (4931) correspond to previously annotated exons; the remaining 8958 are new transcribed sequences that we refer to as transcriptionally active regions, or TARs (5). We located 1566 TARs within previously annotated introns on the same strand, raising the possibility that they correspond to overlooked exons. However, an equal number of TARs (1529) lie on the antisense strand of



**Fig. 1. (A)** Annotated genes aligned with microarray fluorescence intensities. Comparison of the gene structures with intensity data shows strong agreement with expected exon-intron boundaries. The upper two examples illustrate uniform representation across the entire gene, whereas the lower two examples show a slight 3' bias inherent in reverse-transcription labeling of mRNA. Gray segments at the top of each graph indicate the position of oligonucleotide probes tiled

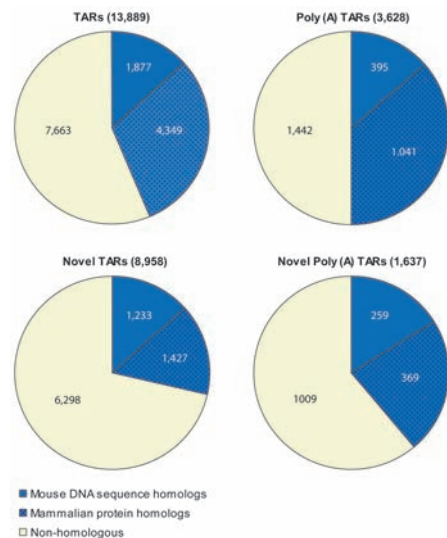
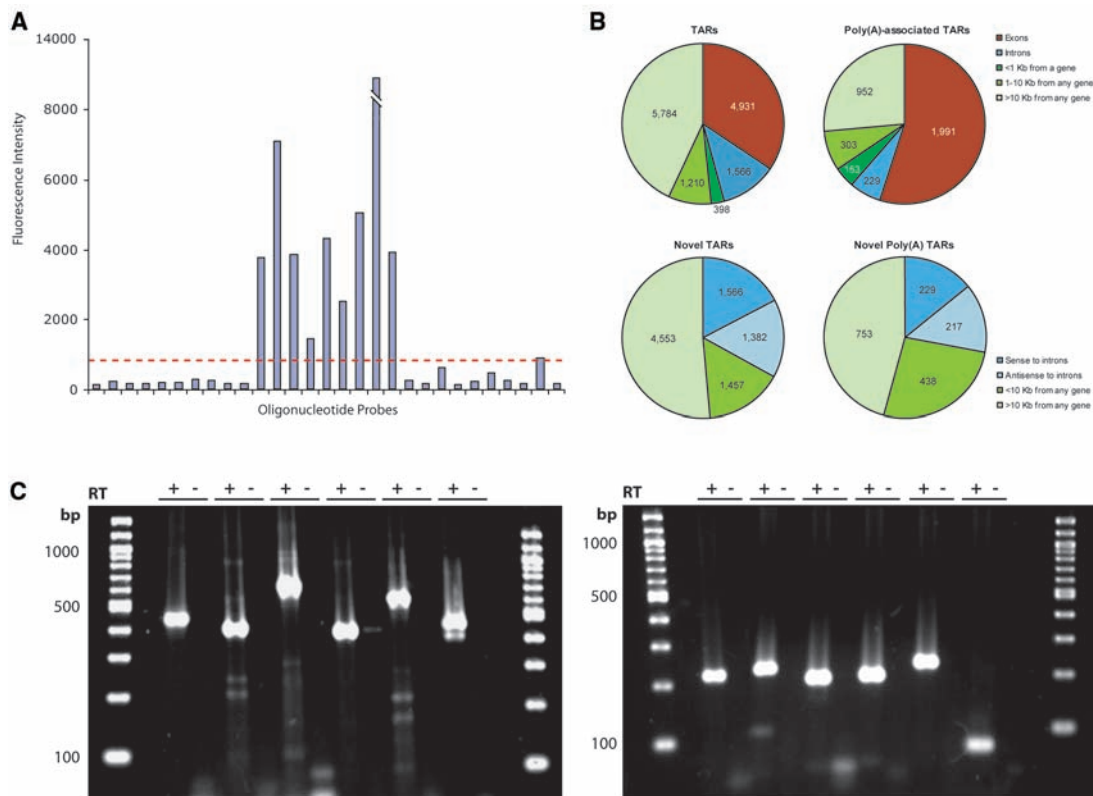
across nonrepetitive regions of the chromosome. **(B)** Proportion of genes detected from each of four annotation sources. The percentages of genes detected from each data set increase as the annotation shifts from solely ab initio predictions (Genscan) to fully characterized genes (RefSeq).

