Sciencexpress Report

# Global Identification of Human Transcribed Sequences with Genome Tiling Arrays

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

[1]Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT 06520–8103, USA. [2]Center for Nanotechnology, NASA Ames Research Center, Moffett Field, CA 94035, USA. [3]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520–8114, USA. [4]Eloret Corporation, Sunnyvale, CA 94087, USA [5]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.)

**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 novel 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 approximately 1.5 Gb of nonrepetitive genomic DNA from each strand of the human genome (*8*, *9*). A total of 51,874,388 36mer oligonucleotide 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 approximately 390,000 probes per array [Fig. S1 (*10*)]. To measure transcriptional activity the arrays were hybridized to fluorescence-labeled cDNA reverse-transcribed from triple-selected poly $(A)^+$ 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 $r^2$ correlations between 0.90 and 0.95 (data not shown) indicating that the experiments are highly reproducible. To further reduce the effect of potential variation across individual cDNA samples, pooled reverse transcription products of 20 separate labeling reactions were used 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 (*11*) and Ensembl (*12*, *13*) 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 (*14*). 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 (*15*). Using 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 9,844 RefSeq genes whose corresponding UniGene (*16*) annotations indicate transcription in liver tissue was also examined; 70% (6,907) of these were detected using our approach (Table 1A).

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 1) unannotated exons from alternatively spliced messages, 2) under-represented 3' and 5' untranslated regions, 3) non-protein coding RNA transcripts, and 4) novel transcripts coding for functional proteins. We considered aggregate transcription units consisting of at least 5 consecutive probes exhibiting fluorescence intensities in the top 90th intensity percentile, and whose genomic coordinates lay within a 250 nt window (Fig. 2A). These were compiled from throughout the genome and their locations compared relative to annotated gene components (Fig. 2B). A total of 13,889 transcription units, ranging in size from 209 to 3,438 nt, were identified in the genome by these criteria; approximately 400 are expected under the null hypothesis of zero transcription. One-third (4,931) correspond to previously annotated exons; the remaining 8,958 are new transcribed sequences that we refer to as transcriptionally active regions, or TARs (*5*). Interestingly, 1,566 TARs were located within previously annotated introns on the same strand, raising the possibility that they correspond to overlooked exons. However, an equal number of TARs (1,529) lie on the antisense strand of 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 5,784 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) 5 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 (*17*). Instances of 'AATAAA' sequences were designated type I, and 'ATTAAA' type II. An additional 3,628 TARs were identified using this method; approximately 100 such instances are expected to occur at random in the genome. Most (1,991) lie within annotated exons, while 952 are located more than 10 Kb from any annotated gene. Of the 1,371 type I and 674 type II poly (A) sequences identified within exons of known genes, 94% (1,289) 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 disctinct from annotated exons (1,637), combined with the 8,948 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 $(A)^+$ 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*), finding that 41% of TARs match the transcribed fragments, or "transfrags," identified in their study. Due to 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 (*18*) comparisons revealed that many TARs are homologous to sequences in the mouse genome. Of the 8,958 novel TARs, 24% (2,185) produced BLAST alignments with *e*-values less than $10^{-5}$, with most of these (1,486) having *e*-values less than $10^{-20}$. This compares to 39% (5,419) 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 $10^{-5}$; 3,761 of these had *e*-values less than $10^{-20}$. Similarly, 32% (532) of the 1,637 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 5,419 TARs and 1,515 poly (A)-associated TARs found to be homologous to sequences in the mouse genome, 27% (1,488) and 21% (321) from each category are located greater 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 1B). A total of 16% (1,427) and 12% (1,091) of novel TARs produced BLAST matches less than $10^{-5}$ and $10^{-20}$, respectively, compared with 31% (4,329) and 24% (3,311) 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% (1,307) 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 due 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 9,408 novel TARs that additionally do not lie antisense to annotated exons to the library of human cDNA sequences in the Ensembl database, finding only 11% (1,034) with at least 95% identity over a stretch of 150 nt. Interestingly, of the remaining 8,374 non-homologous novel TARs, 347 were found to intersect the genomic coordinates of processed pseudogenes (*19*, *20*), 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 is illustrated in Figure 4A, revealing 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 $r^2 = 0.35$, $P$-value < 0.002). Comparing 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, while the distal poly (A)-associated TARs are located farther away, which is expected since the majority 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 ($P$-value < 0.0001).

