Published online 23 August 2016

Nucleic Acids Research, 2016, Vol. 44, No. 21 ***e156***
doi: 10.1093/nar/gkw734

Analysis of *C. elegans* muscle transcriptome using trans-splicing-based RNA tagging (SRT)

Xiaopeng Ma[1,2,†], Ge Zhan[1,†], Monica C. Sleumer[1], Siyu Chen[1], Weihong Liu[1], Michael Q. Zhang[1,3,4] and Xiao Liu[1,*]

[1]MOE Key Laboratory of Bioinformatics, Center for Synthetic and Systems Biology, School of Life Sciences, Tsinghua University, Beijing 100084, China, [2]PTN (Peking University-Tsinghua University-National Institute of Biological Sciences) Joint Graduate Program, Beijing 100084, China, [3]Departmental of Biological Sciences, Center for Systems Biology, The University of Texas at Dallas, TX 75080, USA and [4]Division of Bioinformatics, TNLIST, School of Information Sciences, Tsinghua University, Beijing 100084, China

Received May 30, 2016; Revised August 03, 2016; Accepted August 11, 2016

ABSTRACT

Current approaches to profiling tissue-specific gene expression in *C. elegans* require delicate manipulation and are difficult under certain conditions, e.g. from dauer or aging worms. We have developed an easy and robust method for tissue-specific RNA-seq by taking advantage of the endogenous trans-splicing process. In this method, transgenic worms are generated in which a spliced leader (SL) RNA gene is fused with a sequence tag and driven by a tissue-specific promoter. Only in the tissue of interest, the tagged SL RNA gene is transcribed and then trans-spliced onto mRNAs. The tag allows enrichment and sequencing of mRNAs from that tissue only. As a proof of principle, we profiled the muscle transcriptome, which showed high coverage and efficient enrichment of muscle specific genes, with low background noise. To demonstrate the robustness of our method, we profiled muscle gene expression in dauer larvae and aging worms, revealing gene expression changes consistent with the physiology of these stages. The resulting muscle transcriptome also revealed 461 novel RNA transcripts, likely muscle-expressed long non-coding RNAs. In summary, the splicing-based RNA tagging (SRT) method provides a convenient and robust tool to profile trans-spliced genes and identify novel transcripts in a tissue-specific manner, with a low false positive rate.

INTRODUCTION

C. elegans is a widely used model for development, physiology and aging studies because of its invariable cell lineage, small body size and short life-span (1–3). As in other multicellular organisms, different tissues in *C. elegans* have various developmental programs and physical responses to stimuli, which involve tissue-specific gene expression. Thus, accurate transcriptome measurements of different tissues under various conditions are essential for systematic understanding of complicated biological processes (4).

Tissue-specific gene expression profiling has been very successful in *C. elegans* embryos because it is relatively easy to dissociate embryos for cell sorting (5). However, larvae and adults have a tough cuticle barrier preventing cell isolation. Recently, various methods have been developed to isolate cells or nuclei from postembryonic worms, revealing important insights in development and aging (6–10). Nevertheless, this strategy requires intact cell body or nuclei, which is challenging when worms have fragile cells or thickened cuticles. As a result, no tissue-specific profiling has been reported for aging or dauer worms. Another option is to isolate tissue-specific RNAs by selective expression of epitope-tagged polyA binding protein (PAB), which can bind mRNA tails in target tissues. Tissue-specific mRNAs can be purified using PAB-mediated RNA immunoprecipitation (RNA-IP) (11). This technique has been extensively utilized in many tissues of larvae and adult worms (11–16). However, RNA-IP requires the non-covalent interaction between mRNA and PAB to be fixed by formaldehyde cross-linking, which takes time to optimize the IP protocol and introduces significant background noise (12). Therefore, a more accurate covalent tagging method is desired for convenient and clean purification of mRNAs from a specific tissue.

Trans-splicing between the spliced leader (SL) RNA and pre-mRNA occurs for ∼70% of genes encoded in the *C. elegans* genome (17). As a result, mRNAs of these genes have SL sequences at their 5′ ends (17,18). SL RNAs consist of SL1, SL2 and SL2 variants. SL1-based trans-splicing occurs in 50–60% of *C. elegans* genes, while SL2 and its vari-

*To whom correspondence should be addressed. Tel: +86 10 62792304; Fax: +86 10 62788604; Email: xiaoliu@tsinghua.edu.cn
†These authors contributed equally to the paper as first authors.

ants are responsible for splitting poly-cistronic operon transcripts into individual mRNAs (19), occurring in 10–20% of genes.

The 22 bp 'exon' sequence of SL1 RNA is essential and highly conserved among nematodes (17). However, previous studies have found that some point mutations or permutations in the SL1 sequence are tolerable both *in vitro* and *in vivo*, including adding an extra sequence to the 5′ end (20,21). Furthermore, an SL1 transgene driven by the U2-3 promoter can rescue SL1 null mutants (20). These results suggest the possibility of engineering an SL1 sequence tag and driving its expression using a tissue-specific promoter to covalently tag mRNAs in tissues of interest via SL1-mediated trans-splicing.

Based on this novel idea, we developed a method called trans-Splicing based RNA Tagging (SRT) and profiled the gene expression of body muscle through driving the expression of tagged SL1 transgenes by a muscle-specific *myo-3* promoter. The resulting muscle transcriptome showed high coverage and specificity. We illustrated the robustness of our novel SRT method by applying it to dauer larvae and aging worms, which are refractory to cell isolation strategies. Due to its efficiency and accuracy, our profile also identified 461 novel transcripts, most of which are likely muscle-specific long non-coding RNAs (lncRNAs).

MATERIALS AND METHODS

Plasmid and worm constructs

The PU2-3::SL1 plasmid was a generous gift from the Rothman lab (20), in which a 341-bp U2-3 promoter is fused to a 224-bp fragment that contains the SL1 RNA gene followed by 120-nt of 3′ sequence to ensure proper 3′ end formation of SL1 RNA. We used a fragment of Illumina adapter (ACACGACGCTCTTCCGATCT) as the tag and introduced a mutation, G > C at position 16 or G > A at position 20 in the SL1 sequence by overlap extension PCR. The PU2-3::tag::SL1(G16C) fusion sequence was cloned into T-vector pEASY-Blunt^{TM} (TransGen). Similarly, a 1992 bp *myo-3* promoter and 3589 bp *rgef-1* promoter sequence was cloned from the *C. elegans* genome, fused to tag::SL1 (G16C) or tag::SL1(G20A) and cloned into T-vector pEASY-Blunt^{TM}. The full sequences of the PU2-3::tag::SL1 (G16C), P*myo-3*::tag::SL1(G16C), P*myo-3*::tag::SL1(G20A) and P*rgef-1*::tag::SL1(G16C) have been deposited to Addgene with accession numbers #79002, #79003, #81114, #81113, respectively. The linear polymerase chain reaction (PCR) product of the promoter::tag::SL1 (G16C) was microinjected into HT1593 (*unc-119, ed3*) worms by the standard method (22) using *cbr-unc-119* as a co-injection marker. Strains generated in this study were XIL1119 (thuEx119, *ed3*) containing PU2-3::tag::SL1 (G16C), XIL1115 (thuEx115, *ed3*) containing P*myo-3*::tag::SL1 (G16C), XIL1252 (thuEx252, *ed3*) containing P*myo-3*::tag::SL1(G20A) and XIL1253 (thuEX253, *ed3*) containing P*rgef-1*::tag::SL1(G16C).

Each promoter sequence upstream of a trans-splicing acceptor site (TSAS) of interest was PCR amplified from the *C. elegans* genome and cloned into pJIM20 vector, which contained *his-24::mCherry* and *cbr-unc-119* selection markers (23). The resulting plasmids were micro-injected into HT1593 (*unc-119, ed3*) worms. Information on primer sequences and transgenic strains is available in Supplementary File 5.

Worm culture

Worms were grown on NGM plates seeded with OP50 bacteria at 20°C using the standard culture method (24). Synchronized worms were obtained by treating with sodium hypochlorite, followed by embryo hatching on NGM plates without food. After overnight hatching, synchronized worms were considered as L1 larvae. About 12 h after feeding, worms were considered L2 larvae. About 2 days after feeding, most worms had a couple of eggs in their gonads and were considered as young adults. Synchronized L1, L2 and young adult worms were harvested to store in Trizol for RNA extraction. Aging worms were obtained by growing on Fluorodeoxyuridine containing NGM plates (400 μM) until the 12th day (young adult as day 1) as described previously (25).

The dauer worms were grown and purified according to the previous method (26). Briefly, transgenic worms were cultured in liquid at 20°C with vigorous shaking for 3–4 days. Then they were treated with 1% SDS for 15 min to kill non-dauers, followed by sucrose floatation and Ficoll precipitation to separate live worms from debris. The purity of dauer status was examined under stereo-microscope to make sure that more than 90% of worms were live dauers. The harvested dauers recovered overnight in 0.1 M NaCl at 20°C to eliminate stressful stimuli to worms during the tough treatment.