**Table 1.** Distribution of TARs relative to published gene annotation. Many TARs (40%) correspond to known exons; however, a significant fraction (38%) are located more than 10 kb from any previously annotated gene. BLAST results compare TARs to mammalian protein sequences and to the mouse genome. A total of 6934 (40%) of all TARs are homologs to the mouse genome (*e*-value ≤ 10<sup>-5</sup>), with 5656 (32%) homologous to protein sequences (25 to 30% of TARs belong to both categories), providing evidence for possible functional roles in humans.

|                         | Total               | Exons             | Introns           | <1 kb                     | 1 to 10 kb        | >10 kb            |
|-------------------------|---------------------|-------------------|-------------------|---------------------------|-------------------|-------------------|
| TARs                    | 13,889              | 4,931             | 1,566             | 398                       | 1,210             | 5,784             |
| Poly(A)-associated TARs | 3,628               | 1,991             | 229               | 153                       | 303               | 952               |
| Type I (AATAAA)         | 2,393               | 1,371             | 137               | 105                       | 187               | 593               |
| Type II (ATTAAG)        | 1,325               | 674               | 101               | 51                        | 123               | 376               |
|                         | BLAST: mouse genome |                   |                   | BLAST: mammalian proteins |                   |                   |
|                         | 1e <sup>-5</sup>    | 1e <sup>-10</sup> | 1e <sup>-20</sup> | 1e <sup>-5</sup>          | 1e <sup>-10</sup> | 1e <sup>-20</sup> |
| TARs                    | 5,419               | 4,747             | 3,761             | 4,349                     | 4,008             | 3,311             |
| Poly(A)-associated TARs | 1,515               | 1,247             | 936               | 1,307                     | 1,198             | 995               |
| Type I (AATAAA)         | 1,044               | 862               | 637               | 905                       | 830               | 685               |
| Type II (ATTAAG)        | 517                 | 423               | 328               | 436                       | 401               | 340               |

## REPORTS

**Fig. 2. (A)** Example TAR: A series of consecutive probes in the genome with fluorescence intensities that rank above the 90th percentile over all probes on the array (indicated with a dashed line). **(B)** Distribution of TARs relative to annotated genes. Occupancy within gene components and proximity to known genes are depicted for all TARs (upper charts) and for novel TARs that lie outside annotated exons (lower charts). Most of the novel TARs are located more than 10 kb from any previously annotated gene, suggesting that these correspond to distinct transcribed sequences. **(C)** RT-PCR validation of TAR sequences. A group of variable-length TARs between 400 and 650 bp is shown (left) opposite a group of approximately equal-length poly(A)-associated TARs (right). PCR products are loaded adjacent to their corresponding negative control samples.



**Fig. 3.** Conservation between TARs and other mammalian sequences. Forty-one percent of TARs and 50% of poly(A)-associated TARs were found to be homologous, as were 29% and 39% of novel TARs from each category. A large number of TARs show significant similarity to known proteins (BLAST  $e$ -values  $\leq 10^{-5}$ ), suggesting that many of these may be functional elements. A subset of these exhibited sequence similarity to regions of the mouse genome when restricted to similar  $e$ -values (solid blue sections).

