

## Large-scale RACE approach for proactive experimental definition of *C. elegans* ORFeome

Kourosh Salehi-Ashtiani, Chenwei Lin, Tong Hao, et al.

*Genome Res.* published online October 2, 2009

Access the most recent version at doi:[10.1101/gr.098640.109](https://doi.org/10.1101/gr.098640.109)

---

<b>Supplemental Material</b>	<a href="http://genome.cshlp.org/content/suppl/2009/10/02/gr.098640.109.DC1.html">http://genome.cshlp.org/content/suppl/2009/10/02/gr.098640.109.DC1.html</a>
<b>P&lt;P</b>	Published online October 2, 2009 in advance of the print journal.
<b>Open Access</b>	This manuscript is Open Access.
<b>Accepted Preprint</b>	Peer-reviewed and accepted for publication but not copyedited or typeset; preprint is likely to differ from the final, published version.
<b>Email alerting service</b>	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or <a href="#">click here</a>

---

---

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

---

To subscribe to *Genome Research* go to:  
<http://genome.cshlp.org/subscriptions>

---

# Large-scale RACE approach for proactive experimental definition of *C. elegans* ORFeome

Kourosh Salehi-Ashtiani<sup>1†§</sup>, Chenwei Lin<sup>1†</sup>, Tong Hao<sup>1†</sup>, Yun Shen<sup>1</sup>, David Szeto<sup>1</sup>, Xiping Yang<sup>1</sup>, Lila Ghamsari<sup>1</sup>, HanJoo Lee<sup>1</sup>, Changyu Fan<sup>1</sup>, Ryan R. Murray<sup>1</sup>, Stuart Milstein<sup>1,2</sup>, Nenad Svrzikapa<sup>1,2</sup>, Michael E. Cusick<sup>1</sup>, Frederick P. Roth<sup>3</sup>, David E. Hill<sup>1</sup> & Marc Vidal<sup>1§</sup>

<sup>1</sup>Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, and Department of Genetics, Harvard Medical School, Boston, MA 02115, USA.

<sup>2</sup> Current address: Alnylam Pharmaceuticals, 300 Third Street, Cambridge, MA 02142, USA.

<sup>3</sup>Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA.

<sup>†</sup>These authors contributed equally to this work

<sup>§</sup>Corresponding authors: KSA: kourosh\_salehi-ashtiani@dfci.harvard.edu ; MV: marc\_vidal@dfci.harvard.edu

## Abstract

Although a highly accurate sequence of the *C. elegans* genome has been available for ten years, the exact transcript structures of many of its protein coding genes remain unsettled. Approximately two-thirds of the ORFeome has been verified reactively by amplifying and cloning computationally predicted transcript models; still a full third of the ORFeome remains experimentally unverified. To fully identify the protein coding potential of the worm genome including transcripts that may not satisfy existing heuristics for gene prediction, we developed a computational and experimental platform adapting Rapid Amplification of cDNA Ends (RACE) for large-scale structural transcript annotation. We interrogated two thousand unverified protein coding genes using this platform. We obtained RACE data for approximately two-thirds of the examined transcripts and reconstructed ORF and transcript models for close to one thousand of these. We defined untranslated regions, identified new exons, and redefined previously annotated exons. Our results show that as much as 20% of the *C. elegans* genome may be incorrectly annotated. Many annotation errors could be corrected proactively with our large-scale RACE platform.

[5' and 3' RACE sequences are available at <http://www.wormbase.org> and <http://wormfdb.dfci.harvard.edu/index.php?page=race>. Processed RACE sequences, RACE derived annotations, and other relevant sequences and data are provided in the Supplemental files 2 - 9 accompanying this manuscript]

## Introduction

One of the goals of obtaining full genome sequences in the 1990s was to precisely identify the full complement of proteins, or proteome, used by a few model organisms and humans. The *C. elegans* genome was the first metazoan genome to be sequenced (The *C. elegans* Sequencing Consortium. 1998) and remains today the only complete and contiguous animal genome sequence ever established. The sequence is of excellent quality, with <1 mismatch per 30,000 nucleotides in the original release ten years ago and an error rate currently of <1 mismatch per 40,000 nucleotides (Hillier et al. 2005).

