Letter

Mapping the *C. elegans* noncoding transcriptome with a whole-genome tiling microarray

Housheng He,[1,2,8] Jie Wang,[1,2,8] Tao Liu,[1,2,8] X. Shirley Liu,[3,4] Tiantian Li,[1,2] Yunfei Wang,[1,2] Zuwei Qian,[5] Haixia Zheng,[1,2] Xiaopeng Zhu,[1,2] Tao Wu,[1,2] Baochen Shi,[1,2] Wei Deng,[1] Wei Zhou,[5] Geir Skogerbø,[1,9] and Runsheng Chen[1,6,7,9]

[1]*Bioinformatics Laboratory and National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China;* [2]*Graduate School of the Chinese Academy of Science, Beijing 100080, China;* [3]*Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts 02115, USA;* [4]*Harvard School of Public Health, Boston, Massachusetts 02115, USA;* [5]*Affymetrix, Inc., Santa Clara, California 95051, USA;* [6]*Bioinformatics Research Group, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Science, Beijing 100080, China;* [7]*Chinese National Human Genome Center, Beijing 100176, China*

The number of annotated protein coding genes in the genome of *Caenorhabditis elegans* is similar to that of other animals, but the extent of its non-protein-coding transcriptome remains unknown. Expression profiling on whole-genome tiling microarrays applied to a mixed-stage *C. elegans* population verified the expression of 71% of all annotated exons. Only a small fraction (11%) of the polyadenylated transcription is non-annotated and appears to consist of ~3200 missed or alternative exons and 7800 small transcripts of unknown function (TUFs). Almost half (44%) of the detected transcriptional output is non-polyadenylated and probably not protein coding, and of this, 70% overlaps the boundaries of protein-coding genes in a complex manner. Specific analysis of small non-polyadenylated transcripts verified 97% of all annotated small ncRNAs and suggested that the transcriptome contains ~1200 small (<500 nt) unannotated noncoding loci. After combining overlapping transcripts, we estimate that at least 70% of the total *C. elegans* genome is transcribed.

[Supplemental material is available online at www.genome.org.]

In organisms previously analyzed by tiling microarrays, a substantial part of the non-annotated genome has consistently displayed transcriptional activity (Kapranov et al. 2002; Rinn et al. 2003; Yamada et al. 2003; Bertone et al. 2004; Cheng et al. 2005; Stolc et al. 2005; David et al. 2006; Li et al. 2006; Manak et al. 2006), suggesting the existence of either a larger-than-predicted number of protein-coding genes or a large number of non-protein-coding RNA (ncRNAs) genes. The current annotation of the 100-Mb *Caenorhabditis elegans* genome estimated ~22,000 protein-coding genes and ~1,000 small ncRNA genes (Chen et al. 2005; Stricklin et al. 2005), and computational predictions have suggested the presence of an additional ~3,000 small ncRNA genes in the genome (Deng et al. 2006; Missal et al. 2006). The fact that a 1000-cell nematode appears to contain nearly as many protein coding genes as the far more complex genomes of insects and vertebrates invites the question of whether it is equally rich in ncRNA genes.

Transcriptional analyses employing microarrays that constitute complete nonrepetitive tile paths over a genome or part of a genome, irrespective of the location of annotated genes (genomic tiling microarrays; Bertone et al. 2006), have recently been applied to a number of organisms. With the exception of a recent study of 10 human chromosomes (Cheng et al. 2005), most expression profiling studies on genomic tiling arrays have focused on the polyadenylated fraction of the transcriptome in the respective organisms (Kapranov et al. 2002; Rinn et al. 2003; Yamada et al. 2003; Bertone et al. 2004; Stolc et al. 2005; David et al. 2006; Li et al. 2006; Manak et al. 2006). In an attempt to map out a major fraction of the small noncoding transcriptome in the worm, we adapted a highly efficient protocol for small (<500 nt) ncRNA cloning and microarray sample preparation (Deng et al. 2006; He et al. 2006), and applied this to a newly released Affymetrix *C. elegans* whole genome tiling array. Profiling a small non-polyadenylated (SNPA) RNA sample on this tiling array provided high sensitivity and specificity for detecting small ncRNAs; at a threshold where 97% of the known ncRNAs were detected, >80% of the array-detected, previously unknown transcripts were verifiable by reverse transcription–polymerase chain reaction (RT-PCR) or rapid amplification of c-DNA ends (RACE) (Supplemental Document 1). Incorporating these results with those obtained from polyadenylated (PA) and non-polyadenylated (NPA) total RNA profiled on tiling arrays demonstrated several advantages of this approach with respect to the breadth and depth of the information that could be extracted.

Results

The Affymetrix *C. elegans* Tiling 1.0R array contains ~3.2 million 25-mer oligonucleotide probe pairs covering the Watson strand of the entire nonrepetitive genome at an average resolution (distance between the central position of adjacent probes) of 25 bp. RNA was extracted from a mixed-stage population of wild-type *C. elegans* strain N2 and reverse-transcribed into double-stranded

[8]These authors contributed equally to this work.
[9]Corresponding authors.
E-mail crs@sun5.ibp.ac.cn; fax 86-10-64889892
E-mail zgb@moon.ibp.ac.cn; fax 86-10-64889892.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6611807.

cDNA samples representing either PA RNA, NPA total RNA depleted in both polyadenylated RNAs and rRNAs, or SNPA RNAs.