RNA isolation, amplification, library preparation and sequencing

Total RNA was extracted with the Trizol/phenol/isopropanol method as described (Invitrogen). A total of 300–600 μg RNA was pulled-down with a biotin-labelled probe (antisense to tag, biotin-AGATCGGAAGAGCGTCGTGT) with a PolyATract mRNA kit according to the manufacturer's instruction (Promega, #Z5310) except that the polyT probe in the kit was replaced by the anti-tag probe. All of the pulled-down RNA was then reverse-transcribed with pRT(ATTTAGGTGACACTATAGAAGCAGAAGA CGGCATACGAT(20)V) with PrimeScript^{TM} II Reverse Transcriptase (Takara, #2690A) at 42°C for 90 min to get full length cDNA. Then the cDNAs were PCR amplified for 10 cycles (98°C 10 s, 58°C 30 s, 72°C 1.5 min) with primer pF1 (ACACGACGCTCTTCCGATC) and pR1 (ATTTAGGTGACACTATAGAAGCAGAA) with Ex-Taq HS (takara, #RR006). To get higher specificity, nested PCR was performed (98°C 10 s, 58°C 30 s, 72°C 1.5 min) with pF2 (GACGCTCTTCCGATCTGGTT) and pR2 (CAAGCAGAAGACGGCATACGA) for another 10 cycles (Supplementary Figure S2A). The resulting cDNAs were fragmented and tagged with adapters with Tn5 transposase (Vazyme) at 55°C for 10 min, then amplified with pFY(AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCTGGTT) and pN7(CAAGCAGAAGACGGCAT ACGAGATIIIII

IIIGTCTCGTGGGCTCGG, 'I' represents index) with Q5 polymerase (NEB, #M0494) for 12–15 cycles (pre-PCR: 72°C 3min, PCR-cycle: 98°C 10 s, 66°C 30 s, 72°C 50 s) and purified with Ampure XP beads. PCR product was then sequenced on an Illumina HiSeq2000 using 2 × 100 bp or 2 × 125 bp mode.

Gene specific RT-PCR

Total RNAs from wild-type N2 strain were reverse-transcribed with polyT primer, pRT. In the validation of novel transcripts, first PCR was carried out using pR1 and gene-specific primer_F with Ex-Taq HS, followed by nested PCR using pR2 and gene-specific-primer_nested_F. The resulting products were sequenced by the Sanger method (Primer and transcript sequences are provided in Supplementary File 5).

Microscopy

Promoter::Cherry transgenic lines were imaged on a Zeiss imager A2 using 20X or 40X objective lens for embryos and larvae.

Detection of potential stage or condition specific endogenous SL1 trans-splicing

We analysed the RNA-seq data of 19 stages or conditions from modENCODE (18). For each pair of conditions at each TSAS, we produced a 2 × 2 contingency table containing the number of trans-spliced reads and the number of non-trans-spliced reads at that TSAS under the two conditions. We only analysed tables in which the number of non-trans-spliced reads was more than zero under at least one of the two conditions, and used the two-tailed Fisher exact test to determine the significance of the differences between the counts. In total, we performed 731 531 Fisher Exact tests, and applied a Bonferroni-adjusted *P*-value threshold of 0.05 for significance.

TSASs of modENCODE data

A global analysis of trans-splicing of modENCODE data was carried out by a previous study (18), and we directly downloaded its Supplementary File 1, which contained the TSAS locations and total amount of SL1-containing reads among different stages or conditions. The positions were converted to WormBase WS249 by the remap_gff_between_releases program that was downloaded from the Sanger Center website.

RNA-seq data analysis and TSASs annotation

Firstly, we filtered the reads that did not start with SL1(G16C), then trimmed adapter sequence with the program trimmomatic (27). We trimmed 22-bp SL1(G16C) sequence from each clean reads and aligned them to the *C. elegans* genome WS249 with tophat2 using WS249 gene model annotations (28), allowing 2 mismatches. A Python script was used to annotate the location of TSASs for both modENCODE data (18) and our data.

TSASs on exact or 100 bp upstream or downstream of the 5′-start site of an annotated exon were considered as annotated TSASs. Others were non-annotated, and were further classified into 3 categories: those in the intergenic region that were more than 1 kb upstream of known genes ('intergenic'); those on the antisense strand of known genes ('antisense'); and others ('undetermined'). In the calculations of gene expression, we considered only the reads on the first exon of a gene's transcripts (i.e. 'trans-splicing site') and ignored these on the 'cis-splicing site', similar to a previous study (18). Reads on cis-splicing sites were usually lowly expressed, representing potential different transcripts or trans-splicing inaccuracies (18). A total of 10 bps of sequence before the TSASs were extracted for each data set and submitted to WebLogo (weblogo.berkeley.edu) to generate trans-splicing motifs (29).

Detection of differentially expressed genes (DEG)

Read counts of each profile were normalized by reads per million mapped reads (RPM), then further normalized by upper-quantile normalization as previously described (30). DEGs were detected with the R package DESeq (31) using a negative binomial test, with the parameters of method = 'pooled' and sharingMode = 'fit-only'. Benjamini–Hochberg adjusted *P*-values were used.

The muscle expressed genes identified by reporter or immuno-staining assay were downloaded from WormBase (ftp://caltech.wormbase.org/pub/wormbase/expr_dump).

Enrichment analysis

Tissue enrichment predictions were analysed by the Tissue Expression Predictions for *C. elegans* program, version 1.0 (http://worm-tissue.princeton.edu/search/multi). DEGs were submitted to the DAVID website (version 6.7) for GO enrichment analysis with all trans-spliced genes as background (32,33). The Benjamini-corrected *P*-values < 0.01 were regarded as significant. In addition, the Gene Set Enrichment Analysis (GSEA) program with default parameters was used to test whether muscle marker genes were enriched in young adult muscle versus aging muscle, and larvae muscle versus dauer muscle.

Coding potential analysis

CPC scores were calculated by the Coding Potential Calculator (http://cpc.cbi.pku.edu.cn/). A Python script was used to calculate the distances from TSASs to the first downstream ATG or CAT.

RESULTS

The SRT strategy for tissue-specific gene expression profiling

In a transgene construct, we added a sequence tag (SRT-tag) to the 5′ end of the SL1 gene so that its transcript was a tagged SL1 RNA (Figure 1A). In practice, we used the Illumina adapter sequence as a SRT-tag to facilitate RNA-seq library construction. We also introduced a G to C point mutation at position 16 (G16C) into the SL1 sequence to

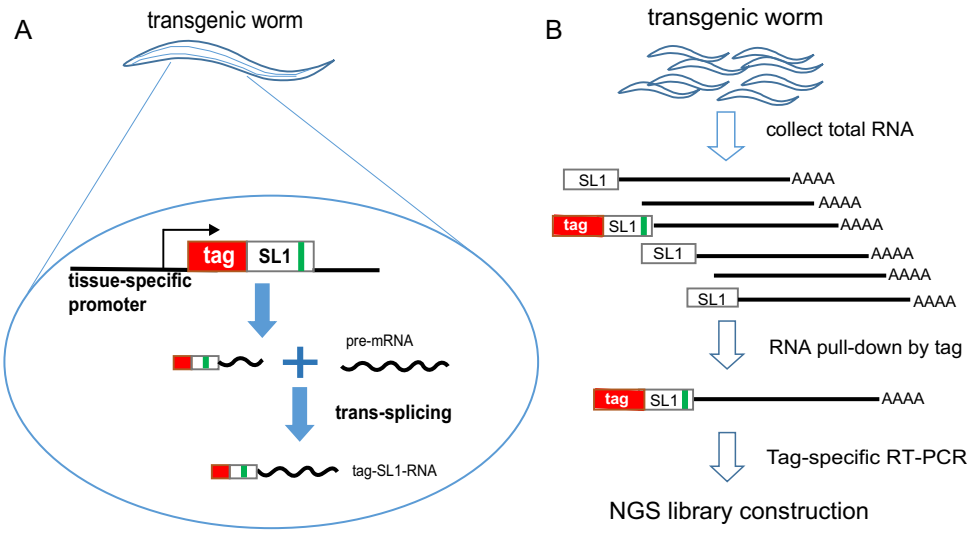

Figure 1. Schematic representation of the SRT method. (**A**) The worm strain carries a sequence-tagged SL1 transgene driven by a tissue-specific promoter. Only in this tissue, the transgenic tag::SL1 will express and trans-splice onto pre-mRNA, resulting in the tag::SL1 fused to the 5′ end of mRNAs. The green bar represents the G16C point mutation introduced to SL1 to distinguish from endogenous wild-type SL1. (**B**) Flowchart of sequencing library construction. After RNAs were purified from a large number of transgenic worms, those from the tissue of interest were pulled down using the tag sequence as bait. These enriched RNAs are reverse transcribed using polyT primer and polymerase chain reaction (PCR) amplified using the tag as a primer. The tag-based PCR further enriches mRNAs from the target tissue. PCR products are used to construct Illumina sequencing libraries (Supplementary Figure S2A shows more detail).

make it further distinguishable from endogenous SL1 (Supplementary Figure S1). The expression of this tagged SL1 was driven by a tissue-specific promoter so that only in the tissue of interest would the tag::SL1 RNA be expressed and trans-spliced onto mRNAs (Figure 1A). To obtain the transcriptome of the tissue, cell sorting or PAB-mediated RNA-IP was not required. Instead, total RNAs were collected from whole worm bodies and then tag::SL1 trans-spliced RNAs from the tissue of interest were enriched by tag-based RNA pulldown, reverse transcribed and amplified using the SRT-tag sequence as primer. Nested PCR was carried out to further enrich tag-SL1-containing cDNAs, followed by next generation sequencing (NGS) library construction (Figure 1B). The amplified cDNAs were fragmented and simultaneously ligated by two adapters (tn5-A1, tn5-A2) using Tn5 transposase. Then the fragments were amplified using primers annealing to the SRT-tag and tn5-A1 (or tn5-A2) (Supplementary Figure S2A). The resulting library was compatible with Illumina sequencing platforms. After sequencing, reads carrying the G16C mutation in the SL1 sequence represented 5′ ends of mRNAs from the tissue of interest.

One concern of this strategy is whether endogenous trans-splicing really occurred in a tissue-specific manner. Previous studies have identified tens of thousands of TSASs at the 5′ ends of annotated *C. elegans* genes (18). We compared their trans-splicing rates across 19 different conditions and stages (18). Only 369 (2.6%) TSASs showed significant differences in trans-splicing rates between at least one pair of stages or conditions (Bonferroni-adjusted P-value < 0.05, Fisher exact test). Furthermore, a very small fraction of these TSASs, 77 (0.54% of all TSASs), had significantly different rates of trans-splicing between at least one pair of developmental stages (early embryo, late embryo, L1, L2, L3, L4 and young adult) (Supplementary File 2). Because worms of different stages have different compositions of various tissues, these results suggested little tissue-specificity of trans-splicing. Next, to examine the consistency in abundance between full length mRNA of trans-spliced genes and their SL1 linked 5′ ends, we analysed modENCODE data (18) and observed strong correlation between the abundance of SL1-containing reads and those of whole genes (Pearson R = 0.72, Supplementary Figure S3), validating the usage of 5′ ends of mRNAs for the measurement of gene expression.