Despite this high-quality genome sequence, the complexity of gene transcription initiation and termination as well as differential splicing has made it extremely challenging to experimentally verify the exact sequence of the full complement of predicted protein-coding open reading frames (ORFs), or ORFeome (Walhout et al. 2000b). More than ten years after the first release of the genome sequence, the ORFeome remains imprecisely defined.

Our first attempt to experimentally verify the *C. elegans* ORFeome used PCR-based cloning of ORFs using the Gateway system (Walhout et al. 2000a) carried out from a high-quality cDNA library (Walhout et al. 2000b) using ORF-specific primers based on the WS9 release of WormPep (August 1999). In this first attempt we demonstrated that the number of protein-coding genes in *C. elegans* exceeds 17,300, a remarkably high number given that the *Drosophila* genome had been annotated with just 13,600 genes (Reboul et al. 2001). Our first genome-wide effort at cloning the *C. elegans* ORFeome experimentally verified approximately 55% of all ~19,000 predicted ORFs, with 4,000 ORFs that had remained strictly computationally defined until that point (Reboul et al. 2003). In a second effort, using re-designed ORF-specific PCR primers for 4,232 re-predicted ORFs based on WS100 release of WormPep (May 2003), we successfully cloned 1,378 re-predicted ORFs and 937 newly predicted ORFs (55%) (Lamesch et al. 2004). In both, the success rate was higher (63%) for ORFs supported by expressed sequence tags (ESTs) or other cDNA sequence evidence (“touched” ORFs), than for “untouched” ORFs lacking experimental support (42%). The remaining 45% of predicted ORFs could not be amplified even with redesigned ORF-specific primers, due most likely to mispredicted ORF boundaries (Reboul et al. 2001; Reboul et al. 2003; Lamesch et al. 2004). Because these verification experiments were done reactively, that is, they were limited to PCR amplification of the predicted computational models, the possibility of extended exon structure beyond the verified portion cannot be overlooked.

Other approaches have provided insight into worm transcript structure. These include: reconstruction of mRNA structure based on available EST and cDNA sequences (Thierry-Mieg and Thierry-Mieg 2006); TEC-RED (Hwang et al. 2004), which makes

use of trans-spliced leader to capture short reads from the 5' end of transcripts; and tiling arrays (He et al. 2007) to investigate which parts of the genome are transcribed. Two recent “next-generation” sequencing efforts of the *C. elegans* transcriptome (Shin et al. 2008; Hillier et al. 2009) provided information on transcribed genome regions and on splice junctions. These efforts have provided further insights into the transcriptional landscape and transcript structure of the worm. However, complete transcript structures, including *cis*-connectivity of exons, remain to be fully defined for all encoded genes.

As for any organism, a complete description of the *C. elegans* full complement of gene products—non-coding RNAs and proteins—including all variants obtained from alternative transcription and splicing is necessary for a comprehensive systems description of its biology. Experimental verifications of computationally derived structural transcript models suggest that current understanding of the rules of transcription initiation and termination, as well as splicing, remain approximate and incomplete. A proactive strategy that can define ORF and transcript structures without relying completely on computational predictions is urgently needed. Here we describe a large-scale RACE platform and demonstrate its value by applying it to ~2,000 unverified ORF models. Using existing ORF annotations as a point of departure, we defined ~1,000 alternative transcript and ORF models. With the strategy described here we are now in a position to proactively obtain models for entire ORFeomes for *C. elegans* and other organisms of interest.