When hybridized to the tiling array (see Methods for details), the PA, NPA, and SNPA samples gave rise to 23.5% (736,710), 18.2% (571,347), and 2.0% (63,292) of the probes with positive signals, respectively, amounting to a total of 917,753 positive probes representing 22.7% of the *C. elegans* genome. As single positive probes are likely to be the result of spurious nonspecific hybridization, we defined a putatively transcribed fragment (transfrag; The ENCODE Project Consortium 2004) as at least two positive probes separated by a gap of no more than 30 bp. The three samples individually produced 108,669, 97,548, and 5738 transfrags (Fig. 1), which after removal of redundancies suggested the presence of at least 146,249 stably expressed regions with an average and median length of 156 and 103 nt, respectively. Among the nonredundant transfrags, 95,928 (65.6%) are annotated protein-coding exons, 875 overlap with known small noncoding transcripts, and 6281 correspond to tandem repeats, pseudogenes, or transposons (Fig. 1). The remaining 43,165 then represent the lowest estimate for transcripts of unknown function (TUFs; The ENCODE Project Consortium 2004) detected by the tiling microarray.

The array detects 70% to 97% of annotated genes

To estimate the sensitivity of the tiling microarray, an annotated genomic element was regarded as detected if 30% or more of the interrogating probes were positive (Kampa et al. 2004). The highest detection rate for annotated genomic loci was observed for the SNPA sample, in which >97% of the known small ncRNA loci were detected. The detection rate depended somewhat on the ncRNA class, with tRNAs and snRNAs showing nearly 100% detection, whereas snoRNAs and uncharacterized RNAs detected at somewhat lower levels (Table 1). In comparison, the detection rate for small ncRNAs in the NPA sample was far lower with an average of 59% (65% for tRNAs, and 47% for other ncRNAs; Fig. 2A). Alternatively, using the poly(A)-tailless histone mRNAs in the NPA sample gave a detection rate of 97%, implying that random hexamer-primed reverse transcription may have biased the sample toward longer transcripts. MicroRNA precursors (pri- and pre-miRNAs) vary in length and polyadenylation status. We altogether detected signals corresponding to 64 out of 115 annotated miRNA precursor loci (55.6%) in the SNPA (46), PA (19), and in the NPA (29) data set, the lower detection rate for miRNA precursor probably caused by the lower stability of these transcripts (Bracht et al. 2004).

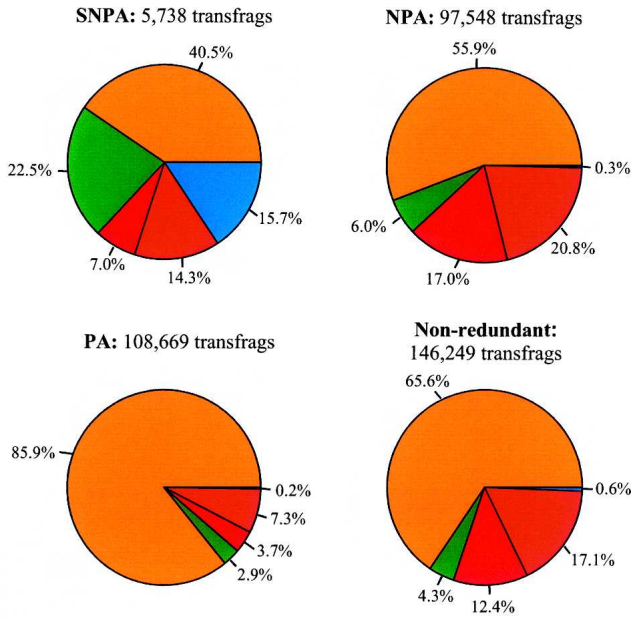

Figure 1. Transfrag distribution in the three different samples. "Other annotated" mainly includes tandem repeats and pseudogenes; "exons" include curated exons; "ncRNAs" include all tRNAs (Lowe and Eddy 1997), rRNAs, and other known noncoding RNA transcripts from NONCODE (Liu et al. 2005), RNAdb (Pang et al. 2005), miRBase (Griffiths-Jones et al. 2006), and recent literature (Deng et al. 2006; Zemann et al. 2006).

Table 1. Detection rates of annotated ncRNAs in the SNPA sample

ncRNA class	Known[a]	Detected	Fraction (%)
tRNAs	629	621	98.70
snoRNAs	146	132	90.41
snRNAs	102	101	99.02
snlRNAs	12	12	100
SRP RNAs	5	5	100
sbRNAs	13	13	100
Uncharacterized RNAs	23	21	91.30
All ncRNAs	**930**	**905**	**97.31**

[a]"Known" indicates known ncRNAs whose loci are interrogated by the tiling microarray.

