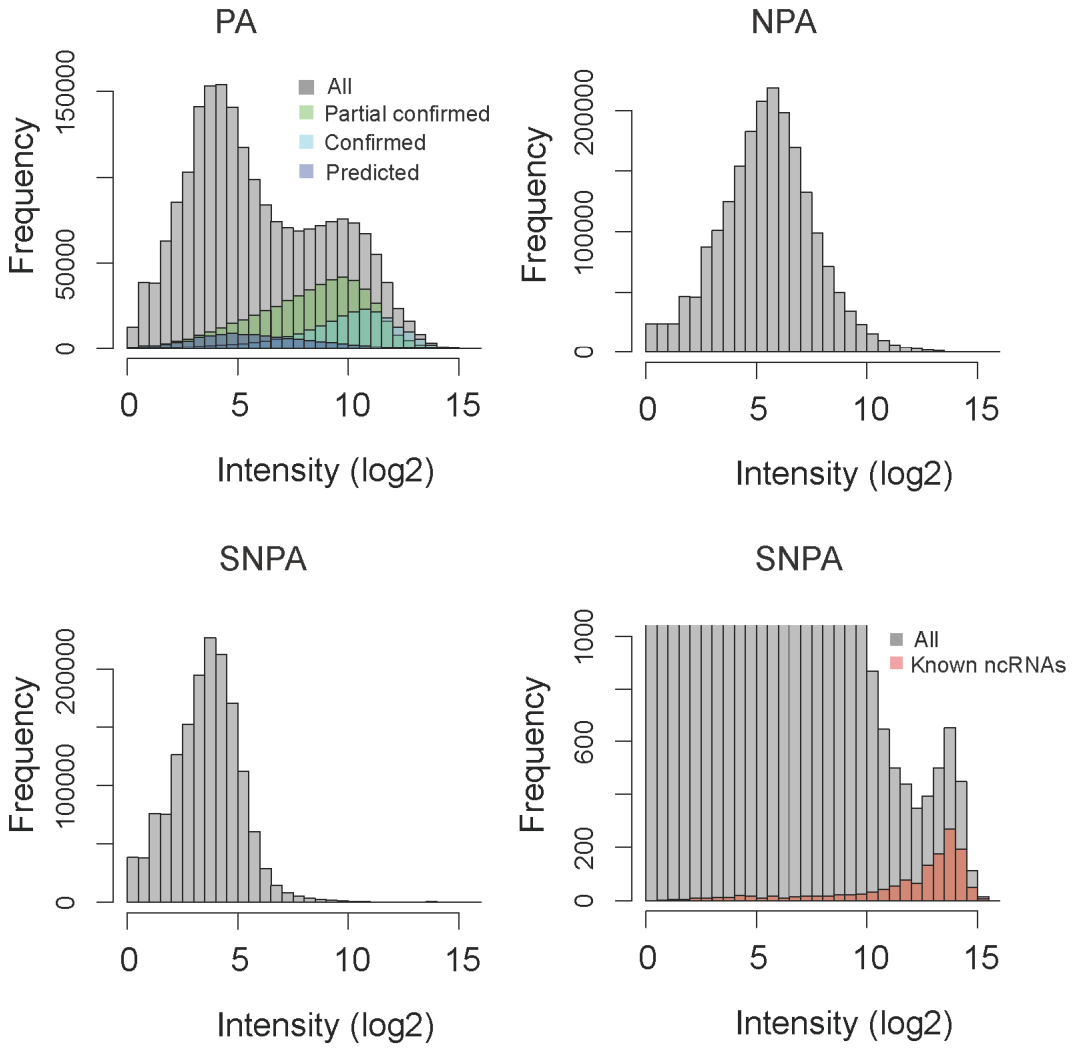
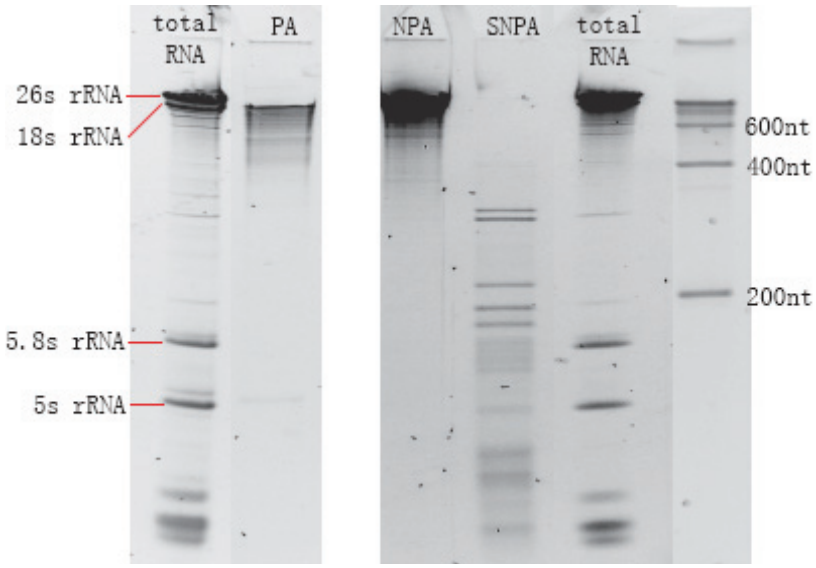
Supplemental Materials for