## Results

### The RACE approach

Since almost half of predicted ORFs in *C. elegans* remain only partially supported by experimental data from cloning, new experimental strategies are needed to improve transcript and ORF annotation. RACE (Rapid Amplification of cDNA Ends) (Bao and Hull 1993) can proactively explore protein coding transcript models. Large-scale RACE has been hampered by low throughput, low specificity and sensitivity, and occasionally false capture of transcript ends. We adapted RACE for genome-wide studies in *C. elegans* by 1) improving throughput by combined use of Gateway cloning technology

(Walhout et al. 2000b) and minipool sequencing (Reboul et al. 2001; Reboul et al. 2003; Rual et al. 2004b); 2) increasing sensitivity and specificity by carrying out nested PCR, and 3) making use of *C. elegans* *trans*-spliced leader sequences to ensure capture of true 5' transcript ends (Fig. 1). About 70% of all *C. elegans* mRNAs have a *trans*-spliced leader sequence (Krause and Hirsh 1987; Blumenthal 2005); mostly the 22 base-long SL1 sequence (Conrad et al. 1995), with SL2 ranking as the next frequently used *trans*-spliced leader (Huang and Hirsh 1989; Blumenthal et al. 2002). For 5' RACE, the use of SL sequences, as opposed to the ligation of an arbitrary sequence to the 5' ends of transcripts in conventional RACE, provides two distinct advantages: no additional manipulation of RNA is needed, and the presence of SL1/SL2 ensures that the mRNA has an intact 5' end. To increase sensitivity and specificity for both 5' and 3' RACE, we carried out nested PCRs, with secondary primers matching to regions inside of the primary amplification regions. The secondary primers were tailed with Gateway sequences to permit recombinational cloning. Established protocols (Walhout et al. 2000b) were used for cloning, and for sequencing of products as minipools (Reboul et al. 2003). Sequencing from minipools, besides improving throughput, advantageously provides sequence information on the dominant transcript species, while still allowing detection of major alternatively spliced variants (showing up as discrete regions of 'mixed called' bases in the sequence traces) when they are present in the minipools.

To establish both a benchmark for RACE and an automated annotation pipeline, we gathered a positive control set (PCS) of 94 well-annotated protein coding transcripts, as well as an experimental reference set (ERS) of 94 transcripts whose ORF annotations had not been experimentally defined previously. The PCS consisted of transcripts for which the corresponding ORFs are: i) previously cloned, ii) shown to contain the SL1 leader sequence, and iii) known to have poly(A) sites. The ERS was picked randomly from transcripts for which either 1) ORF cloning was unsuccessful in the first two worm ORFeome project efforts (Reboul et al. 2003; Lamesch et al. 2004), or 2) cloning was successful but for which exon structure had been mispredicted such that actual coding region was cloned out of frame. Of the 94 transcripts of the ERS, 50 are to have an SL1

trans-splice acceptor based on WormBase WS150 annotation. The only experimental data supporting the remaining 44 transcripts were one or more ESTs.

Primer design and generation of RACE fragments were as described (Methods). We arranged the primers according to the sizes of the expected RACE fragments in each row. Examination of our PCS RACE products (Supplemental Figure 1) shows a general increase of size in each row, in line with the design of the primers. The ERS RACE products also show an increase in size in each row; however, consistent with being inaccurately modeled sometimes, the sizes are less regular than with the PCS (Supplemental Figure 1).

Both 5' and 3' RACE products were cloned as described (Methods). Following transformation, PCR amplimers were either generated from mini-pools made of multiple transformants, or from individual colony isolates. RACE sequence tags (RSTs) were obtained for each 5' and 3' RACE product (5' and 3' reads for the 5' RACE and only 5' reads for the 3' RACE). Each RST was initially aligned to its corresponding annotated CDS using `bl2seq`. We only retained RSTs with an alignment length greater than 100 bp and with high sequence quality (having 200 or more consecutive bases with PHRED  $\geq$  20) for further analysis.

When minipools of transformants were analyzed, we observed approximately 98% success rate (for the combined 5' and 3' reads) for our PCS, and 82% success rate (combined 5' and 3' reads) for the ERS. When individual colonies were sequenced, these numbers increased to 100% and 90%, respectively. For 3' RACE when mini-pools were analyzed, we observed approximately 86% success rate for the PCS and 77% success rate for the ERS. When individual colonies were sequenced these numbers increased to 95% and 86% respectively. The increase in success rate going from minipools to colony sequencing was contributed to by the reduction of: i) failed minipool reads that were redone as multiple single colonies, and ii) mixed sequence traces due to presence of alternative splicing, which interfere with base-calling of the sequences.