The detection rate of an exon in the PA sample was on average 71% but depended on the confirmation status of its corresponding gene and varied from 94% for exons in fully confirmed genes to only 28% in predicted genes (Fig. 2B; Supplemental Fig. 1). To relate the developmental- and environmental-specific expression of genes from the mixed-population RNA hybridized to the tiling microarrays, the genes previously reported to express under a number of given conditions (Jiang et al. 2001; Wang and Kim 2003) were compared to the same genes detected in the PA sample (Fig. 2C). For most tested conditions, expression of between 90% and 97% of the previously reported genes were observed on the tiling array. The exceptions were genes predominantly expressed in males, of which only 60% were detected, most probably reflecting a low number of males present in the mixed *C. elegans* population used for RNA sample preparation.

Major part of non-annotated transcriptome is longer non-polyadenylated transcripts

Compared to most genomes analyzed by tiling microarray (Cheng et al. 2005; David et al. 2006; Li et al. 2006; Manak et al. 2006), only a relatively small fraction (11%) of the detected polyadenylated transcripts occurred outside annotated exons of protein-coding genes, and the majority of the detected non-protein-coding transcripts in *C. elegans* thus appear to be non-polyadenylated. Only a very small fraction of this transcription was detected in the SNPA sample. At a signal probe intensity cutoff of 6.1, the SNPA data contained 1222 transcripts without annotation. RT-PCR and RACE analysis confirmed 77% of a random sample of TUFs from this set (Supplemental Documents 1

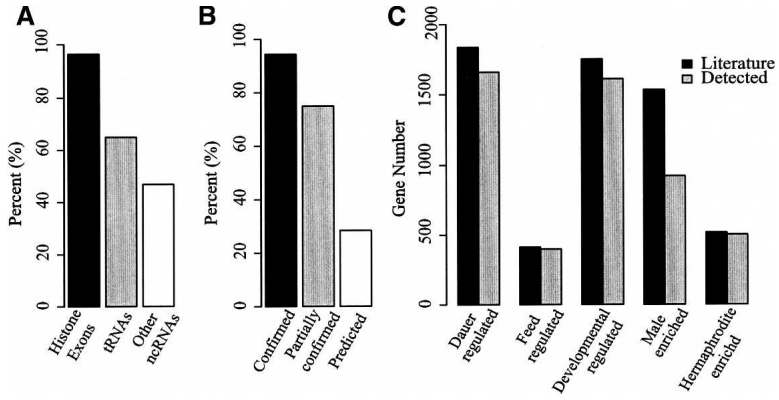

Figure 2. Detection rates of annotated exons and genes in the NPA and PA samples. (*A*) Detection rates for histone exons, tRNAs, and other small ncRNAs in the NPA sample. (*B*) Detection rates for exons in genes with different confirmation status. Confirmed, partially confirmed, and predicted genes relate to genes in which all, some, or no exons, respectively, have experimental verification (WormBase; Chen et al. 2005). (*C*) Number of genes with literature-reported expression under various developmental and environmental conditions compared to number of the same genes detected in the PA sample on the tiling arrays.

and 2). Contrary to an earlier analysis of the small noncoding transcriptome that found chromosome X to be nearly devoid of small ncRNAs (Deng et al. 2006), the SNPA TUF loci are nearly equally distributed on the *C. elegans* chromosomes, with a slight preference for chromosome X. Two-thirds of all TUF loci are intergenic, a higher fraction than for known small ncRNA loci (55%; Deng et al. 2006); however, the intergenic SNPA TUF loci show the same tendency as known ncRNA loci to locate in relative vicinity to annotated coding genes. The SNPA TUF loci appear less conserved than known and recently cloned loci, as only 21% show some conservation (weak WABA; Kent and Zahler 2000) in *C. briggsae,* and none was found to be conserved outside the nematodes. Further sequence analysis suggested that ~10% (126) of the SNPA TUFs may belong to various known ncRNA classes (mainly snoRNAs, snlRNAs, and sbRNAs), thus a far larger fraction of the SNPA loci may represent potentially novel functional categories of short RNAs than hitherto cloned transcripts. Analysis of sequence flanking the SNPA TUFs identified three known (UM1–3; Deng et al. 2006) and one novel (UM4) upstream motif at 143 of the most strongly expressed TUF loci (for further details see J. Wang, H. He, T. Liu, G. Skogerbø, and R. Chen, in prep.).