Validation and characterization of SRT profiling

It had been reported that SL1 RNAs with either a point mutation or extra sequence at the 5′ end were capable of trans-splicing (20,34). However, the trans-splicing activity of SL1 RNAs with both modifications remains unknown. We therefore generated transgenic worms carrying tag::SL1(G16C) that was driven by the promoter of universally expressed snRNA U2-3 or that of muscle-specific *myo-3*. These transgenic strains grew normally, indicating that tag::SL1(G16C) transgenes had no or negligible side effects on worms. We used *act-1*, the first identified trans-spliced gene in *C. elegans* (35), to test the trans-splicing activity of tag::SL1(G16C). As expected, *act-1* mRNA existed in all worms while transcripts of tagged *act-1* were detected only in tag::SL1(G16C) transgenic strains (Figure 2A). Sanger sequencing showed that tagged *act-1* mRNA contained the G16C mutation in its SL1 sequence (Figure 2B), validating the trans-splicing activity of tag::SL1(G16C) RNA.

Next, we used the P*myo-3*::tag::SL1(G16C) strain to validate the SRT method for gene expression profiling. Starting with a 0.5 ml pellet of young adults carrying the P*myo-3*::tag::SL1(G16C) transgene, we purified total RNAs and then reverse transcribed them using a polyT primer. We am-

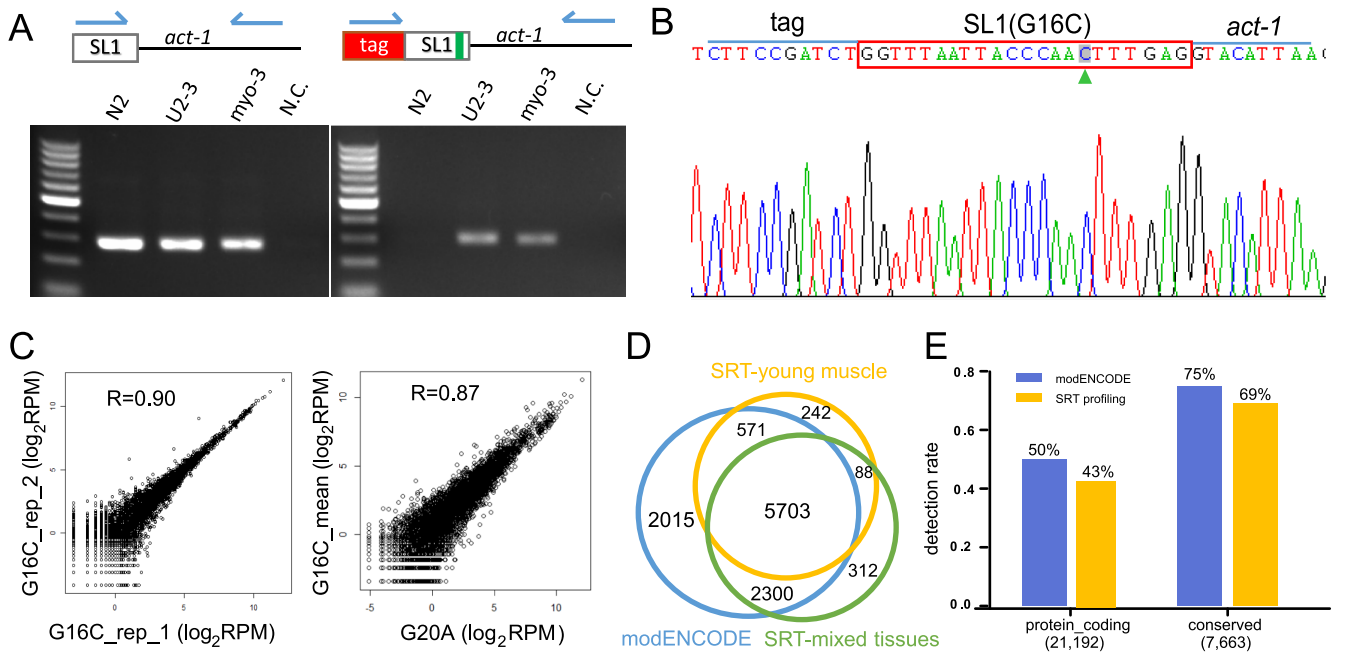

Figure 2. Validation and characterization of the SRT method. (**A**) *act-1* was trans-spliced by wild-type SL1 and tagged SL1(G16C) and detected by RT-PCR using SL1 or the tag as primer, respectively. N2: wild-type strain; U2-3 and *myo-3*: promoters driving the expression of tag::SL1(G16C); N.C. negative control, no PCR template. PCR products showed the expected sizes. Half arrows represent positions and orientation of primer sequences. The green bar represents the G16C point mutation introduced to SL1. (**B**) Sanger sequencing validated the existence of tag sequence and SL1 with the G16C mutation at the 5′ end of *act-1*. The green triangle represents the G16C point mutation. (**C**) Scatterplots of gene expression derived from SRT profiling of young adult muscle. Two biological replicates of P*myo-3*::tag::SL1(G16C) and one replicate of P*myo-3*::tag::SL1(G20A) showed high correlation. Pearson correlation coefficient is shown. (**D**) Gene coverage of SRT profiling. Overlap between the 10 589 mRNAs that were annotated as SL1-trans-spliced by modENCODE (18) and those detected by SRT profiling. (**E**) The fraction of SL1-trans-spliced protein-coding genes revealed by modENCODE studies and SRT profiling of young muscle and mixed tissues among all 21 192 protein-coding genes and 7663 conserved genes with human orthologs.

plified the resulting cDNAs by tag-specific nested PCR. RT-PCR products were used as substrate to construct NGS libraries (Supplementary Figure S2A). NGS generated about 5M clean reads. A total of 27.6% of these reads contained wild-type SL1 sequence. They were likely artefact derived from hybrid PCR (Supplementary Figure S2B) and discarded. The 72.4% of clean reads contained SL1(G16C), representing RNAs from body muscle (Supplementary Table S1). Expression profiles from two biological replicates had a strong correlation (R = 0.90) (Figure 2C). To further examine the reproducibility of the SRT profiling, we generated and profiled a strain carrying a transgene with a different point mutation in SL1, P*myo-3*::tag::SL1(G20A). Its transcriptome also had a strong correlation (R = 0.87) with that of P*myo-3*::tag::SL1(G16C) (Figure 2C), illustrating the robustness of the SRT profiling to different transgenes. Additionally, high correlation between biological repeats was observed in profiling of PU2-3::tag::SL1(G16C) transgenic worms (R = 0.97, Supplementary Figure S4A), further demonstrating the high reproducibility of the SRT method.

Reads from the P*myo-3*::tag::SL1(G16C) and PU2-3::tag::SL1(G16C) libraries were mapped to 6604 and 8403 annotated protein-coding genes, respectively (Supplementary File4 Sheet1). Their union contained 9216 genes, 93.0% of which are known to have TSAS according to a meta-analysis of modENCODE data (18). The reads covered 81.0% of known trans-spliced genes (Figure 2D). Genes whose trans-splicing was detected only by modENCODE or by our SRT profiling showed significantly lower expression than those detected by both (Supplementary Figure S5), indicating that highly expressed genes had more chance to be detected by RNA-seq and tended to be consistent among different studies. For example, mRNAs of ribosome proteins were most highly expressed genes in all libraries (Supplementary File 4 Sheet2). On the other hand, both the lowest expressed genes in the whole-body modENCODE data and trans-spliced genes not detected by our PU2-3-based SRT profiling were enriched in neuron-related GO terms (Supplementary Figure S6). The poor detection of neuron genes in whole body profiling might arise from the small fraction of neuron mRNAs in worm body extract due to their small cell volume. An alternative, but not exclusive, explanation for the different enrichment of various tissue genes is that expression of U2-3 is not totally homogenous. Indeed, the U2-3 reporter had undetectable expression in adults (Supplementary Figure S7). Nevertheless, the U2-3 reporter is active in most, if not all, somatic cells in embryos and early larvae. But it remains to investigate whether our protocol of preparing worms of mixed stages tended to enrich early-stage embryos, which contained fewer neurons than larvae and adults (36,37).

Only 50–60% of worm protein-coding genes are trans-spliced by SL1 (17,18), essentially the upper limit of the coverage of SRT profiling. Nevertheless, 75% of protein-coding genes conserved between worm and human are SL1-spliced

(18,38). Correspondingly, our SRT profiling detected 69% of the conserved genes (Figure 2E, Supplementary Table S4 Sheet1).

To cover more genes, we generated and profiled a P*myo-3*::tag::SL2(T5A) transgenic line. The SL2-based transcriptome detected 984 protein-coding genes, 92.4% of which were annotated as operon genes and 98.8% of which were found SL2-trans-spliced by modENCODE (Supplementary Figure S8A). Intriguingly, 98.2% of these 984 genes were also detected by our SL1 profiling (Supplementary Figure S8C), consistent with the broad dual-trans-splicing of SL2 acceptor sites revealed by modENCODE (Supplementary Figure S8D). Moreover, their expression levels in SL1- and SL2-based transcriptomes had significant correlation (Supplementary Figure S8B). At last, our SRT profile revealed trans-splicing on annotated cis-splicing acceptor sites of 5627 protein-coding genes. A total of 1819 of them were trans-spliced only on annotated cis-splicing sites, which provided a potential mechanism to increase the coverage of the SRT profiling (Supplementary Figure S9, Supplementary Table S1). However, it remains to investigate whether they resulted from trans-splicing onto cis-splicing sites of full-length mRNAs or from transcription of short mRNA isoforms driven by interval promoters in introns. So our further analysis did not include these cis-splicing acceptor sites.