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



Supplemental Figure 1. Overall intensity (log2(PM-MM)) distribution of PA, NPA and SNPA.



Supplemental Figure 2. The RNA samples of PA, NPA and SNPA. The sizes of the SNPA sample range from about 50 to 500 nt, whereas the PA and NPA samples cover a wide range of sizes, mainly above 500 nt.

Supplemental Document 1

Experimental verification of TUFs by RT-PCR

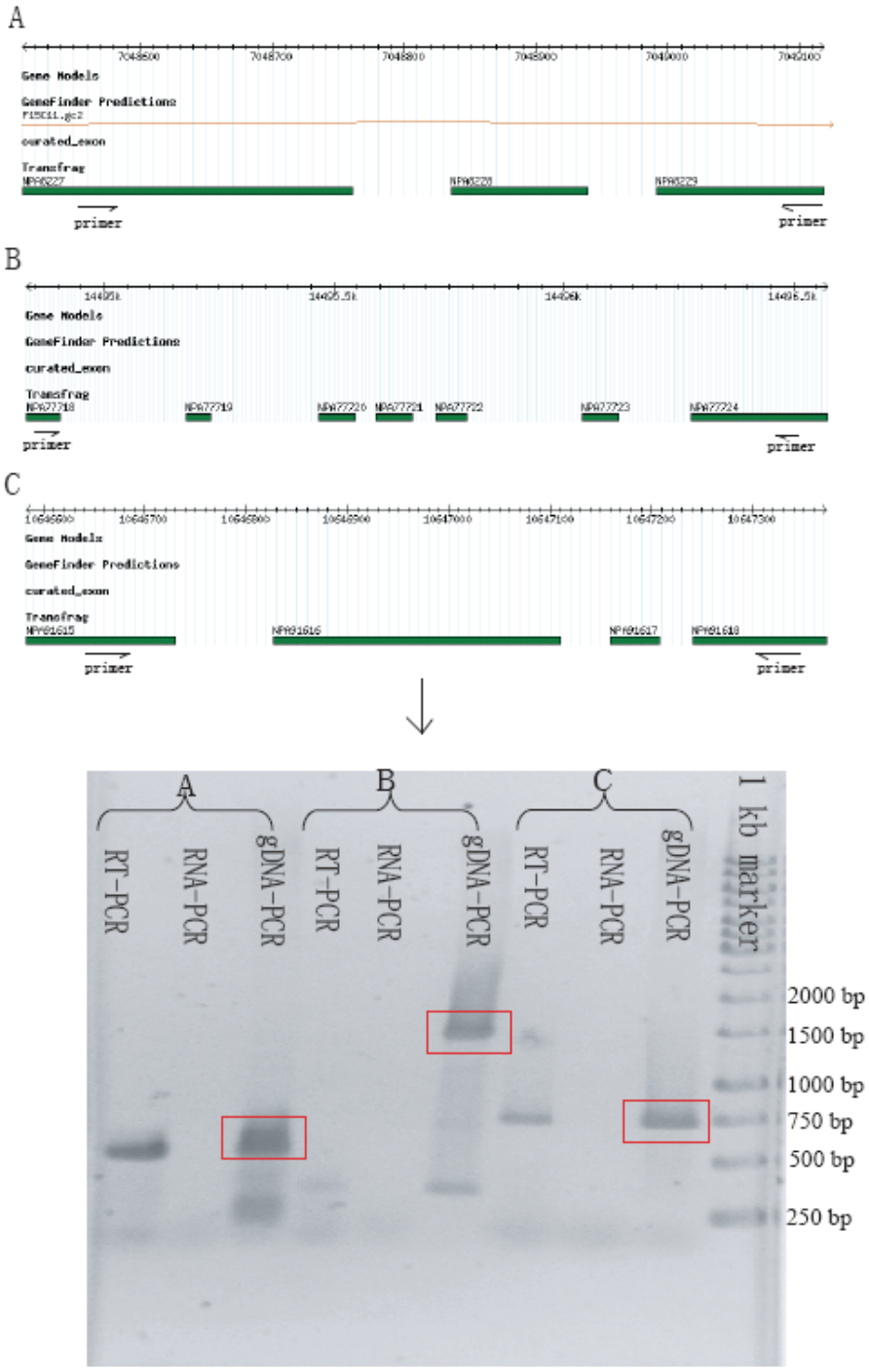
Eight negative controls were selected from regions with signal intensity lower than the lowest cutoff (6.1), and an additional 8 negative controls were selected from regions bridge exon-intron boundaries. From each of the SNPA, NPA and PA data 35 TUFs were randomly selected, and primers were successfully designed for 31, 29 and 24 of them, respectively. Total RNAs and the small ncRNA fraction were digested with DNAse I (Fermentas) at 37^{o}C for 1 hour and then purified by Trizol and chloroform. The purified total RNA fraction was then used as template for the RT-PCRs of the NPA and PA TUFs, and the small ncRNA fraction for the RT-PCR of the SNPA TUFs. A TUF was considered confirmed by RT-PCR in the following conditions:

1. The RT-PCR amplified fragment was at the expected length, which means the same length as amplified by PCR using genomic DNA (gDNA-PCR).
2. A PCR using the same RNA fraction as template ("RNA-PCR") did not amplify any fragment.