The NPA sample produced 97,548 transfrags, all of which could potentially represent noncoding transcripts (except transcripts coding for histones). Nearly 70% of the NPA transcripts overlap with annotated exons (55.9%) or introns (17%) of coding genes, whereas 20.8% are non-annotated intergenic TUFs. The NPA signal-to-background ratio is lower than for the other samples (Supplemental Fig. 1); however, RT-PCR analysis confirmed 90% (26/29) of randomly sampled intronic and intergenic TUFs, effectively excluding the possibility that the majority of the NPA TUFs are a result of nonspecific hybridization. RT-PCRs against eight regions of low signal intensity gave no positive amplification (Supplemental Document 1), further indicating that the NPA data are real and have picked up most of the existent non-polyadenylated transcription. TUFs in the NPA sample are also fairly well conserved, with 54% showing at least some level of conservation (weak WABA; Kent and Zahler 2000) in *Caenorhabditis briggsae.* Although some longer NPA TUFs were observed (the longest being 3579 nt), most are generally short (mean 88 nt, median 75 nt); however, of these only 557 overlapped with the SNPA TUFs, which seems unexpectedly few, considering the high specificity of the latter. This discrepancy may stem from a lower ability of random hexamer priming used for reverse transcription of the NPA sample to capture short ncRNAs (as compared to priming from a 3′-end–ligated adapter used for the SNPA sample), and short NPA TUFs located in close proximity may actually represent longer transcripts. We first tested this by randomly selecting eight pairs of TUFs separated by <500 bp, all of which could be individually validated by RT-PCR. Subsequent RT-PCRs with one primer in each of the paired two TUFs resulted in the amplification of fragments corresponding to the genomic distance between the TUFs in five of the eight pairs. No amplification was observed when reverse transcriptase was omitted from the reaction, indicating that results were not generated from contamination of genomic DNA but were instead results of unspliced transcripts spanning distance between the two TUFs (Supplemental Document 1). To further explore the possibility that NPA TUFs mostly represented longer fragments, we then attempted a nested 5′- and 3′-RACE approach for all 33 TUFs validated by RT-PCR (Supplemental Document 1). Amplified fragments were cloned and sequenced, and 11 of 33 yielded at least one positive 5′- or 3′-RACE sequence. Seven of the RACE fragments extended at least 30 nt beyond the TUF from which they were initiated, and in one case the RACE fragment was 1 kb longer than its corresponding TUF (see Supplemental material for details). Taken together, these data suggest that a considerable fraction of the non-polyadenylated transcripts in *C. elegans* are in the form of longer, unspliced RNAs.

Coding regions are a complex web of overlapping transcripts

The NPA signals overlapping genic (exonic and intronic) sequence are more difficult to interpret. These could be of the same nature as intergenic non-polyadenylated signals (i.e., independent of coding gene transcription) or, conversely, could simply represent fragments from mRNA splicing and degradation. The signal intensity distributions for NPA TUFs and exonic transfrags show little difference (Fig. 3A), and various analyses of the genic PA and NPA data favor a hypothesis that genic non-polyadenylated transcription is not principally different from the intergenic transcriptional output (for further details, see T. Liu, H. He, J. Wang, G. Skogerbø, and R. Chen, in prep.); however, the genic non-polyadenylated transcription appears at least in part to be composed of alternative, unspliced transcripts (possibly antisense) covering both exons and introns of the coding genes. There also appears to be a positive correlation between polyadenylated and non-polyadenylated activity within the same coding gene boundaries. A few annotated coding genes with evidence of both PA and NPA transcription were tested by reverse transcription with single primers in either orientation, followed by PCR. Only two out of 14 cases amplified a fragment corresponding to an antisense transcript (Supplemental Fig. 5F); thus antisense transcription is not likely to make up the bulk of non-

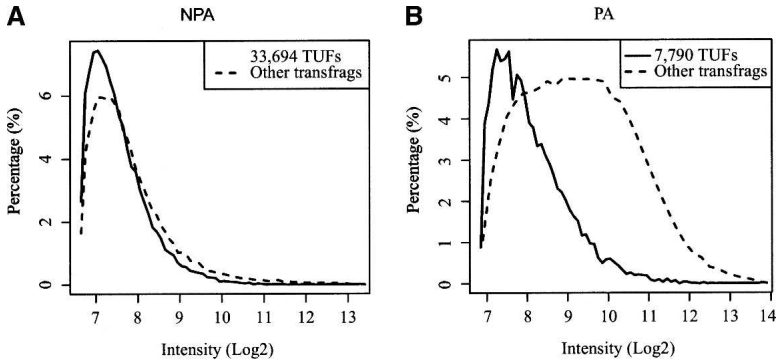

Figure 3. Signal intensity ($\log_2$) distribution for NPA TUFs and annotated transfrags (*A*) and PA TUFs and annotated transfrags (*B*).

polyadenylated transcription overlapping coding exons. Bimorphic transcripts (identical transcripts existing in both polyadenylated and non-polyadenylated form) have been indicated in the human transcriptome (Cheng et al. 2005), but our data cannot distinguish between this and other forms of transcriptional activity occurring at coding loci. Nonetheless, the strong overlap between polyadenylated and non-polyadenylated transcription in annotated protein-coding regions of the genome suggests that the transcriptional complexity in *C. elegans* is similar to that observed in other eukaryotes (Stolc et al. 2005; Engström et al. 2006)

Non-annotated polyadenylated TUFs are composed of novel exons and other transcripts

The PA data included 93,337 transfrags overlapping annotated exons and 11,925 without genomic annotation. To associate unannotated transfrags with known genes or transcripts (Manak et al. 2006), we supplemented the WormBase WS160 RefSeq annotations (v. 21) with 346,064 ESTs from GenBank. Clustering ESTs and RefSeq cDNA overlapping the TUFs (Supplemental Document 3) produced 1938 potential gene regions (PGRs) containing 3192 TUFs. Of these, 1340 TUFs appear to be additional or alternative exons of known annotated genes (Fig. 4), and the remaining TUFs may represent potential exons of unknown genes. An intriguing example of the latter is a PGR on the X chromosome containing 14 TUFs surrounding a locus annotated as noncoding transcript C53C7.5. The PGR lacks extended coding potential, contains an SL1 splicing recognition site, and is detected also by the NPA array, suggesting that this PGR may be a *trans*-spliced, bimorphic (Cheng et al. 2005) noncoding RNA gene (Fig. 5).