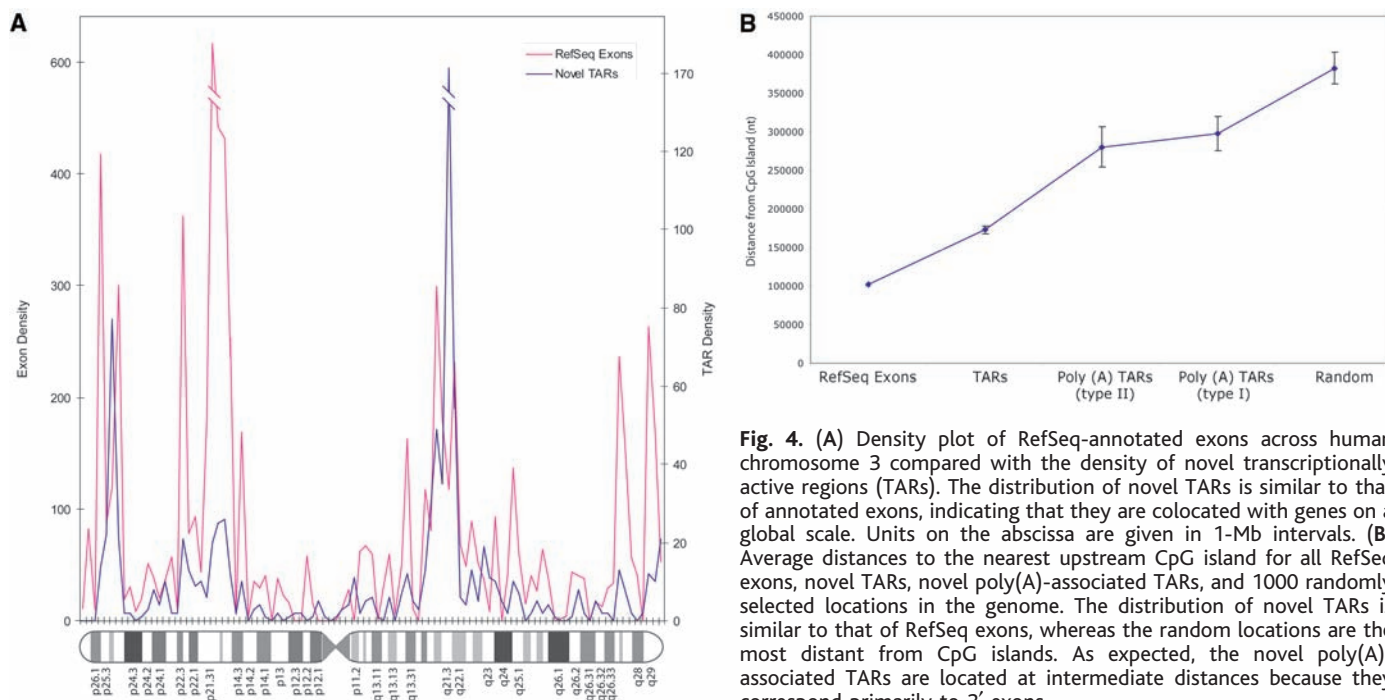
introns, indicating that many of the intronic TARs likely represent novel transcription units. Over half of all TARs were found to

be distal to annotated genes (greater than 10 kb from any gene), indicating the presence of an additional 5784 transcribed elements that are apparently unrelated to known genes.

We also used an independent set of criteria to identify TARs in which probe hybridization intensities were correlated with the presence of a polyadenylation signal 3' of the active region. Here we considered transcription units of (exactly) five consecutive probes with fluorescence intensities in the top 80th intensity percentile appearing in windows of 250 nt, where the 3' region contains or lies near a polyadenylation signal (18). Instances of "AATAAA" sequences were designated type I, and "ATATAA" type II. An additional 3628 TARs were identified using this method; ~100 such instances are expected to occur at random in the genome. Most (1991) lie within annotated exons, whereas 952 are located more than 10 kb from any annotated gene. Of the 1371 type I and 674 type II poly(A) sequences identified within exons of known genes, 94% (1289) of type I and 90% (607) of type II instances occur in the 3' exon of the gene in question, a strong indication of the effectiveness of this approach. The fraction of poly(A) TARs distinct from annotated exons (1637), combined with the 8948 novel TARs identified above, yields a total of 10,585 new transcribed sequences throughout the genome.

To validate the transcription of identified TARs with an independent method, we performed reverse transcriptase polymerase chain reaction (RT-PCR) assays using human liver poly(A)<sup>+</sup> RNA, targeting 48 poly(A)-associated and 48 non-poly(A)-associated TARs (10). Reactions were carried out in the presence and absence of reverse transcriptase; the latter served as a negative control. Of the 96 reactions, 90 (94%) amplified PCR products of the expected size in a single-pass assay with no detectable signal observed in the negative control (Fig. 2C). As a further validation, we compared the novel TARs against data derived from the second phase of the Kapranov *et al.* transcript mapping experiment on chromosomes 21 and 22 (4). We found that 41% of TARs match the transcribed fragments, or "transfrags," identified in their study. Because of the highly stringent selection of TARs in the present study, many low-abundance transcripts are not identified by these criteria, and we expect to have an appreciable false-negative rate.