These studies demonstrate that it is possible to use high-resolution oligonucleotide microarrays for the comprehensive analysis of the human genome. Since many transcribed sequences are located in distinct intergenic regions distant from known genes, their precise mapping can only be accomplished using genomic tiling arrays where nearly all of the nonrepetitive DNA is available for hybridization to RNA transcripts. Several BAC clone-based genomic tiling arrays have been developed for comparative genomic hybridization (CGH) studies in humans (*21*, *22*); 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 reminder may encode small proteins, untranslated exons or RNAs whose functions have yet to be elucidated (*23*, *24*). 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. Experimental data and associated microarray designs are available in the GEO database under series GSE1904, sample records GSM34073-GSM34213, and platform records GPL1539-GPL1673.
11. K. D. Pruitt *et al*., *Trends Genet*. **16**, 44 (2000).
12. T. Hubbard *et al*., *Nucleic Acids Res*. **30**, 38 (2002).
13. E. Birney *et al*., *Genome Res*. **14**, 925 (2004).
14. 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.
15. C. Burge and S. Karlin, *J. Mol. Biol*. **268**, 78 (1997).
16. D. L. Wheeler *et al*., *Nucleic Acids Res*. **31**, 28 (2003).
17. Polyadenylation signals are required to appear downstream of the 15th nucleotide of the 3' oligo in the transcribed region. An additional 51 (46 + 5) downstream nucleotides are included in the calculation to ensure full coverage of the sequence.
18. S. F. Altshul *et al*., *J. Mol. Biol.* **215**, 403 (1990).
19. P. M. Harrison *et al*. *Genome Res*. **12**, 272 (2002).
20. Z. Zhang *et al*., *Genome Res*. **13**, 2541 (2003).
21. P. G. Buckley *et al*., *Hum. Mol. Genet*. **11**, 3221 (2002).
22. A. S. Ishkanian *et al*., *Nat. Genet*. **36**, 299 (2004).
23. J. S. Mattick. *Bioessays* **25**, 930 (2003).
24. D. Kampa *et al*., *Genome Res*. **14**, 331 (2004).
25. Supplementary data are available online at transcriptome.gersteinlab.org.
26. This work was supported by NIH grant P50 HG02357.

**Supporting Online Material**
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.

**Table 1.** (**A**) Distribution of TARs relative to published gene annotation. Many TARs (40%) correspond to known exons; however, a significant fraction (38%) are located greater than 10 Kb from any previously annotated gene. (**B**) BLAST results comparing TARs to mammalian protein sequences and to the mouse genome. A total of 6,934 (40%) of all TARs are homologs to the mouse genome (*e*-value ≤ $10^{-5}$), with 5,656 (32%) homologous to protein sequences (25-30% of TARs belong to both categories), providing evidence for possible functional roles in humans.

**A**

| | **Total** | **Exons** | **Introns** | **<1 Kb** | **1-10 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 (ATTAAA) | 1,325 | 674 | 101 | 51 | 123 | 376 |