Among the 8733 TUFs in the PA sample that cannot be linked to RefSeq and EST data, 943 have gene prediction annotation and therefore have some protein-coding potential. This leaves 7790 TUFs with no additional information. The PA TUFs are generally short with a median (mean) size of 75 (87) bp, respectively, considerably shorter than most *C. elegans* exons. These TUFs have far lower signal intensities than most PA transfrags (Fig. 3B); nonetheless, RT-PCR analysis (Supplemental Document 1) confirmed 75% (18/24) of these, with no difference in confirmation rate between intronic and intergenic loci (see Supplemental materials for details). Further analysis by 5′- and 3′-RACE, cloning and sequencing of the 24 PA TUFs gave a positive 5′- and/or 3′-RACE fragment for six of these (Supplemental Table 1), three of which extended >30 nt beyond the TUF itself, possibly suggesting that also a fraction of the small PA TUFs may represent longer, lowly expressed transcripts.

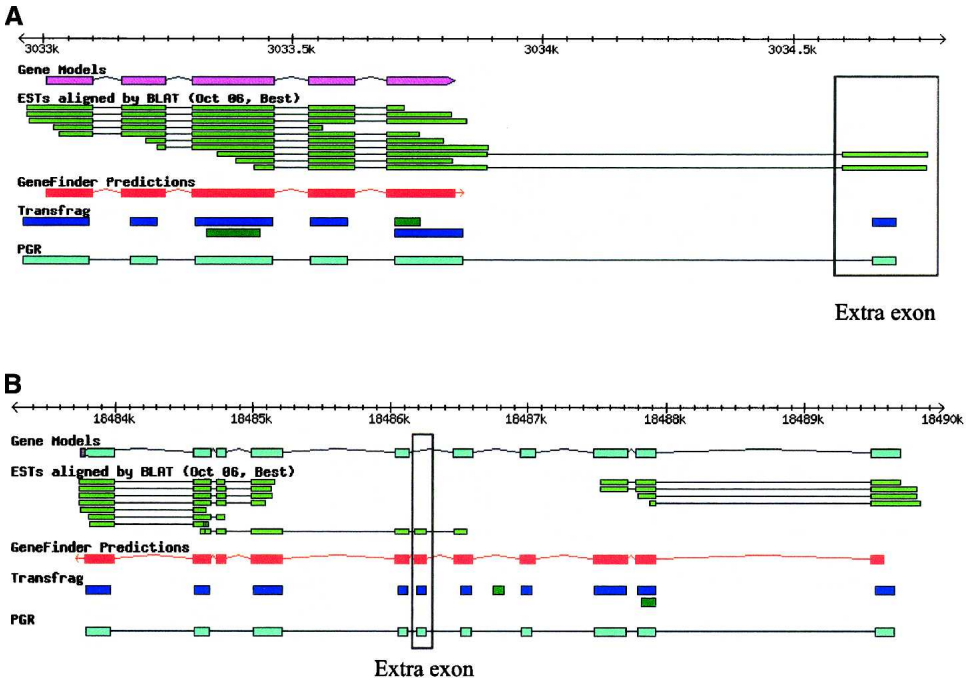

Figure 4. Assignment of additional exons to coding genes. (*A*) Potential 3′-end exon is detected ~700 bp downstream of the 3′-most annotated exon in gene Y54E10BR.2 on chromosome 1. (*B*) Coding gene Y51A2D.18 on chromosome V has a potential novel exon in intron 5.

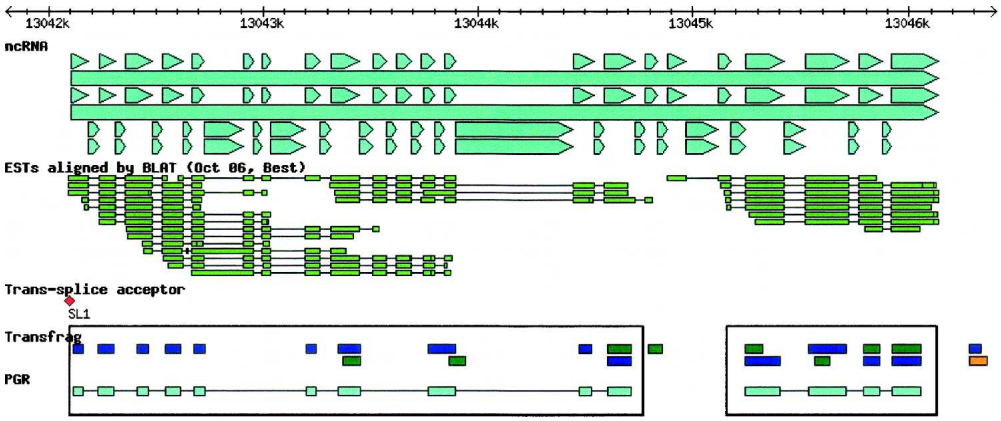

Figure 5. PGRs generated near a non-protein-coding region ("C53C7.5" in WormBase; Chen et al. 2005) on chromosome X. In the "Transfrag" track, blue, green, or orange boxes represent transfrags from the PA, NPA, or SNPA arrays, respectively. An SL1 site is annotated at the 5′ end of this region.