None of the 8 low signal intensity controls or the 8 exon-intron boundaries controls amplified RT-PCR fragments. Of the selected TUFs (above), 24 out of 31, 26 out of 29 and 18 out of 24 tested TUFs were confirmed according to these criteria, resulting in positive rates of 77%, 90% and 75% for the SNPA, NPA and PA samples, respectively.

An additional 8 pairs of NPA TUFs (16 TUFs) located in relative genomic vicinity were selected for RT-PCR. Individual RT-PCRs against these 16 TUFs all resulted in amplification

of a fragment of expected size, and when RT-PCRs were run with one primer in each TUF of a pair, five of the eight pairs amplified a fragment corresponding to the genomic distance between the priming sites, See figure below.



Supplemental Figure 3. *Test of adjacent NPA TUFs belonging to identical transcriptional units. Red boxes show the position of the gDNA-PCR bands. In all the paired TUFs (i.e. two TUFs located more than 500bp apart) A, B and C, the negative control (RNA-PCR) have not produced a band. In the paired TUFs A and C, the RT-PCR have bands the same size as the gDNA-PCR bands, suggesting that these polyA- transcripts in C. elegans are in the form of*

longer, unspliced RNAs. In the paired TUFs B, RT-PCR does not amplify the expected band (seen in the gDNA-PCR lane) and was counted as negative.