Characterisation of body muscle gene expression in young adult worms

The *myo-3* promoter is active specifically in 95 body wall muscle cells, 16 sex muscle cells, four enteric muscle cells and five pairs of gonad sheath cells in a hermaphrodite (39). Therefore, our SRT RNA-seq data derived from P*myo-3*::tag::SL1(G16C) worms represented mostly the body muscle transcriptome. To evaluate the coverage of our muscle transcriptome, we searched WormBase and found 1144 muscle-expressed genes defined either by fluorescent reporter or immunostaining assay (Supplementary File 4 Sheet3). A total of 66.3% of them (758) were detected in our young adult muscle profile. If only trans-spliced genes were counted (984), the coverage increased to 77.0% (Figure 3A). Embryonic muscle gene expression had been profiled using cell-sorting-based RNA-seq (40), which detected 6837 muscle-expressed genes, including 5339 trans-spliced ones, 79.2% (4229) of which were detected by our SRT-derived muscle transcriptome (Figure 3A). Generally speaking, the more highly expressed a gene was in our SRT-based profile, the more likely it had been previously detected as muscle-expressed (Figure 3B). For example, 88.7% of the most highly expressed trans-spliced genes (RPM > 17.8) in our muscle SRT profiling were supported by previous evidence, while only 23.0% of trans-spliced genes undetected by our muscle SRT profiling were considered as muscle-expressed in previous studies (Figure 3B).

To evaluate the tissue specificity of the SRT profiling, we examined the expression in our SRT-derived muscle transcriptome of known marker genes of muscle (40,41), intestine (41,42) and hypodermis (43). These genes were responsible for tissue-specific function and mainly enriched in muscle, intestine and hypodermis, respectively. Nearly all 30 trans-spliced sarcomere-related muscle marker genes showed significant expression in both the PAB-RNA-IP-derived muscle profile (44) and the SRT-derived one. In the PAB-RNA-IP-derived muscle profile, many hypodermis or intestine marker genes showed significant expression, indicating the high background of the RNA-IP based method. On the contrary, in our SRT-derived muscle profile most hypodermal or intestinal genes showed undetectable or low level expression (Figure 3C), but there were some intestine or hypodermis makers showing significant expression. However, most of these were detected to be muscle-expressed by previous studies, suggesting they are not intestine- or hypodermis-exclusive (Figure 3C).

We identified 712 muscle-enriched genes by comparing the SRT-based muscle transcriptome from P*myo-3*::tag::SL1 transgenic worms with that of a mixed tissue transcriptome from PU2-3::tag::SL1 transgenic worms (Benjamini–Hochberg adjusted P-value < 0.001 and fold change > 8). Tissue Expression Predictions analysis (45) predicted that these genes were most enriched in muscle (Figure 3D). GO analysis of these genes showed an over-representation of muscle-related terms, such as muscle cell development, striated muscle cell differentiation, regulation of locomotion, myofibril assembly, as well as mitochondria and glycolysis, which are consistent with the physiological status of active body muscle (Figure 3E). In particular, the top 10 most enriched genes included 7 sarcomere related genes: *unc-87*, *pfn-3*, *tnt-2*, *mup-2*, *mlc-2*, *unc-27* and *dim-1* (Supplementary File 4 Sheet 5). These genes are involved in maintaining the structure of myofilament, muscle thin filament assembly, muscle contraction, muscle organization and troponin.

We then compared our muscle-enriched genes with previously reported ones identified by PAB-mediated RNA-IP (11) and by nuclear purification (41). Using the set of sarcomere-related muscle marker genes as a gold standard (40,41), ROC/AUC analysis showed that our SRT result was about as accurate as that of nuclear purification, and better than those derived from PAB-mediated RNA-IP (Figure 3F). A total of 50.1% of our muscle-enriched genes were not supported by any previous study (Figure 3G). Nevertheless, the promoters of these genes were significantly enriched in binding targets of transcription factor HLH-1, which is a master regulator of muscle development (46,47) (Figure 3G). Therefore, our SRT-based profiling likely identified novel muscle-specific genes.

Characterization of body muscle gene expression in aging and dauer worms

Our SRT method does not require any chemical or physical treatments of worms and therefore it can be applied to any worm status with little optimization needed. To test the robustness of our SRT method, we profiled muscle gene expression in aging worms using essentially the same protocol as in young adults. Two biological replicates had a correlation of 0.71, moderately weaker than that of young muscle (Supplementary Figure S4C).

Our SRT profiling detected 6005 annotated genes in muscle of day-12 worms. A total of 94.5% of these genes were also annotated as trans-spliced by modENCODE stud-

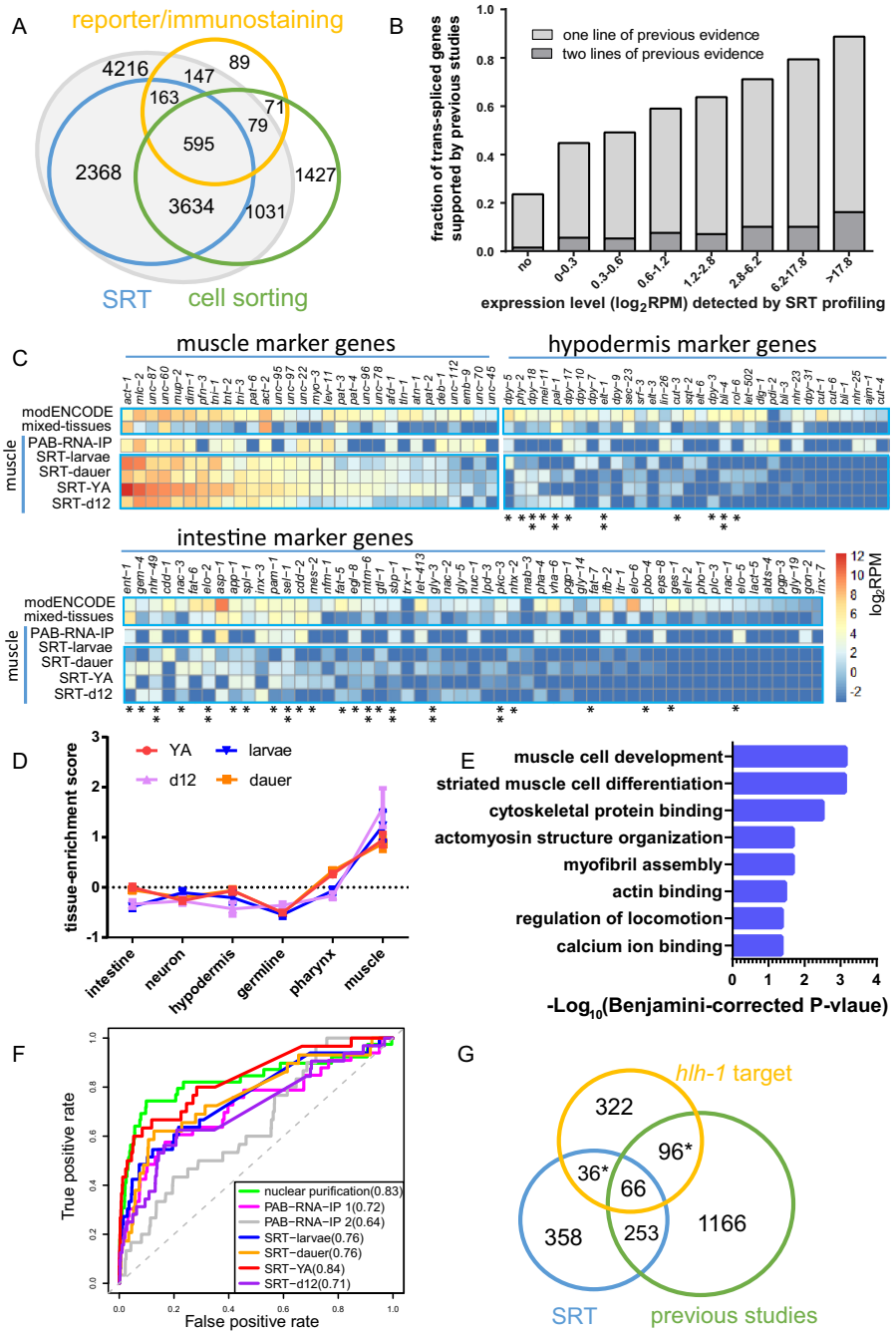

Figure 3. Characterization of SRT-derived muscle transcriptome. (**A**) Muscle-expressed genes detected by various methods. SRT: genes detected by our SRT profiling of muscle in young adults; Reporter/immunostaining: genes described as muscle-expressed in WormBase; cell sorting: genes detected by profiling of muscle cells isolated from embryos (40). Grey circle represents trans-spliced genes. (**B**) Correlation of expression levels in muscle measured by our SRT assay and probability of genes described as muscle-expressed by previous studies. Two lines of previous evidence: considered as muscle-expressed by both cell-sorting-based profiling and reporter/immunostaining assay; one line of previous evidence: considered as muscle-expressed either by cell-sorting-based profiling or by reporter/immunostaining assay, but not both. (**C**) The expression levels of the muscle, intestine and hypodermis marker genes detected in various muscle transcriptomes. modENCODE: mean level of whole body expression across 19 stages and conditions; mixed tissues: SRT-based expression profile derived from PU2-3::tag::SL1(G16C) transgenic worms; PAB-RNA-IP: RNA-seq data derived from PAB-mediated RNA-IP (44). According to WormBase, some intestine or hypodermis marker genes were also described as muscle, expressed by fluorescent reporter (**) or other muscle-specific high-throughput profile (*) (13). Data table is in Supplementary File 4 Sheet 4. (**D**) Tissue enrichment of our muscle profiles. Tissue enrichment score was predicted by the online program Tissue Expression Prediction for C.elegans (45). YA: young adult; d12: day-12; larvae: L1 to L2 stage larvae; dauer: dauer worms. Mean ± s.e.m. (**E**) Young-adult-muscle-enriched genes were over-represented on muscle related GO terms. (**F**) ROC curve of the prediction of known muscle marker genes by different muscle transcriptomes. Nuclear purification: RNA-seq of purified muscle nuclei (41). PAB RNA-IP1 and PAB RNA-IP2: microarray and RNA-seq data of PAB-mediated RNA-IP assays (11,44), respectively. AUC is shown in parentheses. G. Overlap of muscle-enriched genes revealed by different profiling methods. Previous studies included muscle enriched genes detected either by PAB-mediated RNA-IP (11) or by the nuclear purification approach (41). Asterisks indicate that the non-overlapped genes were significantly enriched with *hlh-1* target gene set. Fisher exact test, one-sided, P-value < 0.01.

ies (Supplementary Figure S10), even though they did not profile aging worms. This consistency corroborated previous observations that trans-splicing was largely stage-independent (17).