We carried out manual evaluation of the PCS and ERS sequence data followed by automated evaluation (see next section and Methods). For manual analysis, after clipping vector sequences as well as SL and poly(A) sequences from the original traces, RSTs were aligned to the *C. elegans* genomic sequence version WS150 using the Acembly program of AceDB (A *C. elegans* DataBase). For each set of RSTs that aligned to the expected genomic region, a transcript and ORF model was generated and compared with the existing WS150 models (complete listings in Supplemental files 2 and 3). Examination of the re-modeled ORFs enabled detection of alternatively spliced variants for both the ERS and the PCS (Table 1 and Table 2). Several of these variants had different internal exon structures, while others had alternative 5' and 3' exons. Many in the experimental set had mispredicted start/stop codon positions which could explain the previous failed attempts at cloning these ORFs. These results confirm previous suggestions (Reboul et al. 2001) that many of the genes that failed to amplify have mispredicted models.

To compare the obtained 5' RACE results to the known trans-splice acceptor sites of PCS and ERS transcripts, the clipped sequences of 5' RSTs were also aligned to the *C. elegans* genome with the BLAT program. RSTs containing sequences that could not be aligned to the genome were discarded. Out of 144 known trans-splice acceptor sites, 129 were verified by the RSTs; the other 15 were different but located within 25 kb of the RST starting site. BLAT results also located trans-splice acceptor sites for 35 ORFs among 44 ORFs without known trans-splice sites. In addition, 22% of genes have more than one trans-splice acceptor site.

### **Computational pipeline for generating RACE-defined transcript and ORF models**

To process and assemble RSTs and to ultimately construct transcript and ORF models we established a computational pipeline with multiple quality control filters (Fig. 1). We obtained both 5' RACE and 3' RACE products for all 94 PCS transcripts. Our automated pipeline generated one or more ORF models for 87 of the 94 PCS transcripts, whereas we were able to generate ORF models for all 94 manually (Supplemental files 1 - 3). For the ERS, we were able to generate models for 78 transcripts using our automated algorithm

and for 81 transcripts manually. Out of all the transcripts with at least one ORF model, there were 12 PCS and 35 ERS transcripts with only new ORF models (*i.e.*, no confirmatory WormBase models were found); and 25 PCS transcripts and 18 ERS transcripts containing both new models and WormBase matching models (Table 1). In total, 252 ORF models were generated for the PCS and ERS sets (details in Table 2). The RACE data also defined new ORF structures for ~13% of the PCS ORFs, indicating alternative models remain to be discovered even for “well annotated” worm transcripts.

To experimentally verify the ORF models derived from the RACE results, we generated RACE-defined transcript models for PCS and ERS sets, and then used these to generate primer pairs to PCR amplify 94 full-length ORFs. This set included 34 ORFs from the PCS, and 60 ORFs from the ERS. Following RT-PCR on RNA prepared from N2 worms (Supplemental Fig.1, Panel C), we cloned the generated PCR products. Sequencing of these revealed a 96% (90/94) success rate in amplifying the targeted ORFs. Hence, our RACE strategy efficiently delineates transcript boundaries and can guide more accurate ORF cloning.

### **Large scale RACE on ~2,000 unverified gene models**

With the necessary experimental and computational pipelines in place, we next scaled up the RACE experiments to interrogate 2,039 ORF models for which previous attempts at ORFeome cloning had failed (Reboul et al. 2003), representing ~25% of the unverified worm ORFeome (Supplemental Fig. 2). Of these, 1,569 ORF models were touched and 470 were untouched. Each RACE product was cloned and then sequenced from an individual minipool, unidirectionally from the 5' end. Of the 1,569 touched ORF models, 74% of the 5' RSTs and 82% of the 3' RSTs passed all quality control filters. Of the untouched genes, 39% of the 5' RSTs and 60% of the 3' RSTs passed the filters (Table 3; Supplemental File 4).