Experimental verification of antisense transcript

Fourteen transfrags annotated as coding genes with evidence of both PA and NPA transcription were tested by reverse transcription with single primers in either orientation, followed by PCR. Antisense transcription was considered confirmed by RT-PCR under the following conditions:

1. The RT-PCR amplified a fragment of the expected length, i.e. the same length as that amplified by PCR using genomic DNA (gDNA-PCR) as template.
2. PCR using total RNA or the small ncRNAs fraction (RNA-PCR) as template did not amplify any fragment.
3. Reverse transcription using each primer followed by PCR showed positive amplification of the expected fragment.

In only 2 out of 14 cases gave amplification of fragments corresponding to both sense and antisense transcripts.

Experimental verifications of TUFs by RACE

SNPA TUFs RACE was performed by PCR amplification of a previously prepared small ncRNA cDNA library (Deng et al. 2006), with one primer specific to the ncRNA sequence and the other primer complementary to either the 5' (5CD) or 3' (3RT) adapters, respectively. Twenty six TUFs from SNPA sample were selected for RACE analysis; 20 amplified fragments at the expected length range for both 5' and 3' RACE, and 4 have amplified fragments at the expected length range for either 5' or 3' RACE. Sequencing of 4 of the RACE fragments yielded the expected result in each of the 4 cases.

In the NPA and PA data, 57 TUFs (33 NPA and 24 PA) which had been tested by RT-PCR experiment were selected for 3' and 5' RACE experiments to determine the actual length of the transcripts. For each TUF, two additional primers (F1, R1, see figure below) were designed for nested PCR. We also designed two common primers, 3'-CDS primer (5'-GGCCACGCGTCGACTAGTAC(T)17VN-3' (V=A,C,G)) and UPS primer (5'-GGCCACGCGTCGACTAGTAC-3').

Supplemental Figure 4. In order to carry out the RACE experiment for PA and NPA TUFs, four primers were designed for each TUF.

Polyadenylated RNA and non-polyadenylated RNA extracted from total RNA were treated

with DNAse I for 30 minutes and purified with Trizol and phenol respectively. A number of standard PCRs (i.e. non-RT PCRs) were tested with these RNAs as a template (below), with none showing amplification, thus effectively ruling out DNA contamination as a source of the RACE/RT-PCR results reported below. Polyadenylated RNA was used for PA RACE, and non-polyadenylated RNA was used for NPA RACE.

3' RACE of PA TUFs

The 3' RACE reactions were performed on cDNA templates generated from an oligo-dT primer. First strand cDNA was synthesized using reverse transcriptase Superscript III (invitrogen) at 55 ^{o}C with the 3'-CDS primer (5'-GGCCACGCGTCGACTAGTAC(T)17VN-3' (V=A,C,G)). In each case, cDNA from 100ng total RNA, UPS primer (5'-GGCCACGCGTCGACTAGTAC-3') and primer F or R were used for the two first round PCRs, respectively, using the 1:100 diluted PCR product as template. The direction of the TUF was determined by second round PCRs with primer F and R using the diluted (1:100) first round PCRs as templates. Once the direction was determined, a nested PCR using the UPS primer and the corresponding nested primer (F1 or R1) was performed. The PCR product was analysed on PAGE or agarose gel, and candidate bands were recovered and cloned into the vector pGEM-T (Promega) and sequenced at Invitrogen, Beijing.

5' RACE of PA TUFs

The TUF specific primers F or R were used for reverse transcription and the cDNA was tailed with polyA using terminal DNA transferase (Fermentas). First round PCR was done with 3'CDS primer (5'-GGCCACGCGTCGACTAGTAC(T)17VN-3' (V=A,C,G)) and each of the TUF specific F and R primers. Nested PCR was then carried out using the diluted first round PCR product, the UPS primer (5'-GGCCACGCGTCGACTAGTAC-3') and corresponding nested primer F1 or R1. The second round PCR product was analysed on PAGE or agarose gel, and candidate bands were recovered and sequenced as above.