We identified 203 aging muscle-enriched genes by comparing the SRT-based transcriptome of old muscle with that of mixed tissues at mixed stages (Benjamini–Hochberg adjusted *P*-value < 0.001 and fold change > 8, Supplementary File 4 Sheet 6). GO analysis (Supplementary Table S2), tissue expression prediction (Figure 3D) and ROC/AUC analysis (Figure 3F) all suggested that these aging muscle-enriched genes had muscle specificity comparable to, although moderately weaker than, that of young muscle. Taken together, the profile of day-12 muscle retained muscle signatures, illustrating that our SRT method could conveniently generate tissue-specific gene expression profiles, even for worms at a status refractory to cellular manipulation.

Like mammals, old worms undergo sarcopenia, the decline of muscle structure and progressive locomotion impairment during aging (48). Comparing muscle profiles of young and day-12 adults, we identified 274 up-regulated genes and 149 down-regulated muscle-expressed genes during aging (*P*-value < 0.05 and fold change > 2, Supplementary Figure S11A, Supplementary File4 Sheet7). Several GO terms are significantly enriched among down-regulated genes, including larvae development, growth, translation, ribosome protein, glycolysis and mitochondrial matrix (Supplementary Figure S11E). Interestingly, these down-regulated genes included 4 sarcomere related genes: *act-1*, *tnt-2*, *unc-22* and *act-2*. GSEA also revealed the significant decline of sarcomere genes in aging muscle (Supplementary Figure S11B). For further characterization of tissue-specific aging, we utilized a pan-neuronal marker *rgef-1* to profile neuron gene expression. We generated and profiled a transgenic strain of P*rgef-1*::tag::SL1(G16C). Transcriptomes of both young and old neurons showed extensive gene coverage and strong neuron specificity (Supplementary Figure S12). But unlike in muscle, differentially expressed genes in neurons were not significantly related to any functional term (Supplementary Figure S11C–E). The molecular features revealed by our tissue-specific profiling during aging were consistent with previous observation of deteriorated muscle but intact neurons in old worms (49).

The decreased metabolism observed in aging muscle was reminiscent of the dauer stage, a long-lived hibernation status highly related to aging in molecular mechanisms (50). We profiled the gene expression of muscle in dauers using the same SRT protocol as above. Two biological replicates had a correlation of 0.75, similar to that of aging worms (Supplementary Figure S4B). This profile detected 7795 annotated genes, 93.4% of which were also defined as trans-spliced by modENCODE studies (Supplementary Figure S10). We identified 588 dauer-muscle-enriched genes by comparing the dauer muscle transcriptome with that of mixed tissues at mixed stages (Benjamini–Hochberg adjusted *P*-value < 0.001 and fold change > 8, Supplementary File 4 Sheet 8). GO analysis (Supplementary Table S2), tissue enrichment prediction (Figure 3D) and ROC/AUC analysis (Figure 3F) all suggested that these genes had muscle specificity comparable to those of young and aging muscle. Then, we profiled the gene expression of muscle in early larvae as a background, and identified 193 up-regulated and 138 down-regulated muscle-expressed genes in dauers (*P*-value < 0.05 and fold change > 2, Supplementary Figure S13A, Supplementary File 4 Sheet 9). GSEA analysis revealed that upregulated gene set in dauer muscle were significantly correlated with those measured from the whole body of dauer worm (26) (Supplementary Figure S13C). Similar correlation was observed between downregulated gene sets of dauer muscle and those of whole body dauer worms (Supplementary Figure S13D). But unlike the aging muscle, dauer muscle did not show decreased expression of sarcomere genes (Supplementary Figure S13B, Figure 3C). This molecular signature was consistent with a previous observation of intact muscle structure in dauer larvae despite the remodelling or shrinkage of other tissues (51).

Detection of muscle-enriched novel transcripts

The union of our four SRT-based muscle transcriptomes detected 29719 TSASs in the genome (Supplementary File3), 78.3% of which were located on annotated genes (Figure 4A, Supplementary Figure S14). These TSASs were enriched in the classical trans-splicing motif (TTTTCAG/R) at their upstream sequence (Figure 4B), indicating that transgenic tag::SL1(G16C) RNA followed the same principle of trans-splicing as endogenous SL1. A total of 21.7% of detected TSASs did not represent annotated genes. Nevertheless, these TSASs of unknown genes were still enriched in the canonical trans-splicing motif (Figure 4B, Supplementary Figure S15), suggesting these novel TSASs were authentic. A total of 29.5% of these non-annotated TSASs from muscle profiling were located at intergenic regions more than 1 kb upstream of protein-coding genes, while 37.8% were located within protein-coding genes, but at their antisense strand (Figure 4A). The high proportion of antisense orientation is reminiscent of non-coding RNAs, especially the lncRNAs (52). Consistent with previous report that trans-splicing could occur in lncRNAs (52), our SRT-based profiling detected 170 known non-coding RNAs (ncRNAs) in muscle, including 36 out of 170 previously defined long intergenic ncRNAs (52) and 16 out of 100 antisense ncRNAs (Supplementary Table S3). A total of 34.7% of these ncRNAs were also detected in modENCODE data (Figure 4C). Further validation by fluorescent reporter showed that the three annotated lincRNAs detected by our muscle profiling (linc-5, linc-55 and linc-57) were muscle-enriched (Supplementary Figure S16). All these results illustrated that our SRT-derived transcriptome contained useful information on ncRNAs.

The highly expressed non-annotated TSASs (RPM > 1) were likely to be real transcripts because they were enriched in the canonical trans-splicing motif to the same degree as the annotated ones (Supplementary Figure S15). Using the criteria of RPM > 1 and recurrent in both biological replicates, we identified 461 novel transcripts from our muscle transcriptome, 69 from that of mixed tissues and 415 from modENCODE studies (Figure 4D, Supplementary File 5). The expression levels of these novel transcripts were lower than protein-coding genes but slightly higher than known ncRNAs in all of these profiles (Figure 4E, Supplementary Figure S17), further suggesting that these novel tran-

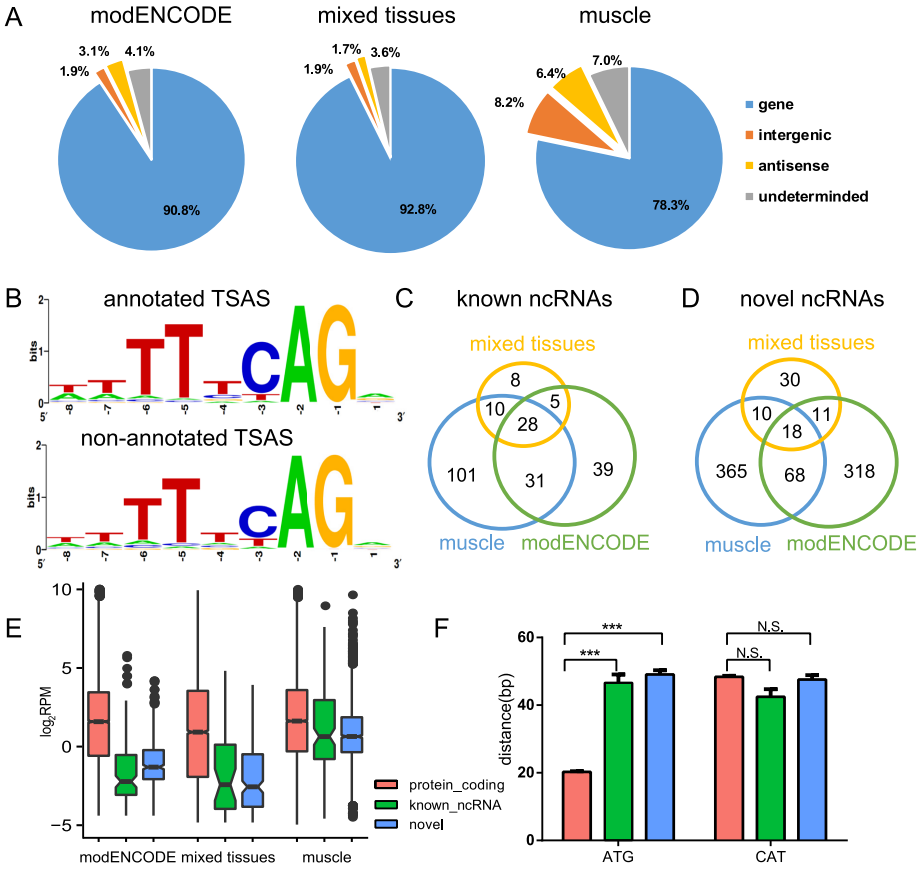

Figure 4. Novel transcripts identified by SRT profiling. (**A**) The composition of trans-splicing acceptor site (TSAS) classes detected by different transcriptomes. Mixed tissues: the expression profile derived from PU2-3::tag::SL1(G16C) transgenic worms. Muscle: the union of expression profile derived from P*myo-3*::tag::SL1(G16C) transgenic worms at various stages. Gene: TSASs located in an annotated protein-coding or ncRNA gene and at its sense strand; antisense: TSASs located at the antisense-strand of an annotated protein-coding or ncRNA gene; intergenic: TSASs located more than 1 kb away from annotated protein-coding or ncRNA genes. (**B**) Motif logos enriched in the annotated and non-annotated TSASs detected by SRT profiling. (**C** and **D**) Overlap of (C) known ncRNAs and (D) novel transcripts identified by various transcriptomes. Gene's names are in Supplementary File 4 Sheet 1 and File 5, respectively. (**E**) The expression levels of protein-coding genes, known non-coding (ncRNAs), and novel transcripts revealed by SRT profiling. (**F**) The distance from TSAS to first downstream ATG (left) or CAT (right) for protein-coding genes, known ncRNAs and novel transcripts. ***: significant difference (t-test, *P*-value $< 10^{-4}$); N.S.: not significant.