Discussion

Relative to its genome size, the transcriptional output in *C. elegans* appears no less complex than those of other eukaryotes subjected to full genome scans. The tiling microarray detected ~200,000 transcribed regions corresponding to 22.7% of the *C. elegans* genome. When transcribed introns of all annotated coding genes are included in this figure, an estimated 62.4% of the *C. elegans* genome could be transcribed. Including the possibility that 60% of the detected genes may also have additional transcripts (antisense, bimorphic, or other) would further increase the amount of transcriptional output per base pair genomic sequence to 70%. The very likely possibility that the intervening regions between non-annotated NPA TUFs are also transcribed might, however, further increase this figure.

This amount of transcription is comparable to what has been estimated for a number of other eukaryotes (Willingham and Gingeras 2006). There are nevertheless a number of differences that set *C. elegans* apart from other organisms. Most tiling microarray studies have found an amount of non-annotated polyadenylated transcription several times higher than that expected to arise from annotated genes. Cumulative transcription detected in eight human cell lines covered 10.5% of all interrogated nucleotides, which is four times the annotated 2.5% exonic sequence in the human genome (Cheng et al. 2005). In rice, 58% of the positive probes represented regions of the genome annotated as intergenic (Li et al. 2006), and in the 24 first hours of *Drosophila* embryo development, 30% of the polyadenylated transcription does not correspond to known exons (Manak et al. 2006). Even in yeast, where annotated genes constitute ~70% of the genome, ~20% of the polyadenylated transcription arise outside annotated exons (David et al. 2006). In comparison, only 11% of the detected *C. elegans* polyadenylated transcripts could not be referred to annotated loci.

Non-polyadenylated transcription has thus far only been studied by tiling microarray analysis in the human genome (Cheng et al. 2005). The amount of *C. elegans* non-polyadenylated transcription was 44% of the total observed transcriptional output on the array, comparable to the almost 50% reported for 10 human chromosomes (Cheng et al. 2005). Also similar to the human data, a major fraction (70%) of the non-polyadenylated transcription falls within the limits of coding loci (i.e., overlapping either exons or introns), the majority of this at least partially overlapping exonic sequence.

Our main aim with this study was to obtain an overview of the non-annotated (and potentially noncoding) elements of the *C. elegans* transcriptome, and in particular the complement of small noncoding RNAs. Computational predictions based on sequence conservation of potential secondary structure had indicated the presence of ~3600 such loci in the *C. elegans* and *C. briggsae* genomes (Missal et al. 2006). To explore this set of ncRNAs, we employed a preparation procedure that enriched the hybridized sample in small non-polyadenylated RNAs. Contrary to the expectations from the computational and other estimates (Deng et al. 2006; Missal et al. 2006), the *C. elegans* genome appears not to encode any larger number of small non-polyadenylated RNAs. Also, of the ~1200 novel SNPA TUFs, only 4.2% overlapped or fell within close reach of the computationally predicted sites, thus, neither DNA sequence conservation nor secondary-structure potential appear to have high predictive value when it comes to identifying novel ncRNA genes. We cannot exclude the possibility that RNA samples harvested from mixed-stage worm culture are not representative for the full small noncoding transcriptome, but as judged from the polyadenylated array data there does not appear to be any major fraction of the transcriptome that is not represented in a mixed-stage culture. Contrarily, in *C. elegans* the major bulk of the potentially noncoding RNAs seems to be either polyadenylated or in the form of longer non-polyadenylated transcripts.

Does the picture of the *C. elegans* transcriptome deviate from those obtained from other organisms studied by tiling arrays? In the sense that the observed non-annotated polyadenylated transcription is just 11% of all the PA detected transcripts would imply that it does, and had we not included non-polyadenylated data in our analysis the answer to Hillier et al. (2005) as to why such a small worm needs so many (coding) genes would have been that it is because it has so few other genes. The non-polyadenylated transcription data completes the picture in the sense that when it comes to the fraction of total transcriptional output, the worm is as rich in non-polyadenylated transcripts as is man. However, when taking into consideration that the human polyadenylated transcriptome is probably several times larger than the coding part of its genome, the worm falls short also in this respect. Thus, it may be that the worm has received a nearly full complement of protein coding genes, but when it

comes to participation in the new RNA world of regulatory complexity (Mattick 2004), its transcriptome betrays its organismal simplicity.