3' RACE of NPA TUFs

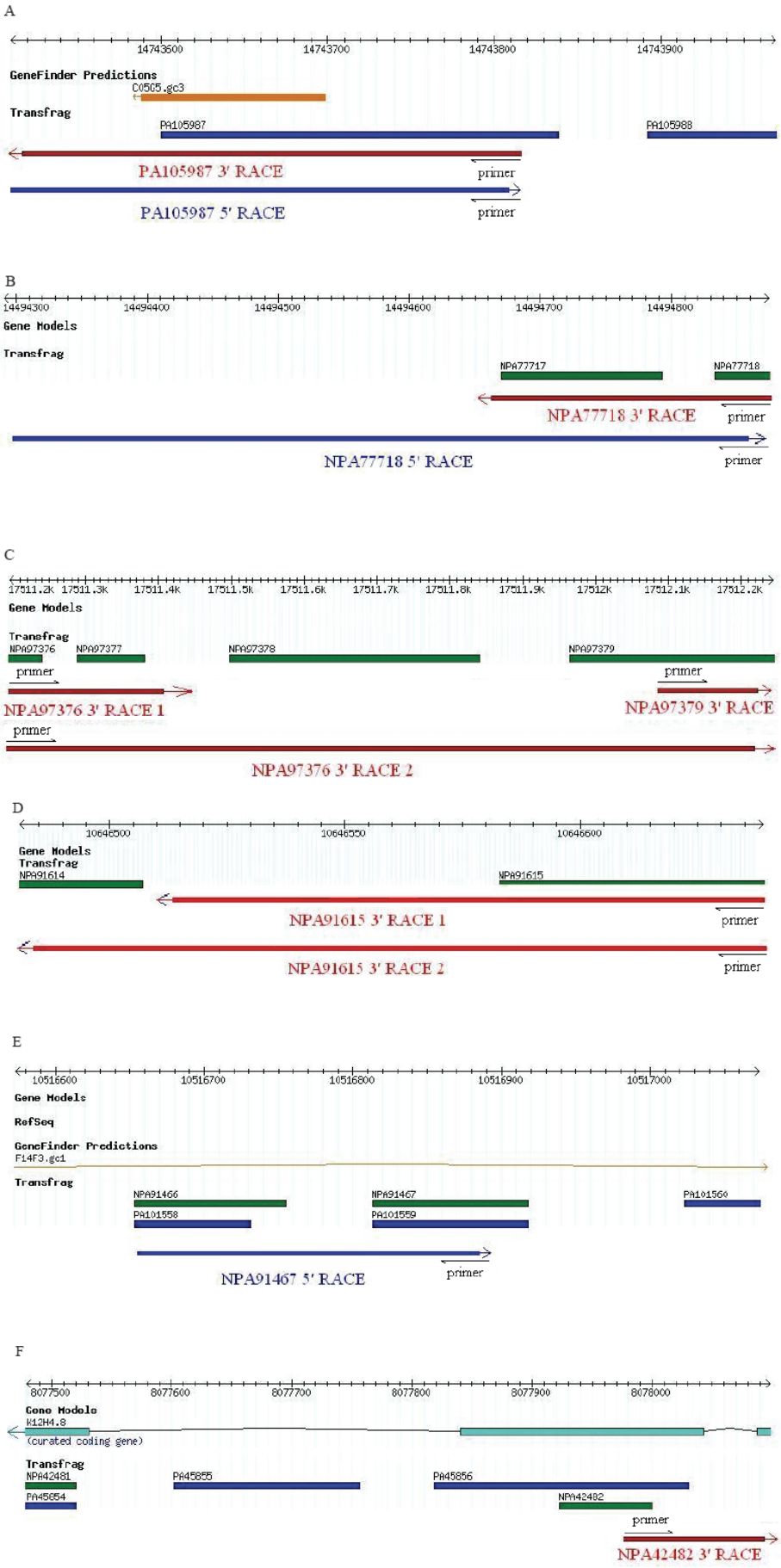
The non-polyadenylated RNA was ligated to a 3' adaptor 3AD as described for the SNPA in Method (see main text), with an additional 3 minutes at 94 oC before dephosphorylation and ligation. The ligated RNA was reverse transcribed into cDNA with a primer (3RT) complementary to the 3' adaptor 3AD. First round PCRs were performed by using primer 3RT and each of primers F or R. Direction of the TUF was determined by a second round PCR as above. Once the direction was determined, a nested PCR using 3RT and the corresponding nest primer F1/R1 was performed, and then the PCR products were analyzed, cloned and sequenced as above.