**B**

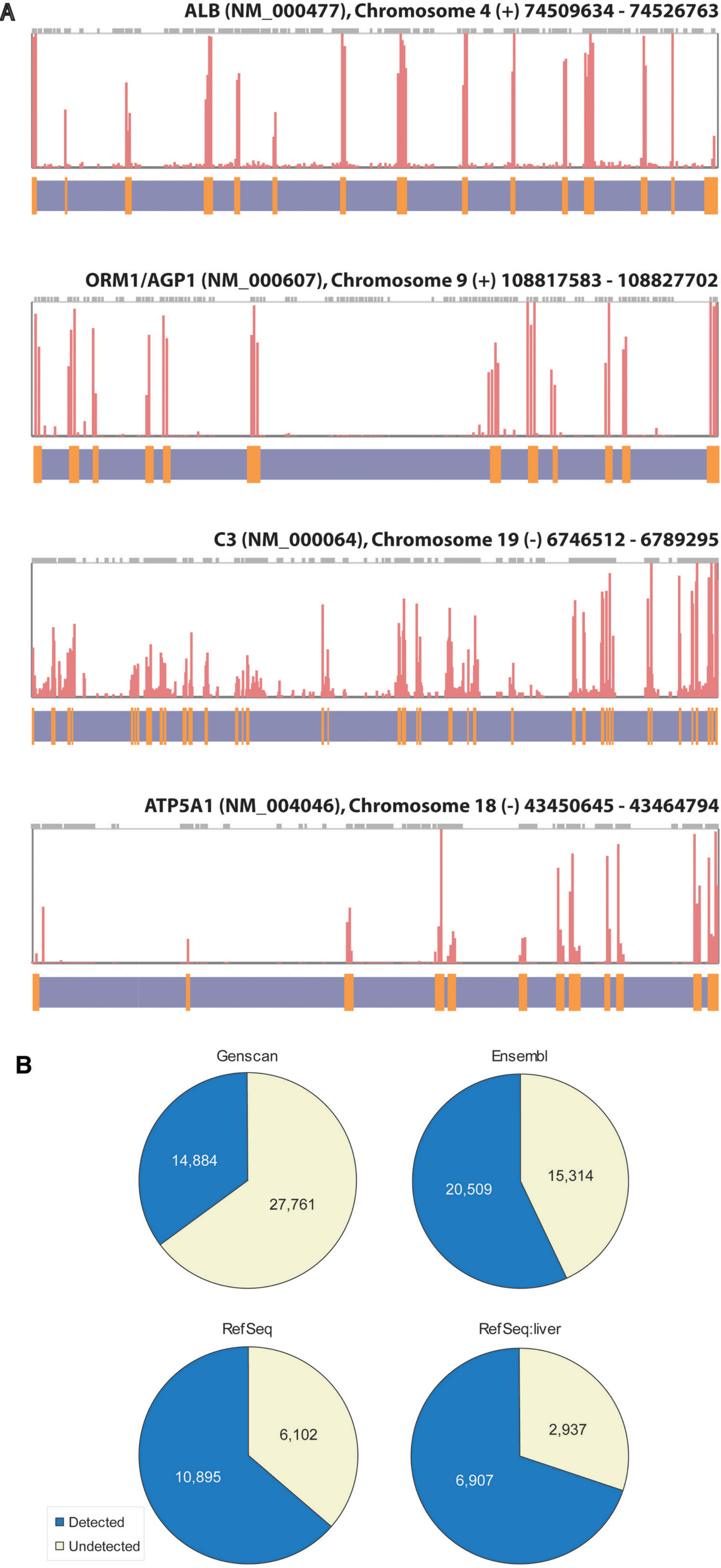
| | **BLAST: Mouse Genome** | | | **BLAST: Mammalian Proteins** | | |
|---|---|---|---|---|---|---|
| | $1e^{-5}$ | $1e^{-10}$ | $1e^{-20}$ | $1e^{-5}$ | $1e^{-10}$ | $1e^{-20}$ |
| **TARs** | 5,419 | 4,747 | 3,761 | 4,349 | 4,008 | 3,311 |
| **Poly(A)-associated TARs** | 1,515 | 1,247 | 936 | 1,307 | 1,198 | 995 |
| Type I (AATAAA) | 1,044 | 862 | 637 | 905 | 830 | 685 |
| –Type II (ATTAAA) | 517 | 423 | 328 | 436 | 401 | 340 |