Methods

Sample preparation

RNA was extracted from mixed-stage wild-type N2 strain worms cultivated at 20°C according to the Trizol (Invitrogen) protocol. Small RNAs (<500 nt, SNPA sample) were isolated using a QIAGEN tip (QIAGEN), and the Poly(A)Purist MAG (Ambion) and MicrobExpress kits (Ambion) were adapted to remove remaining mRNAs and rRNAs (Deng et al. 2006). The enriched ncRNA pool was cloned using an adaptor-mediated library construction protocol. RNAs were dephosphorylated with calf intestine alkaline phosphatase (Fermentas) and then ligated to the 3′-adaptor (3AD) oligonucleotide by T4 RNA ligase (Fermentas) (He et al. 2006). Polyadenylated RNA (PA sample) was isolated from total RNA using the Poly(A)Purist MAG kit (Ambion). Non-polyadenylated RNA (NPA sample) was prepared by removing polyadenylated RNA using the Poly(A)Purist MAG kit and rRNA using the MicrobExpress kit (Supplemental Fig. 2). The PA and NPA RNA samples were reverse-transcribed (RT) using random hexamers, and the SNPA RNA sample was reverse-transcribed using a primer complementary to 3′-adaptor (oligo 3RT). First-strand cDNA was then used for second-strand DNA synthesis; the double-strand DNA was fractioned, labeled, and hybridized to the tiling array according to Affymetrix's GeneChip Whole Transcript (WT) Double-Stranded Target Assay Manual (http://www.affymetrix.com). The microarrays were scanned on a 3000 7G GeneChip Scanner. Hybridization of the PA, NPA, and SNPA samples were started from 140 µg, 140 µg, and 1 mg total RNA, respectively. Each prepared sample was hybridized once to the array, and the entire process of sample preparation and hybridization was carried out twice for each type of sample.

RT-PCR and 5′- and 3′-RACE

Total RNA digested with DNase I (Fermentas) was used as template for RT-PCR (QIAGEN OneStep RT-PCR kit). SNPA TUFs RACE were performed by PCR amplification of previously prepared small ncRNA cDNA library (Deng et al. 2006), with one primer designed specific for the ncRNA sequence and another primer being either 5CD or 3RT for 5′- or 3′-RACE, respectively. The PA and NPA TUF RACE reactions were carried out on the polyadenylated RNA and non-polyadenylated RNA fractions, respectively (see Supplemental Document 1).

Computational analyses

C. elegans genome annotation and sequence data and *C. briggsae* genome data were downloaded from WormBase (version WS140) (Harris et al. 2003). Raw data analysis and transfrag determination were performed as described by Kampa et al. (2004) with minor modifications (Supplemental Document 2). Briefly, the replicates are performed quantile-normalization and then scaled to the median intensity of 60. $\text{Log}_2[\max(PM - MM,1)]$ is calculated for each probe as an estimate of the expression level at each genomic position. The probes are considered significant over background if their signals are above a threshold associated with a false-positive rate of 4.6% estimated from the negative bacterial controls on the arrays. A transfrag is produced by the signal intensity threshold, a maximum gap between positive probe pairs (maxgap = 30), and a minimum length of the stretch-positive probe pairs (minrun = 13, at least two probe pairs). The analysis is implemented by the Affymetrix Tiling Analysis Software version 1.1.02.

All data underlying the study have been made available in the Supplemental material and on our server at http://bioinfo.ibp.ac.cn/tiling_array/.

Acknowledgments

We thank Yi Zhao, Yudong Wang, and Dandan He for early experiment discussion. The *C. elegans* strain N2 used in this work was provided by the *Caenorhabditis* Genetics Center, which is funded by the NIH National Center for Research Resources. This work was supported by the National Sciences Foundation of China (grant 30630040); National Key Basic Research & Development Program 973 (grants 2002CB713805 and 2003CB715900).

References

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., et al. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306:** 2242–2246.

Bertone, P., Trifonov, V., Rozowsky, J.S., Schubert, F., Emanuelsson, O., Karro, J., Kao, M.-Y., Snyder, M., and Gerstein, M. 2006. Design optimization methods for genomic DNA tiling arrays. *Genome Res.* **16:** 271–281.

Bracht, J., Hunter, S., Eachus, R., Weeks, P., and Pasquinelli, A.E. 2004. *Trans*-splicing and polyadenylation of *let-7* microRNA primary transcripts. *RNA* **10:** 1586–1594.

Chen, N., Harris, T.W., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., Bradnam, K., Canaran, P., Chan, J., Chen, C.-K., et al. 2005. WormBase: A comprehensive data resource for *Caenorhabditis* biology and genomics. *Nucleic Acids Res.* **33:** D383–D389. doi: 10.1093/nar/gki066.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308:** 1149–1154.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., and Steinmetz, L.M. 2006. A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci.* **103:** 5320–5325.

Deng, W., Zhu, X., Skogerbø, G., Zhao, Y., Fu, Z., Wang, Y., He, H., Cai, L., Sun, H., Liu, C., et al. 2006. Organization of the *Caenorhabditis elegans* small non-coding transcriptome: Genomic features, biogenesis, and expression. *Genome Res.* **16:** 20–29.

The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science* **306:** 636–640.

Engström, P.G., Suzuki, H., Ninomiya, N., Akalin, A., Sessa, L., Lavorgna, G., Brozzi, A., Luzi, L., Tan, S.L., Yang, L., et al. 2006. Complex loci in human and mouse genomes. *PLoS Genet.* **2:** e47. doi: 10.1371/journal.pgen.0020047.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34:** D140–D144. doi: 10.1093/nar/gkj112.