5' RACE of NPA TUFs

With the exception that non-polyadenylated RNA was used as template for the cDNA synthesis, 5’ RACE on NPA TUFs, was performed as for PA 5’ TUFs (above).

Table 1. All the RACE result.

TUFs	RACE name	TUFs location	RACE location	RACE length
NPA91615	NPA91615_3'RACE_1	X:10646583..10646729	X:10646638..10646520	119
NPA91615	NPA91615_3'RACE_2	X:10646583..10646729	X:10646638..10646481	158
NPA97376	NPA97376_3'RACE_1	X:17511160..17511238	X:17511219..17511456	238
NPA97376	NPA97376_3'RACE_2	X:17511160..17511238	X:17511219..17512243	1025
NPA1499	NPA1499_3'RACE_1	I:1315255..1315383	I:1315332..1315274	59
NPA1499	NPA1499_3'RACE_2	I:1315255..1315383	I:1315260..1315332	73
NPA97379	NPA97379_3'RACE	X:17511965..17512303	X:17512146..17512243	98
NPA13392	NPA13392_3' RACE	I:11203401..11203662	I:11203340..11203532	193
NPA77718	NPA77718_3'RACE	V:14494833..14494904	V:14494873.. 14494654	221
NPA50867	NPA50867_3'RACE	IV:1826657..1826830	IV:1826729..1826666	64
NPA71494	NPA71494_5'RACE	V:7953758..7953862	V:7953806..7953979	174
NPA78186	NPA78186_5'RACE	V:14914886..14914987	V:14914941..14914887	55
NPA77718	NPA77718_5'RACE	V:14494833..14494904	V:14494873..14494291	583
NPA18321	NPA18321_5'RACE	II:1201281..1201513	II:1201399..1201312	88
NPA91467	NPA91467_5'RACE	X:10516813..10516917	X:10516714..10516862	149
NPA42482	NPA42482_3'RACE	III:8077923..8077999	III:80777970..8078097	128
PA105987	PA105987_3'RACE	X:14743600..14743838	X:14743510..14743668	159
PA45669	PA45669_3'RACE	III:7953740..7953890	III:7953722..7953774	53
PA45855	PA45855_3'RACE	III:8077602..8077756	III:8077701..8078097	397
PA66946	PA66946_3'RACE	IV:12511704..12511807	IV:12511734..12511699	36
PA81061	PA81061_3'RACE	V:11054100..11054459	V:11054199..11054258	60
PA59552	PA59552_5'RACE	IV:6884458..6884580	IV:6884498..6884643	146
PA105987	PA105987_5'RACE	X:14743600..14743838	X:14743509..14743646	138


A
14743600
14743700
14743800
14743900
GeneFinder Predictions
C0505.gc3
Transfrag
PA105987
PA105988
PA105987 3' RACE
primer
PA105987 5' RACE
primer
B
14494300
14494400
14494500
14494600
14494700
14494800
Gene Models
Transfrag
NPA77717
NPA77718
NPA77718 3' RACE
primer
NPA77718 5' RACE
primer
C
17511.2K
17511.3K
17511.4K
17511.5K
17511.6K
17511.7K
17511.8K
17511.9K
17512K
17512.1K
17512.2K
Gene Models
Transfrag
NPA97376
NPA97377
NPA97378
NPA97379
primer
primer
NPA97376 3' RACE 1
NPA97379 3' RACE
primer
NPA97376 3' RACE 2
D
10646500
10646550
10646600
Gene Models
Transfrag
NPA91614
NPA91615
NPA91615 3' RACE 1
primer
NPA91615 3' RACE 2
primer
E
10516600
10516700
10516800
10516900
10517000
Gene Models
RefSeq
GeneFinder Predictions
F14F3.gc1
Transfrag
NPA91466
NPA91467
PA101560
PA101558
PA101559
NPA91467 5' RACE
primer
F
8077500
8077600
8077700
8077800
8077900
8078000
Gene Models
K12H4.8
(curated coding gene)
Transfrag
NPA42481
PA45855
PA45856
PA45854
NPA42482
primer
NPA42482 3' RACE