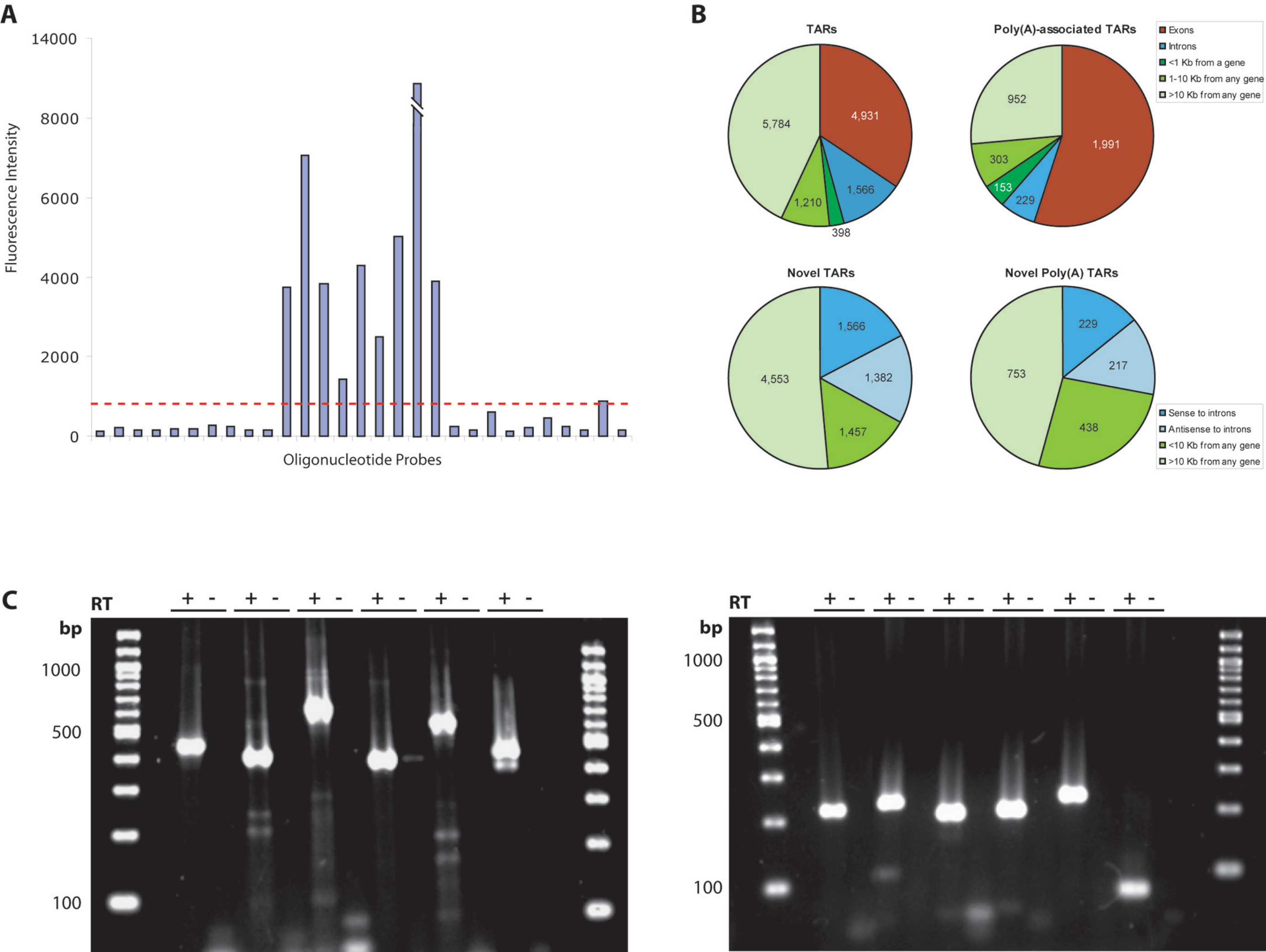
**Fig. 1.** (**A**) Annotated genes aligned with microarray fluorescence intensities. Comparison of the gene structures with intensity data shows strong agreement with expected exon-intron boundaries. The upper two examples illustrate uniform representation across the entire gene, while the lower two examples show a slight 3' bias inherent in reverse-transcription labeling of mRNA. Grey 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).