We note that the lower observed success rate for these RACE reactions relative to the benchmarking experimental sets (ERS) is not unexpected. All ERS transcript models were supported by EST evidence, so that their expression level is likely to be higher and

their internal exons correspondingly more likely to be accurately annotated. Also, individual colonies were sequenced from each minipool in the ERS set. Minipool sequencing provides several advantages such as rapidity, cost effectiveness, and evidence for the presence of major isoform (if present). When alternative forms are present in more or less equal ratios, base calling becomes difficult and the apparent success rate decreases. Individual colony sequencing deconvolutes the mixture and provides readable sequence, increasing the success rate. Future primer walking experiments, guided by available tiling-array data and transcriptome sequencing data, would help obtain RACE information on such transcripts.

### **Large-scale RACE derived structural annotation**

To construct transcript and ORF models using RSTs, we considered only the 5' and 3' RSTs that passed our quality control filters, from which we constructed 1,090 RACE-defined transcripts (or RD-transcripts) (Table 4; Supplemental file 4). Out of the 1,090 RD-transcripts, 973 constituted a full-length ORF, that is, an ORF with a recognizable ATG start codon and a stop codon (Table 4; Supplemental file 1; see Supplemental File 5 for complete listing with full-length sequences). Among these 973 generated ORF models, 627 (64%) confirmed WormBase release WS150 ORF models (or splice variants), while 346 (36%) were new, not present in WS150. Of the 346 new ORF models, 328 (or ~95%) were with redefined ends: 225 (or ~65%) had redefined 5' ends, 53 (or ~15%) had redefined 3' ends, and 50 (or ~15%) had redefined 5' and 3' ends (Fig. 2A). The remaining 18 (or 5%) represented internal alternative splice variants of models with unchanged ends. Among the 328 ORF models with redefined ends, 93 (28%) showed internal exon changes as well.

We further investigated the 80% (262/328) of new ORF models with 5' boundary changes, discovering that 30% (98/328) of these represented in-frame extensions or reductions of the 5' end. In 50% (164/328) of these the change was more complex, with the 5' end eliminated and replaced by a new end. For some the redefinition involved significant changes in chromosomal span. While most newly redefined ORFs showed less than 1 kb change in chromosomal span (Fig. 2B), we observed three ORFs (C10E2.3,

T02C5.5c and C37F5.1) with large changes in chromosomal span (over 10 kb). The C10E2.3 ORF model was 2.6 kb with a chromosomal span that was 10 kb longer than that of the original ORF model, while the C37F5.1 ORF model was 1.2 kb with a chromosomal span 21 kb shorter than the original model. A shorter transcript for C37F5.1 is also annotated in AceView (<http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/>; (Thierry-Mieg and Thierry-Mieg 2006), and the confirmation of the shorter form does not necessarily indicate that the longer annotation is incorrect. ORF model T02C5.5c was 6 kb long with a predicted span of 18 kb. A new 5' start site ~18 kb upstream of the annotated site was found for T02C5.5c, but the ORF size changed only slightly. We tested C10E2.3 and C37F5.1 by RT-PCR (T02C5.5c was not tested due to its long ORF length). We confirmed the complete ORF of C37F5.1 and the ends of C10E2.3 by cloning and sequencing.

Comparison of the 973 newly generated ORF models with those in WS150 found 84 new exons in 69 of the ORFs (Supplemental file 6). These exons were entirely novel, in that no part of the corresponding genomic sequences had previously been defined as exonic. We additionally identified 313 exons in 280 ORFs that modified the annotated exons by extending or truncating previously predicted exons (Supplemental file 7). We examined the splice sites of introns for both the new and modified exons. In WS150 transcript models, 99,592 out of 100,506 (or ~99%) exons corresponded to the most common GT/AG splicing signals. Among all the introns we examined, 99.5% (391/393) have GT/AG (386) or GC/AG (5) splicing signals (Supplemental file 8), the latter being the second most abundant splice site in *C. elegans* (Sheth et al. 2006).