scripts were authentic, rather than transcriptional noise. To evaluate their coding potential, we calculated the distance from TSAS to the first downstream ATG in these novel transcripts. Regardless of their expression levels, TSASs of protein-coding genes tended to be significantly closer to their first ATG (median distance of 11 bp) than those of annotated ncRNAs (median distance of 39 bp) and those of these novel transcripts (median distance of 42 bp) (Figure 4F, Supplementary Figure S18). The difference in distance from the TSAS to the first ATG between coding and non-coding genes is unlikely because of A/T content because these two types of genes have similar distances from the TSAS to the first CAT. Therefore, the novel transcripts detected by our SRT assays were likely ncRNAs. In contrast to the protein-coding genes or known ncRNAs, 79.2% of these novel genes were not detected in whole-body RNA-seq data, including those from extensive modENCODE studies and those from our PU2-3::tag::SL1(G16C) profiling (Figure 4D), suggesting that a large fraction of novel TSASs discovered in this study represented muscle-specific novel ncRNA transcripts.

To further validate our prediction, we tried to obtain full-length sequences of 26 predicted novel RNAs, most of which were expressed strongly and uniformly in all stages of muscle but lowly or not-expressed in whole body profiles, by SRT based profiling (Figure 5A). All 26 were successfully cloned by RT-PCR (Figure 5B). Sanger sequencing revealed that they all had polyA tails, 11 of them (42%) had introns, 3 transcripts even had two isoforms (nc-1, nc-16 and nc-17) and 3 intergenic ones had previous EST evidence (nc-11, nc-13 and nc-17) (Supplementary Figure S19, Supplementary File 5). Their average length was 384 bp and all but two were longer than 200 bp. All the evidences strongly indicated that they represented real transcripts, excluding genome DNA contaminants or other artefacts. Furthermore, their coding potential scores were very low. Most (20 out of 26) transcripts were predicted non-coding based on a threshold used by a previous study (CPC < -1) (52) (Supplementary File 5). For example, there was a typical novel lncRNA (nc-6) detected by muscle profiling on the antisense strand of *glr-2* (Figure 5C). It contained a trans-splicing consensus sequence, one intron, a polyA tail and no ORF

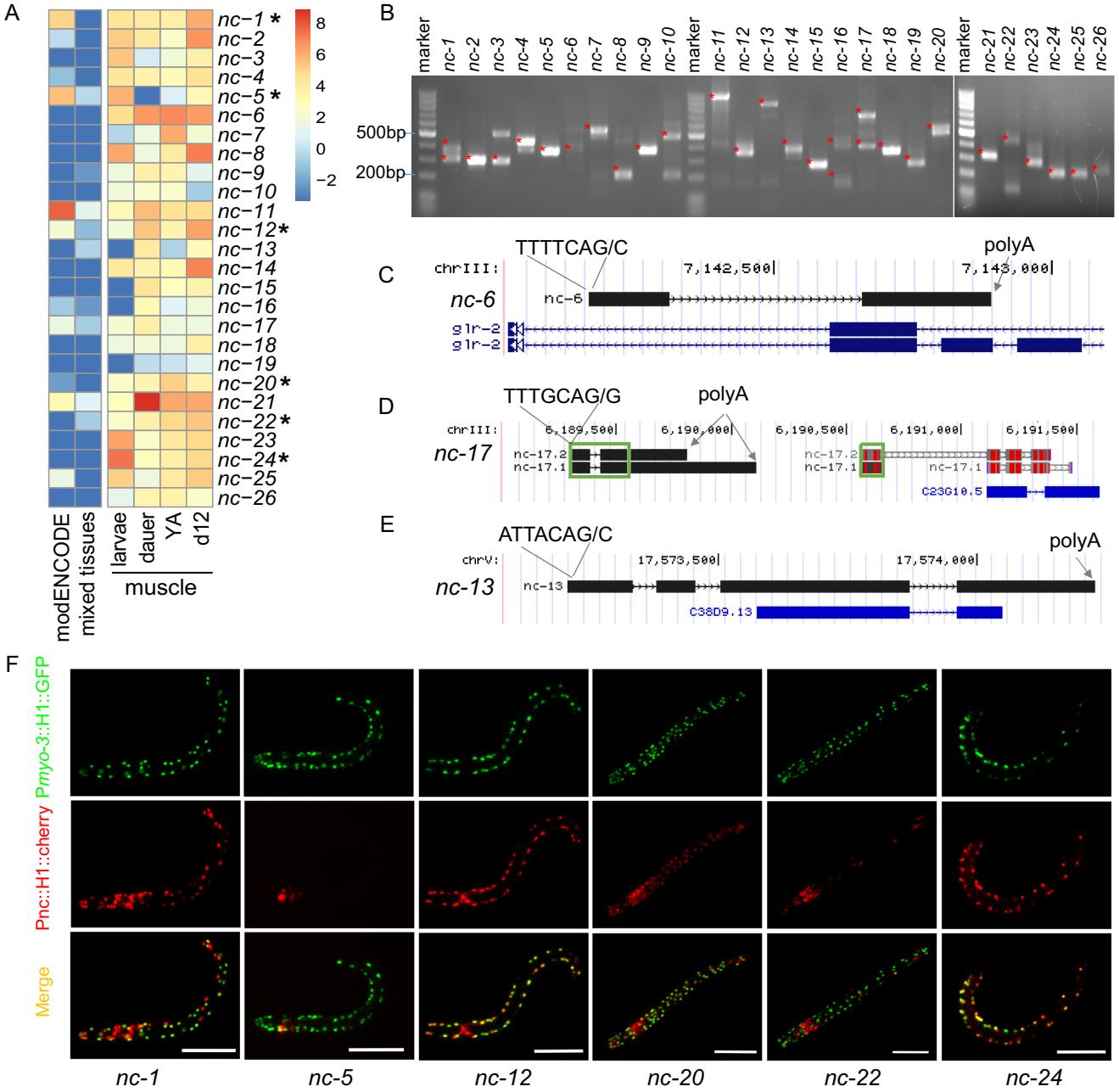

Figure 5. Experimental validation of novel transcripts revealed by SRT profiling. (**A**) Expression heatmap of 26 validated novel transcripts. Asterisks represent genes whose fluorescent reporters were constructed and analyzed. (**B**) A total of 26 novel transcripts validated by RT-PCR and cloning. Red asterisks represent RT-PCR products that were successfully cloned and sequenced. Two labeled bands in a lane represent different isoforms. (**C**) An example of novel transcript *nc-6* in muscle from the antisense strand of *glr-2*, with a trans-splicing consensus sequence, intron and polyA tail. (**D**) Novel transcript *nc-17* was a potential pseudogene, whose downstream gene is its paralog *C23G10.5*. BLAT alignment of nc-17 transcripts to *C23G10.5* was shown. Red bars represent sequence mismatches revealed by BLAT. Green dash line boxes represent the predicted ORF. (**E**) Novel transcript *nc-13* represented annotated gene *C38D9.13* with two extra exons. (**F**) Expression patterns of novel lncRNA genes revealed by their promoter reporters. The muscle cells were labeled by P*myo-3*::H1::GFP. Scale bar: 50 μm.

candidate. Five transcripts were predicted weak non-coding ($-1 < \text{CPC} < 0$), whereas only one novel gene (nc-17) had CPC score > 1. However, careful examination found that nc-17 was a paralog of its downstream protein-coding gene *C23G10.5*, but with dramatically deviated sequence (Figure 5D). The putative ORF of nc-17 was not homologous to that of *C23G10.5* (Figure 5D, Supplementary Figure S19) and no nc-17 homolog could be identified in other nematodes by BLAST. Furthermore, the intron of *C23G10.5* had homologous sequence in nc-17 transcripts, suggesting it was not spliced during maturation of nc-17 transcripts (Figure 5D). Therefore, nc-17 is likely an undefined pseudogene derived from duplication of protein-coding gene *C23G10.5*. Finally, novel transcript nc-13 turned out to be annotated

gene *C38D9.13* with two extra exons at its 5′ end (Figure 5E). Although *C38D9.13* was annotated as protein-coding, the new-exon-containing transcript (nc-13) had weak coding potential (CPC = −0.93), a long distance from the TSAS to an in-frame ATG (272 bp) and no homologue in other nematode species. These data indicated that nc-13 could be a lncRNA as well.

To further validate the existence of novel transcripts and reveal their expression patterns, we constructed reporter transgenic worms for 8 of these novel genes. Six of them showed detectable reporter signal. All except nc-5 were mainly expressed in muscle, consistent with our RNA-seq results (Figure 5A and F). In summary, the SRT method identified a large number of novel lncRNA candidates, which are otherwise hard to detect by whole worm profiling.

DISCUSSION

An increasing number of non-coding RNAs have been identified and deciphered as critical regulators of various biological processes. Many non-coding RNAs have special molecular biology activities such that several useful research tools have been developed based on them, e.g. miRNA/siRNA and CRISPR/gRNA (53). SL RNAs mediate trans-splicing in several metazoans so that they are covalently linked to the 5′ ends of mRNAs (17). A SL RNA can maintain its trans-splicing activity even if its sequence contains some variance and its expression is driven by heterogeneous promoters in previous and this study (20,34). Taking advantage of this property, we developed an approach called SRT to *in vivo* tag the 5′ end of mRNAs in a tissue of interest. Combining the tagging with RNA-seq, one can profile tissue-specific transcriptomes. SRT profiling is easy to manipulate, robust in different physiological status of worms and can accurately identify 5′-ends of even low level transcripts.