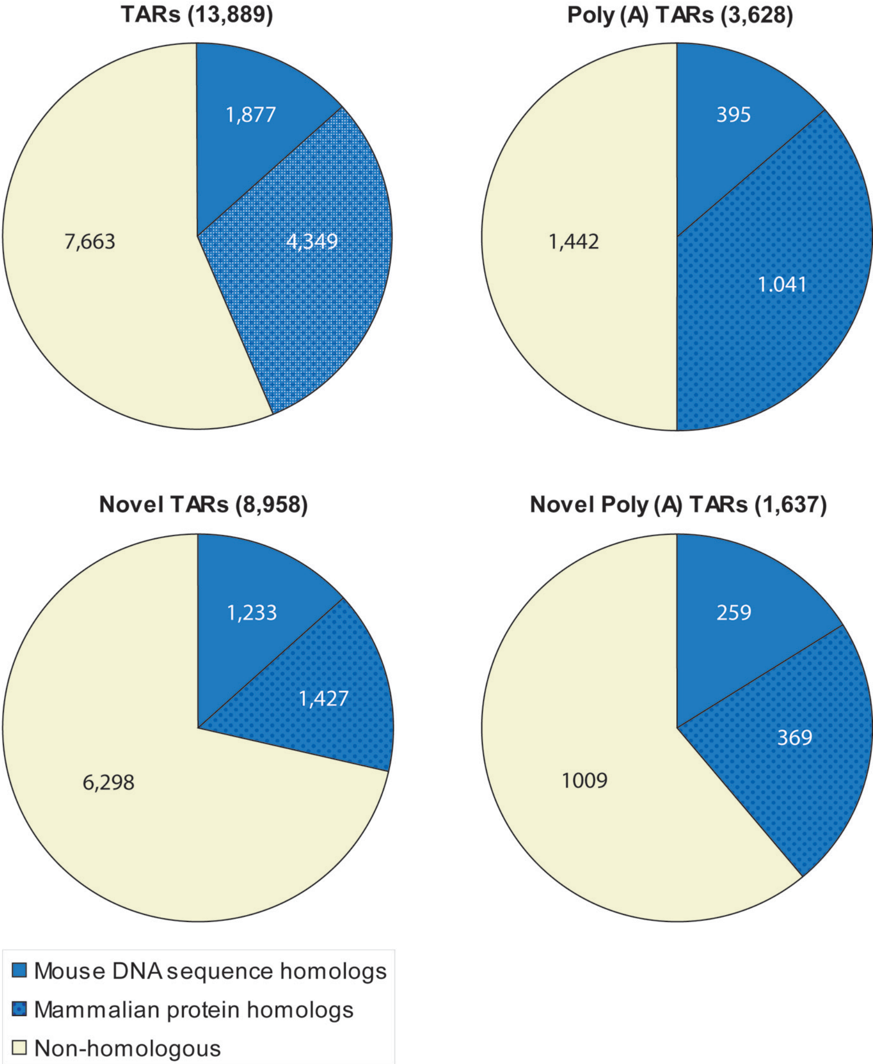
**Fig. 2.** (**A**) Example TAR: a series of consecutive probes in the genome whose fluorescence intensities rank above the 90th percentile over all probes on the array (indicated with a dashed line). (**B**) Distribution of TARs relative to annotated genes. Occupancy within gene components and proximity to known genes is 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.

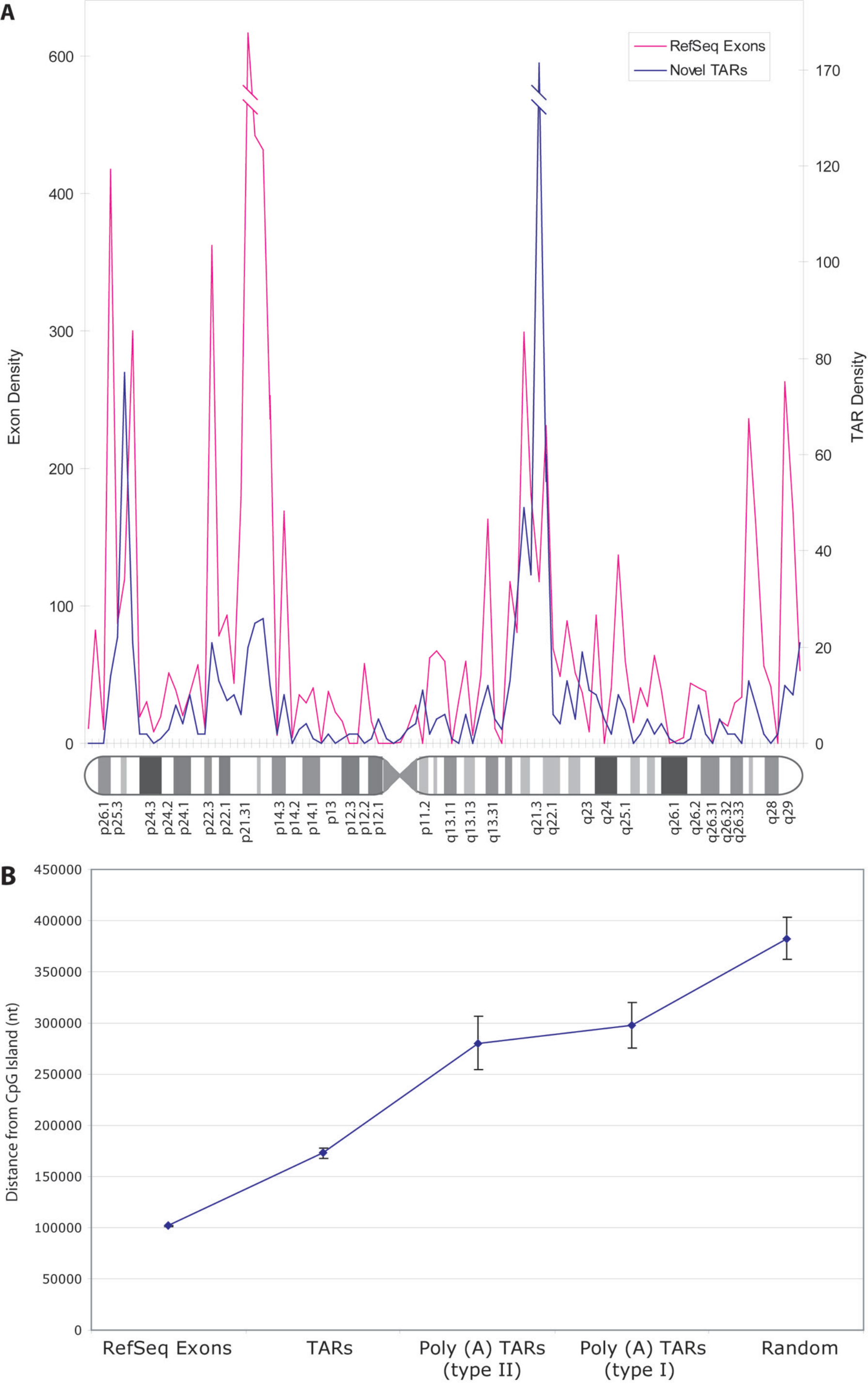
**Fig. 3.** Conservation between TARs and other mammalian sequences. 41% of TARs and 50% of poly (A)-associated TARs were found to be homologous, as were 29% and 39% of novel TARs from each category. A large number of TARs show significant similarity to known proteins (BLAST *e*-values $\leq 10^{-5}$), 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 (plain blue sections).