We next compared the novel TARs with other mammalian DNA sequences to assess their potential for coding functional elements. BLAST (19) comparisons revealed that many TARs are homologous to sequences in the mouse genome. Of the 8958 novel TARs, 24% (2185) produced BLAST alignments with  $e$ -values less than  $10^{-5}$ , with most of these (1486) having  $e$ -values less than  $10^{-20}$ . This compares to 39%



**Fig. 4.** (A) Density plot of RefSeq-annotated exons across human chromosome 3 compared with the density of novel transcriptionally active regions (TARs). The distribution of novel TARs is similar to that of annotated exons, indicating that they are colocalized with genes on a global scale. Units on the abscissa are given in 1-Mb intervals. (B) Average distances to the nearest upstream CpG island for all RefSeq exons, novel TARs, novel poly(A)-associated TARs, and 1000 randomly selected locations in the genome. The distribution of novel TARs is similar to that of RefSeq exons, whereas the random locations are the most distant from CpG islands. As expected, the novel poly(A)-associated TARs are located at intermediate distances because they correspond primarily to 3' exons.

(5419) of the initial set of 13,889 TARs (i.e., novel TARs and those corresponding to exons of known genes) that produced BLAST scores with  $e$ -values less than  $10^{-5}$ ; 3761 of these had  $e$ -values less than  $10^{-20}$ . Similarly, 32% (532) of the 1637 novel poly(A)-associated TARs yielded BLAST alignments with  $e$ -values less than  $10^{-5}$ , with 342 less than  $10^{-20}$  (Fig. 3). Of the initial set of 5419 TARs and 1515 poly(A)-associated TARs found to be homologous to sequences in the mouse genome, 27% (1488) and 21% (321) from each category are located more than 10 kb from any previously annotated gene.

In addition to assessing the degree of genome conservation, we compared mouse proteins with TAR sequences that were translated in all possible reading frames (Table 1). A total of 16% (1427) and 12% (1091) of novel TARs produced BLAST matches less than  $10^{-5}$  and  $10^{-20}$ , respectively, compared with 31% (4329) and 24% (3311) of the total number of TARs with matches below these  $e$ -values. Higher percentages of poly(A)-associated TARs were found to be homologous to mouse proteins: 23% (369) of the novel subset and 36% (1307) of the total set of poly(A) TARs matched protein sequences with  $e$ -values less than  $10^{-5}$ , with 19% (305) and 27% (995) in each category having  $e$ -values less than  $10^{-20}$ . Thus, although many TARs are expected to encode proteins, novel TARs generally exhibit a lesser degree of sequence conservation than those intersecting known genes. This is particularly true for poly(A)-associated TARs owing to the higher degree

of conservation of protein-coding sequences relative to 3'-untranslated regions.

To estimate the number of TARs potentially arising from the cross-hybridization of mRNA transcripts to sequences elsewhere in the genome, we compared 9408 novel TARs that additionally do not lie antisense to annotated exons to the library of human cDNA sequences in the Ensembl database. We found only 11% (1034) with at least 95% identity over a stretch of 150 nt. Of the remaining 8374 nonhomologous novel TARs, 347 were found to intersect the genomic coordinates of processed pseudogenes (20, 21), providing evidence for possible pseudogenic transcription.

Finally, we examined the distribution of TARs relative to the locations of known genes and CpG islands. A density plot comparing TARs and RefSeq-annotated exons along chromosome 3 (Fig. 4A) revealed that TARs are located in the same regions as known genes. The density of TARs is correlated with the distribution of RefSeq-annotated genes along each chromosome (Pearson correlation coefficient  $r^2 = 0.35$ ,  $P < 0.002$ ). Comparison of distances to the nearest upstream CpG island indicates that the relative locations of novel TARs distal to annotated genes are similar to those of RefSeq exons, whereas the distal poly(A)-associated TARs are located farther away, which is expected because most of these should correspond to the 3' ends of genes (Fig. 4B). The distances of all distal TARs to CpG islands were found to be significantly less than those of randomly selected locations ( $P < 0.0001$ ).