A widely used *in vivo* tagging approach is PAB-mediated RNA-IP (11), in which tagged-PAB non-covalently interacts with the polyA tail of mRNAs. The non-covalent interaction needs to be fixed by cross-linking before immunoprecipitation. Unfortunately, cross-linking usually introduces significant background noise (12). In stark contrast, our SRT approach is based on trans-splicing between tagged SL1 RNA and pre-mRNA. Resulting mRNAs have a tagged SL1 sequence at their 5′-ends such that the tag can serve both as a bait for RNA purification and as a primer for RNA-seq. Because of the cross-linking-free protocol, mRNAs from other tissues have little chance to contaminate RNA-seq data. In addition, our tag::SL1 transgene contains a G16C mutation so that its transcript can be distinguished from hybrid PCR products (Supplementary Figure S2B). As a result, the muscle transcriptome derived from our SRT profiling had much higher specificity than that derived from PAB-mediated RNA-IP method. The top-ranked genes were muscle-related and very few genes exclusively expressed in other tissues were actually present in our SRT-derived muscle transcriptome.

Although the tough cuticle covering the worm body is difficult to break, cell isolation protocols have been dramatically improved in recent years (7,9). One can even dissociate bodies of adult worms to purify GFP-expressing neurons using FACS (6,8). An alternative approach is to isolate nuclei from the worm body and purify those from the tissue of interest using FACS or immunoprecipitation (10,41). Although the strategy of cell or nucleus sorting is elegant and successful, significant effort is required to optimize their protocols for different tissues or conditions to minimize damage to cells and disturbance to gene expression. Therefore, methods that are robust to various cell statuses are still highly desirable. Compared to immunoprecipitation and cell sorting, RNA purification and sequencing of SRT are simple, straightforward and do not require complicated optimization for different tissues or stages. Most importantly, SRT does not require intact cells so that it can be applied to tissues refractory to cellular manipulation. Dauer is a stress-resistant developmental stage. It undergoes many morphological changes, including a specialized cuticle, closed oral orifices and stopped pharyngeal pumping to resist environmental insults, including 1% SDS treatment (54). It is therefore challenging to isolate specific tissues, given that tissue isolation depends on SDS to remove the tough outer cuticle (7,8). Similarly, old worms have some features that make it difficult to isolate their intact cells, such as an enlarged body size, fragile cells and a thickened cuticle (49). Consequently, no tissue-specific transcriptome had been reported for worms at these two stages even though these two interesting biological processes have been intensively studied (8,54,55). As a proof of principle, we used the SRT method to profile muscle gene expression in dauers and day-12 worms. These two transcriptomes showed significant muscle molecular signatures comparable to that of young adults. Furthermore, they were enriched in genes revealed by whole-body profiling of dauer and aging worms. These results illustrated the robustness of the SRT method to various physiological statuses of worm body. A previous study revealed that dauer and aging worms share some common expression changes and mechanisms (55). Nevertheless, the morphology and activity of body muscles are significantly different between them. In dauers, muscles have well-ordered sarcomeres, with enhanced muscle contraction in response to stimuli (54). On the contrary, body muscles dramatically deteriorate and shrink in old worms. During aging, muscle mass significantly shrinks and sarcomeres become progressively disorganized and losing their function (49). Consistent with these morphological and physiological observations, our SRT-derived muscle transcriptome revealed that the expression levels of sarcomere-related genes remained largely similar between dauers and normal larvae, while day-12 muscle showed significant down-regulation of sarcomere genes. These results illustrate the sensitivity of our SRT-based gene expression profiling and the importance of determining tissue-specific transcriptomes in investigating biological processes involved in various cell types and developmental stages.

The major limitation of the SRT method is its incomplete coverage because half of protein-coding genes are not trans-spliced by SL1 (17,18). Fortunately, conserved genes tend to be trans-spliced so that our SRT profiling detected 69% of genes that have human orthologues, indicating that the SRT profiling can be very helpful to reveal conserved molecular mechanisms. The combination of SL1- and SL2-based SRT could not improve the coverage because al-

most all SL2 acceptor sites can also be trans-spliced by SL1 (18). Our SRT profiling revealed some transcripts starting at cis-splicing acceptor sites, usually the non-first exons of genes, consistent with previous observation that trans-splicing also occurs on cis-splicing sites (18). So it may be possible for the SRT profiling to cover more genes by converting the cis-splicing sites into trans-splicing ones, e.g. in a U1 RNA mutant with downregulated cis-splicing activity. It is also intriguing to investigate whether the detected trans-splicing on cis-splicing accepter sites were due to transcription driven by internal promoters.

The reads derived from SRT profiling are strand-specific and concentrate on 5′-ends of transcripts. Furthermore, because the reads contain SL RNA sequences that occur only in transcripts, SRT-based RNA-seq is robust to genomic DNA contamination. SRT-based profiling therefore has a good dynamic range and be very sensitive to low abundance transcripts. Surprisingly, our SRT profiling not only covered most annotated genes, but also discovered a large number of novel genes. As the first sequenced metazoan genome and one of the most intensively investigated genetic and genomic models, 21192 protein-coding genes have been annotated in the *C. elegans* genome, comparable to those in the human genome (56). So, there are likely very few worm protein-coding genes left unannotated. Indeed, none of the 26 novel transcripts cloned and fully sequenced in this study represented a potential protein-coding gene.

Although *C. elegans* has one of the most comprehensively annotated genomes, less than 300 worm lncRNAs have been annotated. The ratio of lncRNA genes to protein-coding genes in *C. elegans* is far less than other animals (52,56). One plausible explanation is that tissue-specific gene expression has not been profiled in worm as much as in other model organisms. As a result, low abundance lncRNAs expressed in only a fraction of cells are difficult to detect by RNA-seq of whole worm bodies (56). Supporting this hypothesis, a transcriptome of worm sperm revealed and validated 8 novel lncRNAs that the whole-body gene expression profile generated by the modENCODE Consortium failed to identify (57). Similarly, our muscle transcriptome revealed 461 novel transcripts with high confidence. RT-PCR cloning validated all 26 novel transcripts, and reporter assays on six of these lncRNAs suggested that nearly all of them were muscle-enriched. Therefore, in addition to efficiently discovering novel lncRNAs, characterization of their expression provides a starting point to investigate their biological function.

In summary, SRT takes advantage of endogenous trans-splicing to profile tissue-specific gene expression. Despite its natural limitation of inaccessibility to non-trans-spliced genes, it is conceptually straightforward, easy to manipulate and essentially applicable to all developmental stages and physiological conditions. Additionally, it can accurately reveal 5′-ends of mature mRNAs and efficiently discover novel tissue-specific lncRNAs. We envision that SRT will provide a useful alternative to current methods of profiling tissue-specific gene expression in this classic model animal. It is noteworthy that SL-RNA-mediated trans-splicing occurs in the nematode phylum, trypanosomes, flatworms, hydra and even primitive chordates (17,58). It might therefore be possible for scientists to employ our SRT strategy for functional genomics studies in these species.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Jianhuo Fang, Jidong Lang and Lina Zhang from the Tsinghua Genomics & Synthetics Biology Core Facility for generating sequencing data. The Tsinghua University School of Information Science and Technology provided computational facilities and assistance with the data analysis. Dr Greg Vatcher edited the manuscript. The authors are very grateful to Stuart Kim, Kang Shen, Zhongying Zhao, Jack Chen, Yongbin Li for helpful discussions and comments on the manuscript.

FUNDING

Ministry of Science and Technology of China [20131970194, 2012CB316503]; National Natural Science Foundation of China [20141300429, 91519326]; Tsinghua 985 funding. Funding for open access charge: Ministry of Science and Technology of China [20131970194].
Conflict of interest statement. None declared.

REFERENCES

1. Jorgensen,E.M. and Mango,S.E. (2002) The art and design of genetic screens: Caenorhabditis elegans. *Nat. Rev. Genet.*, **3**, 356–369.
2. Fielenbach,N. and Antebi,A. (2008) C. elegans dauer formation and the molecular basis of plasticity. *Genes Dev.*, **22**, 2149–2165.
3. Gonzalez-Aguilera,C., Palladino,F. and Askjaer,P. (2014) C. elegans epigenetic regulation in development and aging. *Brief. Funct. Genomics*, **13**, 223–234.
4. Handley,A., Schauer,T., Ladurner,A.G. and Margulies,C.E. (2015) Designing cell-type-specific genome-wide experiments. *Mol. Cell*, **58**, 621–631.
5. Zhang,S. and Kuhn,J.R. (2013) Cell isolation and culture. *WormBook*, 1–39.
6. Wang,J., Kaletsky,R., Silva,M., Williams,A., Haas,L.A., Androwski,R.J., Landis,J.N., Patrick,C., Rashid,A., Santiago-Martinez,D. *et al.* (2015) Cell-Specific Transcriptional Profiling of Ciliated Sensory Neurons Reveals Regulators of Behavior and Extracellular Vesicle Biogenesis. *Curr. Biol.*, **25**, 3232–3238.
7. Zhang,S., Banerjee,D. and Kuhn,J.R. (2011) Isolation and culture of larval cells from C. elegans. *PLoS One*, **6**, e19505.
8. Kaletsky,R., Lakhina,V., Arey,R., Williams,A., Landis,J., Ashraf,J. and Murphy,C.T. (2016) The C. elegans adult neuronal IIS/FOXO transcriptome reveals adult phenotype regulators. *Nature*, **529**, 92–96.
9. Spencer,W.C., McWhirter,R., Miller,T., Strasbourger,P., Thompson,O., Hillier,L.W., Waterston,R.H. and Miller,D.M. 3rd (2014) Isolation of specific neurons from C. elegans larvae for gene expression profiling. *PLoS One*, **9**, e112102.
10. Haenni,S., Ji,Z., Hoque,M., Rust,N., Sharpe,H., Eberhard,R., Browne,C., Hengartner,M.O., Mellor,J., Tian,B. *et al.* (2012) Analysis of C. elegans intestinal gene expression and polyadenylation by fluorescence-activated nuclei sorting and 3′-end-seq. *Nucleic Acids Res.*, **40**, 6304–6318.
11. Roy,P.J., Stuart,J.M., Lund,J. and Kim,S.K. (2002) Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. *Nature*, **418**, 975–979.
12. Von Stetina,S.E., Watson,J.D., Fox,R.M., Olszewski,K.L., Spencer,W.C., Roy,P.J. and Miller,D.M. 3rd (2007) Cell-specific microarray profiling experiments reveal a comprehensive picture of gene expression in the C. elegans nervous system. *Genome Biol.*, **8**, R135.