**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 co-located 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 1,000 randomly selected locations in the genome. Novel TARs are distributed similarly to RefSeq exons, while the random locations are the most distant from CpG islands. As expected, the novel poly (A)-associated TARs are located at intermediate distances since they correspond primarily to 3' exons.


A
ALB (NM_000477), Chromosome 4 (+) 74509634 - 74526763
ORM1/AGP1 (NM_000607), Chromosome 9 (+) 108817583 - 108827702
C3 (NM_000064), Chromosome 19 (-) 6746512 - 6789295
ATP5A1 (NM_004046), Chromosome 18 (-) 43450645 - 43464794
B
Genscan
14,884
27,761
Ensembl
20,509
15,314
RefSeq
10,895
6,102
RefSeq:liver
6,907
2,937
Detected
Undetected


A
14000
8000
6000
4000
2000
0
Fluorescence Intensity
Oligonucleotide Probes
B
TARs
5,784
4,931
1,566
1,210
398
Poly(A)-associated TARs
952
1,991
303
153
229
Exons
Introns
<1 Kb from a gene
1-10 Kb from any gene
>10 Kb from any gene
Novel TARs
1,566
1,382
4,553
1,457
Novel Poly(A) TARs
229
217
753
438
Sense to introns
Antisense to introns
<10 Kb from any gene
>10 Kb from any gene
C
RT
+ - + - + - + - + - + -
bp
1000
500
100
RT
+ - + - + - + - + - + -
bp
1000
500
100


TARs (13,889)
1,877
4,349
7,663
Poly (A) TARs (3,628)
395
1.041
1,442
Novel TARs (8,958)
1,233
1,427
6,298
Novel Poly (A) TARs (1,637)
259
369
1009
Mouse DNA sequence homologs
Mammalian protein homologs
Non-homologous


A
RefSeq Exons
Novel TARs
Exon Density
TAR Density
0
100
200
300
400
600
0
20
40
60
80
100
120
170
p26.1
p25.3
p24.3
p24.2
p24.1
p22.3
p22.1
p21.31
p14.3
p14.2
p14.1
p13
p12.3
p12.2
p12.1
p11.2
q13.11
q13.13
q13.31
q21.3
q22.1
q23
q24
q25.1
q26.1
q26.2
q26.31
q26.32
q26.33
q28
q29
B
Distance from CpG Island (nt)
0
50000
100000
150000
200000
250000
300000
350000
400000
450000
RefSeq Exons
TARs
Poly (A) TARs (type II)
Poly (A) TARs (type I)
Random