Our findings demonstrate that it is possible to use high-resolution oligonucleotide microarrays for the comprehensive analysis of the human genome. Because many transcribed sequences are located in distinct intergenic regions distant from known genes, their precise mapping can only be accomplished using genomic tiling arrays in which nearly all of the nonrepetitive DNA is available for hybridization to RNA transcripts. Several bacterial artificial chromosome (BAC) clone-based genomic tiling arrays have been developed for comparative genomic hybridization (CGH) studies in humans (22, 23); however, the identification of short transcription units requires interrogating the genome sequence at a resolution of tens of base pairs, a measurement that is not possible to obtain with BAC technology.

In summary, we identified thousands of new transcribed regions and confirmed the transcription of predicted genes on a global scale. Our results provide a draft expression map for the entire genome, revealing a much more extensive and diverse set of expressed sequences than was previously annotated. Conservation between many of the novel transcribed sequences and well-characterized mouse proteins provides strong evidence that a large number of them are likely to encode functional transcripts. Many conserved transcribed sequences are located in regions distal to known genes, and a notable fraction of these are of sufficient length to encode proteins of 300 or more amino acids. The remainder may encode small proteins, untranslated exons, or RNAs whose functions have yet to be elucidated

(24, 25). These latter RNAs may serve alternate regulatory or structural roles and await detailed characterization.

References and Notes

1. D. W. Selinger *et al.*, *Nat. Biotechnol.* **18**, 1262 (2000).
2. B. Tjaden *et al.*, *Nucleic Acids Res.* **30**, 3732 (2002).
3. K. Yamada *et al.*, *Science* **302**, 842 (2003).
4. P. Kapranov *et al.*, *Science* **296**, 916 (2002).
5. J. L. Rinn *et al.*, *Genes Dev.* **17**, 529 (2003).
6. E. F. Nuwaysir *et al.*, *Genome Res.* **12**, 1749 (2002).
7. T. J. Albert *et al.*, *Nucleic Acids Res.* **31**, e35 (2003).
8. E. S. Lander *et al.*, *Nature* **409**, 860 (2001).
9. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
10. Materials and methods are available as supporting material on Science Online. Additional information can be found at <http://transcriptome.gersteinlab.org>. Experimental data and associated microarray designs are available in the NCBI Gene Expression Omnibus (GEO) under series GSE1904, sample records

- GSM34073 to GSM34213, and platform records GPL1539 to GPL1673.
11. P. Bertone *et al.*, data not shown.
12. K. D. Pruitt *et al.*, *Trends Genet.* **16**, 44 (2000).
13. T. Hubbard *et al.*, *Nucleic Acids Res.* **30**, 38 (2002).
14. E. Birney *et al.*, *Genome Res.* **14**, 925 (2004).
15. Each probe is assigned a value of 1 if its fluorescence intensity is greater than the median intensity of all probes on the array, and 0 otherwise. For a given gene, the expected count of 1's within annotated exons follows a binomial distribution; an unusually high count of 1's therefore yields low *P* values (sign test). Genes having *P* values < 0.05 were regarded as demonstrating positive hybridization.
16. C. Burge, S. Karlin, *J. Mol. Biol.* **268**, 78 (1997).
17. D. L. Wheeler *et al.*, *Nucleic Acids Res.* **31**, 28 (2003).
18. Polyadenylation signals are required to appear downstream of the 15th nucleotide of the 3' oligo-nucleotide in the transcribed region. An additional 51 (46 + 5) downstream nucleotides are included in the calculation to ensure full coverage of the sequence.

19. S. F. Altshul *et al.*, *J. Mol. Biol.* **215**, 403 (1990).
20. P. M. Harrison *et al.*, *Genome Res.* **12**, 272 (2002).
21. Z. Zhang *et al.*, *Genome Res.* **13**, 2541 (2003).
22. P. G. Buckley *et al.*, *Hum. Mol. Genet.* **11**, 3221 (2002).
23. A. S. Ishkanian *et al.*, *Nat. Genet.* **36**, 299 (2004).
24. J. S. Mattick, *Bioessays* **25**, 930 (2003).
25. D. Kampa *et al.*, *Genome Res.* **14**, 331 (2004).
26. This work was supported by NIH grant P50 HG02357.