Supplemental Figure 5. Examples of the RACE results. In each case, arrows show the transcriptional direction inferred from the RACE result. "Primer" indicates the location and direction of the TUF specific primer used for RACE. In figure A and B, both 3' and 5' RACE fragments were obtained with same primer, indicating two oppositely oriented transcripts of similar length arising from the same locus. In figure C and D, 3' and 5' RACE from the same primer gave two fragment of different size, both at least twice as long as the TUF. Figure B, C, D and E show that RACE from one TUF extended to neighboring transfrags with no evidence of splicing. Figure F shows a RACE fragment antisense to the gene K12H4.8, spanning an intron between two adjacent exons in this gene.

Supplemental Document 2

The overall log2 signal distribution of the PA data is shown in supplementary Figure 2, and has an obvious bimodal intensity distribution. Applying a statistical model that rejects noise probes at a 90% confidence level[2], the intensity threshold for this rejection was determined to be 6.8, resulting in a false positive rate of 4.6% calculated from the negative control probe pairs designed against bacterial sequences. The bimodal intensity distribution of NPA and SNPA data is less obvious (supplementary Figure 2), and we therefore fixed the false positive rate at 4.6% to determine the intensity threshold for NPA (6.7), while for SNPA we use a more stringent criteria with a false positive of 2.5%, and the intensity threshold is 6.1.

Supplemental Document 3

Of the 108,669 PA transfrags, 93,337 overlap annotated exons, and 11,925 without genomic annotation. In an attempt to associate unannotated transfrags with known genes or transcript, we used gene annotations from wormbase WS140 along with its RefSeq cDNA (v. 21; 23,172 mRNAs and miscRNAs), and supplemented this with 346,064 ESTs from Genbank. These sequences were mapped to C. elegans genome with Blat[3] and used to identify which transfrags mapped to the same annotated gene, cDNA or EST transcript, Overlapping transcript with their corresponding transfrags were then clustered to produce potential gene regions (PGRs).

The process produced altogether 16,296 PGRs, of which 14,358 (12,781 +1,577) included only previously known transfrags (Table 2). Among the 11,925 TUFs there are 3,192 that have RefSeq cDNA (1,605) or EST (1,587) support. Of these were 1,340 included in PGRs corresponding to annotated genes, located either as additional 5' or 3' end exons (692), or as novel or alternative internal exons (648). An additional 416(245+171) PGRs were assembled from 956(785+171) TUFs using only RefSeq cDNA and EST data. The majority of these has received annotation in the WS160 version of Wormbase, and represents various coding genes. The 580(205+375) PGRs assembled from EST data alone have the greatest potential of representing novel spliced transcripts.

Table 2: PGR summary

		PGR	transfrag	TUFs
≥2 transfrags	all by known annotation	12,781	85,108	0
	by known annotation/EST/RefSeq	942	7,992	1,340
	by RefSeq cDNA	245	785	785
	by EST	205	521	521
single transfrag	by known annotation	1,577	1,577	0
	by RefSeq cDNA	171	171	171
	by EST	375	375	375
	total	**16,296**	**96,529**	**3,192**

References

1. **Deng, W. et al. Organization of the Caenorhabditis elegans small non-coding transcriptome: Genomic features, biogenesis, and expression. *Genome Res.* 16, 20-29 (2006).**
2. **Li, L. et al. Tiling microarray analysis of rice chromosome 10 to identify the transcriptome and relate its expression to chromosomal architecture. *Genome Biol* 6, R52 (2005).**
3. **Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res* 12, 656-64 (2002).**