Harris, T.W., Lee, R., Schwarz, E., Bradnam, K., Lawson, D., Chen, W., Blasier, D., Kenny, E., Cunningham, F., Kishore, R., et al. 2003. WormBase: a cross-species database for comparative genomics. *Nucleic Acids Res.* **31:** 133–137.

He, H., Cai, L., Skogerbø, G., Deng, W., Liu, T., Zhu, X., Wang, Y., Jia, D., Zhang, Z., Tao, Y., et al. 2006. Profiling *Caenorhabditis elegans* non-coding RNA expression with a combined microarray. *Nucleic Acids Res.* **34:** 2976–2983. doi: 10.1093/nar/gkl371.

Hillier, L.W., Coulson, A., Murray, J.I., Bao, Z., Sulston, J.E., and Waterston, R.H. 2005. Genomics in *C. elegans*: So many genes, such a little worm. *Genome Res.* **15:** 1651–1660.

Jiang, M., Ryu, J., Kiraly, M., Duke, K., Reinke, V., and Kim, S.K. 2001. Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci.* **98:** 218–223.

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., et al. 2004. Novel RNAs identified from an in-depth analysis of the

transcriptome of human chromosomes 21 and 22. *Genome Res.* **14:** 331–342.
Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P., and Gingeras, T.R. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296:** 916–919.
Kent, W.J. and Zahler, A.M. 2000. Conservation, regulation, synteny, and introns in a large-scale *C. briggsae–C. elegans* genomic alignment. *Genome Res.* **10:** 1115–1125.
Li, L., Wang, X., Stolc, V., Li, X., Zhang, D., Su, N., Tongprasit, W., Li, S., Cheng, Z., Wang, J., et al. 2006. Genome-wide transcription analyses in rice using tiling microarrays. *Nat. Genet.* **38:** 124–129.
Liu, C., Bai, B., Skogerbo, G., Cai, L., Deng, W., Zhang, Y., Bu, D., Zhao, Y., and Chen, R. 2005. NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res.* **33:** D112–D115. doi: 10.1093/nar/gki041.
Lowe, T.M. and Eddy, S.R. 1997. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25:** 955–964.
Manak, J.R., Dike, S., Sementchenko, V., Kapranov, P., Biemar, F., Long, J., Cheng, J., Bell, I., Ghosh, S., Piccolboni, A., et al. 2006. Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. *Nat. Genet.* **38:** 1151–1158.
Mattick, J.S. 2004. RNA regulation: A new genetics? *Nat. Rev. Genet.* **5:** 316–323.
Missal, K., Zhu, X., Rose, D., Deng, W., Skogerbø, G., Chen, R., and Stadler, P.F. 2006. Prediction of structured non-coding RNAs in the genomes of the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *J. Exp. Zool. B Mol. Dev. Evol.* **306:** 379–392.
Pang, K.C., Stephen, S., Engström, P.G., Tajul-Arifin, K., Chen, W., Wahlestedt, C., Lenhard, B., Hayashizaki, Y., and Mattick, J.S. 2005. RNAdb—A comprehensive mammalian noncoding RNA database. *Nucleic Acids Res.* **33:** D125–D130. doi: 10.1093/nar/gki089.
Rinn, J.L., Euskirchen, G., Bertone, P., Martone, R., Luscombe, N.M., Hartman, S., Harrison, P.M., Nelson, F.K., Miller, P., Gerstein, M., et al. 2003. The transcriptional activity of human chromosome 22. *Genes & Dev.* **17:** 529–540.
Stolc, V., Samanta, M.P., Tongprasit, W., Sethi, H., Liang, S., Nelson, D.C., Hegeman, A., Nelson, C., Rancour, D., Bednarek, S., et al. 2005. Identification of transcribed sequences in *Arabidopsis thaliana* by using high-resolution genome tiling arrays. *Proc. Natl. Acad. Sci.* **102:** 4453–4458.
Stricklin, S.L., Griffiths-Jones, S., and Eddy, S.R. 2005. *C. elegans* noncoding RNA genes. In *WormBook* (ed. The *C. elegans* Research Community). http://www.wormbook.org/chapters/www_noncodingRNA/noncodingRNA.html.
Wang, J. and Kim, S.K. 2003. Global analysis of dauer gene expression in *Caenorhabditis elegans*. *Development* **130:** 1621–1634.
Willingham, A.T. and Gingeras, T.R. 2006. TUF love for "junk" DNA. *Cell* **125:** 1215–1220.
Yamada, K., Lim, J., Dale, J.M., Chen, H., Shinn, P., Palm, C.J., Southwick, A.M., Wu, H.C., Kim, C., Nguyen, M., et al. 2003. Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302:** 842–846.
Zemann, A., op de Bekke, A., Kiefmann, M., Brosius, J., and Schmitz, J. 2006. Evolution of small nucleolar RNAs in nematodes. *Nucleic Acids Res.* **34:** 2676–2685. doi: 10.1093/nar/gkl359.

Received April 13, 2007; accepted in revised form July 12, 2007.