Supporting Online Material

[www.sciencemag.org/cgi/content/full/1103388/DC1](http://www.sciencemag.org/cgi/content/full/1103388/DC1)  
 Materials and Methods  
 Microarray hybridization protocols  
 DNA sequences of transcriptionally active regions  
 Fig. S1

29 July 2004; accepted 3 November 2004  
 Published online 11 November 2004;  
 10.1126/science.1103388  
 Include this information when citing this paper.

# Use of Logic Relationships to Decipher Protein Network Organization

Peter M. Bowers,<sup>1,2</sup> Shawn J. Cokus,<sup>3</sup>  
 David Eisenberg,<sup>1,2</sup> Todd O. Yeates<sup>2,4\*</sup>

A major focus of genome research is to decipher the networks of molecular interactions that underlie cellular function. We describe a computational approach for identifying detailed relationships between proteins on the basis of genomic data. Logic analysis of phylogenetic profiles identifies triplets of proteins whose presence or absence obey certain logic relationships. For example, protein C may be present in a genome only if proteins A and B are both present. The method reveals many previously unidentified higher order relationships. These relationships illustrate the complexities that arise in cellular networks because of branching and alternate pathways, and they also facilitate assignment of cellular functions to uncharacterized proteins.

The sequencing of multiple genomes from diverse species has tremendous potential to impact our understanding of biology, both by providing a census of all proteins and by enabling subsequent analysis of their functions (1–6). Various patterns across multiple complete genomes have been used to infer biological interactions and functional linkages between proteins (6–14). These include observations of two distinct proteins from one organism being genetically fused into a single protein in another organism (13, 14) and the tendency of two proteins to occur in chromosomal proximity across multiple organisms (12, 15). When a sufficiently large number of genomes were fully sequenced, it became possible with the phylogenetic profile approach (11, 16, 17) to detect functional relationships between proteins exhibiting statistically similar patterns of presence or absence. Because

sequenced genomes allow us to catalog all of the proteins encoded in each organism, we can determine the pattern describing a protein's presence or absence by searching for its homologs across *N* organisms, the result of which is an *N*-dimensional vector of ones (present) and zeros (not present) referred to as its phylogenetic profile.

Original implementations of the phylogenetic profile method sought to infer “links” between pairs of proteins with similar profiles (11). A subsequent variation on that idea linked proteins if their profiles represented the negation of each other (18, 19). These ideas are consistent with the simplest notion of how two proteins might be related in a cell, with the presence of one protein implying the presence or absence of another. Such simple patterns might be expected when two proteins are required to form a structural complex or when two proteins carry out sequential steps in an unbranched metabolic pathway. However, such simple relationships cannot adequately describe the full complexity of cellular networks that involve branching, parallel, and alternate pathways.

The observed complexity of cellular networks leads one to expect the existence of higher order logic relationships involving a pattern of presence or absence of multiple proteins. Furthermore, evolutionary divergence, convergence, and horizontal transfer events lead us to expect relationships between multiple gene families that are more complex than can be described by pairwise phylogenetic similarity. Analysis of cellular pathways and networks in terms of logic relations has attracted recent interest (20, 21), and the growing number of sequenced genomes now makes it possible to search for logic relations.

Here, we perform a complete analysis of the logic relations possible between triplets of phylogenetic profiles and demonstrate the power of the resulting logic analysis of phylogenetic profiles (LAPP) in illuminating relationships among multiple proteins and inferring the coarse function of large numbers of uncharacterized protein families. There are eight possible logic relationships combining two phylogenetic profiles to match a third profile (Fig. 1A). For instance, protein C might be present if and only if proteins A and B are both present (denoted here as a type 1 logic relationship), from which we would infer that the function of protein C is necessary only when the functions of proteins A and B are both present. Alternatively, gene C may be present if and only if either A or B is present (a type 7 logic relationship), which is seen when different organisms use two different protein families in combination with a common third protein to accomplish some task (for example, a combination of A and C or B and C). Several of the eight possible logic relationships can be intuitively understood to describe commonly observed biological scenarios, whereas a few of the logic relationships are not easily related to real biological situations.

To identify protein triplets that exhibit the logic relationships described in Fig. 1, we first created a set of binary-valued vectors describing the presence or absence of

<sup>1</sup>Howard Hughes Medical Institute, <sup>2</sup>Institute for Genomics and Proteomics, <sup>3</sup>Department of Mathematics, <sup>4</sup>Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, CA 90095, USA.

\*To whom correspondence should be addressed.  
 E-mail: yeates@mbi.ucla.edu