13. Spencer,W.C., Zeller,G., Watson,J.D., Henz,S.R., Watkins,K.L., McWhirter,R.D., Petersen,S., Sreedharan,V.T., Widmer,C., Jo,J. *et al.* (2011) A spatial and temporal map of C. elegans gene expression. *Genome Res.*, **21**, 325–341.
14. Pauli,F., Liu,Y., Kim,Y.A., Chen,P.J. and Kim,S.K. (2006) Chromosomal clustering and GATA transcriptional regulation of intestine-expressed genes in C. elegans. *Development*, **133**, 287–295.
15. Takayama,J., Faumont,S., Kunitomo,H., Lockery,S.R. and Iino,Y. (2010) Single-cell transcriptional analysis of taste sensory neuron pair in Caenorhabditis elegans. *Nucleic Acids Res.*, **38**, 131–142.
16. Smith,C.J., Watson,J.D., Spencer,W.C., O'Brien,T., Cha,B., Albeg,A., Treinin,M. and Miller,D.M. 3rd (2010) Time-lapse imaging and cell-specific expression profiling reveal dynamic branching and molecular determinants of a multi-dendritic nociceptor in C. elegans. *Dev. Biol.*, **345**, 18–33.
17. Blumenthal,T. (2005) Trans-splicing and operons. *WormBook*, 1–9.
18. Allen,M.A., Hillier,L.W., Waterston,R.H. and Blumenthal,T. (2011) A global analysis of C. elegans trans-splicing. *Genome Res.*, **21**, 255–264.
19. Blumenthal,T., Evans,D., Link,C.D., Guffanti,A., Lawson,D., Thierry-Mieg,J., Thierry-Mieg,D., Chiu,W.L., Duke,K., Kiraly,M. *et al.* (2002) A global analysis of Caenorhabditis elegans operons. *Nature*, **417**, 851–854.
20. Ferguson,K.C. and Rothman,J.H. (1999) Alterations in the conserved SL1 trans-spliced leader of Caenorhabditis elegans demonstrate flexibility in length and sequence requirements in vivo. *Mol. Cell. Biol.*, **19**, 1892–1900.
21. Maroney,P.A., Hannon,G.J., Shambaugh,J.D. and Nilsen,T.W. (1991) Intramolecular base pairing between the nematode spliced leader and its 5′ splice site is not essential for trans-splicing in vitro. *EMBO J.*, **10**, 3869–3875.
22. Mello,C.C., Kramer,J.M., Stinchcomb,D. and Ambros,V. (1991) Efficient gene transfer in C.elegans: Extrachromosomal maintenance and integration of transforming sequences. *EMBO J.*, **10**, 3959–3970.
23. Murray,J.I., Bao,Z., Boyle,T.J., Boeck,M.E., Mericle,B.L., Nicholas,T.J., Zhao,Z., Sandel,M.J. and Waterston,R.H. (2008) Automated analysis of embryonic gene expression with cellular resolution in C. elegans. *Nat. Methods*, **5**, 703–709.
24. Brenner,S. (1974) The genetics of Caenorhabditis elegans. *Genetics*, **77**, 71–94.
25. Mitchell,D.H., Stiles,J.W., Santelli,J. and Sanadi,D.R. (1979) Synchronous growth and aging of Caenorhabditis elegans in the presence of fluorodeoxyuridine. *J. Gerontol.*, **34**, 28–36.
26. Sinha,A., Sommer,R.J. and Dieterich,C. (2012) Divergent gene expression in the conserved dauer stage of the nematodes Pristionchus pacificus and Caenorhabditis elegans. *BMC Genomics*, **13**, 254.
27. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
28. Kim,D., Pertea,G., Trapnell,C., Pimentel,H., Kelley,R. and Salzberg,S.L. (2013) TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.
29. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
30. Dillies,M.A., Rau,A., Aubert,J., Hennequet-Antier,C., Jeanmougin,M., Servant,N., Keime,C., Marot,G., Castel,D., Estelle,J. *et al.* (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.*, **14**, 671–683.
31. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
32. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
33. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
34. Xie,H. and Hirsh,D. (1998) In vivo function of mutated spliced leader RNAs in Caenorhabditis elegans. *PNAS*, **95**, 4235–4240.
35. Agabian,N. (1990) Trans splicing of nuclear pre-mRNAs. *Cell*, **61**, 1157–1160.
36. Sulston,J.E. and Horvitz,H.R. (1977) Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. *Dev. Biol.*, **56**, 110–156.
37. Sulston,J.E., Schierenberg,E., White,J.G. and Thomson,J.N. (1983) The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev. Biol.*, **100**, 64–119.
38. Shaye,D.D. and Greenwald,I. (2011) OrthoList: a compendium of C. elegans genes with human orthologs. *PLoS One*, **6**, e20085.
39. Ardizzi,J.P. and Epstein,H.F. (1987) Immunochemical localization of myosin heavy chain isoforms and paramyosin in developmentally and structurally diverse muscle cell types of the nematode Caenorhabditis elegans. *J. Cell Biol.*, **105**, 2763–2770.
40. Fox,R.M., Watson,J.D., Von Stetina,S.E., McDermott,J., Brodigan,T.M., Fukushige,T., Krause,M. and Miller,D.M. 3rd (2007) The embryonic muscle transcriptome of Caenorhabditis elegans. *Genome Biol.*, **8**, R188.
41. Steiner,F.A., Talbert,P.B., Kasinathan,S., Deal,R.B. and Henikoff,S. (2012) Cell-type-specific nuclei purification from whole animals for genome-wide expression and chromatin profiling. *Genome Res.*, **22**, 766–777.
42. McGhee,J.D. (2013) The Caenorhabditis elegans intestine. *Wiley Interdiscip. Rev. Dev. Biol.*, **2**, 347–367.
43. Page,A.P. and Johnstone,I.L. (2007) The cuticle. *WormBook*, 1–15.
44. Blazie,S.M., Babb,C., Wilky,H., Rawls,A., Park,J.G. and Mangone,M. (2015) Comparative RNA-Seq analysis reveals pervasive tissue-specific alternative polyadenylation in Caenorhabditis elegans intestine and muscles. *BMC Biol.*, **13**, 4.
45. Chikina,M.D., Huttenhower,C., Murphy,C.T. and Troyanskaya,O.G. (2009) Global prediction of tissue-specific gene expression and context-dependent gene networks in Caenorhabditis elegans. *PLoS Comput. Biol.*, **5**, e1000417.
46. Fukushige,T. and Krause,M. (2005) The myogenic potency of HLH-1 reveals wide-spread developmental plasticity in early C. elegans embryos. *Development*, **132**, 1795–1805.
47. Niu,W., Lu,Z.J., Zhong,M., Sarov,M., Murray,J.I., Brdlik,C.M., Janette,J., Chen,C., Alves,P., Preston,E. *et al.* (2011) Diverse transcription factor binding features revealed by genome-wide ChIP-seq in C. elegans. *Genome Res.*, **21**, 245–254.
48. Collins,J.J., Huang,C., Hughes,S. and Kornfeld,K. (2008) The measurement and analysis of age-related changes in Caenorhabditis elegans. *WormBook*, 1–21.
49. Herndon,L.A., Schmeissner,P.J., Dudaronek,J.M., Brown,P.A., Listner,K.M., Sakano,Y., Paupard,M.C., Hall,D.H. and Driscoll,M. (2002) Stochastic and genetic factors influence tissue-specific decline in ageing C. elegans. *Nature*, **419**, 808–814.
50. Tissenbaum,H.A. (2015) Using C. elegans for aging research. *Invertebr. Reprod. Dev.*, **59**, 59–63.
51. Wolkow,C.A. and Hall,D.H. (2013) The Dauer Muscle. *WormAtlas*, doi:10.3908/wormatlas.XXX.
52. Nam,J.W. and Bartel,D.P. (2012) Long noncoding RNAs in C. elegans. *Genome Res.*, **22**, 2529–2540.
53. Wright,A.V., Nunez,J.K. and Doudna,J.A. (2016) Biology and applications of CRISPR systems: Harnessing nature's toolbox for genome engineering. *Cell*, **164**, 29–44.
54. Hu,P.J. (2007) Dauer. *WormBook*, 1–19.
55. Lund,J., Tedesco,P., Duke,K., Wang,J., Kim,S.K. and Johnson,T.E. (2002) Transcriptional profile of aging in C. elegans. *Curr. Biol.*, **12**, 1566–1573.
56. Gerstein,M.B., Rozowsky,J., Yan,K.K., Wang,D., Cheng,C., Brown,J.B., Davis,C.A., Hillier,L., Sisu,C., Li,J.J. *et al.* (2014) Comparative analysis of the transcriptome across distant species. *Nature*, **512**, 445–448.
57. Ma,X., Zhu,Y., Li,C., Xue,P., Zhao,Y., Chen,S., Yang,F. and Miao,L. (2014) Characterisation of Caenorhabditis elegans sperm transcriptome and proteome. *BMC Genomics*, **15**, 168.
58. Hastings,K.E.M. (2005) SL trans-splicing: easy come or easy go? *Trends Genet.*, **21**, 240–247.