To validate the generated ORF models we probed 143 of the models by RT-PCR followed by cloning and sequencing. We only attempted validation for ORFs less than 3 kb due to low processivity of the reverse transcriptase. The generated clones were sequenced and analyzed as minipools, or sometimes as single colony isolates. For ORF models shorter than 1.2 kb, we verified the entire ORF length by sequencing from both ends. For ORF models longer than 1.2 kb, we confirmed both ends by single pass sequencing. We confirmed 134 out of 143 (or ~94%) of the tested ORF models (

Of the tested cases, 112 represented new ORF models, and we confirmed 103 of these (or ~92% success rate). Amongst the tested models we did not observe a statistically significant difference between the confirmation rate of touched and untouched models (104 of 112 (or 95%) vs 30 of 33 (or 91%),  $p = 0.164$ ). In contrast, we had previously observed a significant difference between the confirmation rate of touched and untouched (Reboul et al. 2003; Lamesch et al. 2004), indicating that once an ORF model is confirmed by RACE the two classes behave similarly in verification experiments.

Our RACE platform captures the full-length 5' UTR elements, but does not ensure capture of all full-length 3' UTR elements due to low sequence complexity or extended length of the 3' UTR. Among the 973 generated RD-transcripts with full-length ORFs, 366 (36%) confirmed the 5' UTR in WormBase, 205 (21%) added new variants and 402 (43%) had no 5' UTR information available in WormBase (Fig. 2C). Nine percent of our ORFs had no associated 5' UTRs. In contrast, among genes with defined 5' transcript boundaries in WormBase, six percent have no 5' UTR. This difference is statistically significant ( $p = 0.000042$ ). The higher percentage of genes without a 5' UTR found by experimental RACE is consistent with *trans*-splicing of many *C. elegans* mRNAs, resulting in short 5' UTRs with the SL located near the start of the ORF (Page et al. 1997). We carried out a similar analysis for the 3' UTRs that we could completely define (~49% (479/973) of generated ORF models). Among these only 10% (49 out of 479) confirmed previously annotated 3' UTRs. The other 90% (430 out of 479) had new 3' UTRs which either redefined previously annotated 3' UTRs or had not been described previously (Fig. 2D).

### **Alternative *trans*-splice leader usage**

For ~15% of the 5' RSTs the quality score of the SL sequence was poor, while the surrounding sequences (both the Gateway tails and downstream transcript sequences) had high quality scores. These differences indicate possible alternatively spliced transcripts in the sequenced minipools. Because we had used a mixture of SL1 and SL2 for 5' RACE, and because competitive (or alternative) SL1-SL2 *trans*-splicing in operons has been reported (Graber et al. 2007), these RACE products likely contained a mixture of

transcripts derived from the same gene but linked either to SL1 or SL2, in other words, alternatively *trans*-spliced. To investigate the extent to which such putative competitive splicing is associated with alternative splicing of downstream transcript sequences, we examined 28 such 5' RSTs, sequencing 12 colonies isolated from each respective minipool (Supplemental file 9). We found that 16 of the 28 5' RST minipools investigated contained mixtures of homologous genes due to mispriming. Two categories of mispriming are observed, 1) mispriming of gene-specific primer to a homologous gene, and 2) internal priming by SL1/SL2 in place of the gene-specific primer. These mispriming events were not further investigated.

Clones from the remaining 12 minipools did contain transcripts of the same gene linked to either SL1 or SL2, confirming alternative *trans*-splicing. Among the isolated clones, 60 contained the SL1 sequence and 62 had the SL2. We manually searched upstream sequences for the most common 3'-splicing signal sequences (UUCAG, UUUCA, AUUU, and UUUUC) within a window from -10 to 0, and the U-rich elements (UAUUUU, UACUU, UAUCU, UAUUU, CUUUU and UUUCU) from -100 to 0 associated with *trans*-splicing (Graber et al. 2007). Of the 60 SL1 transcripts, 57 (95%) have both splicing signal and U-rich elements (Supplemental file 9). Of 62 SL2 transcripts, 58 (94%) have U-rich elements and 39 (63%) have both (Supplemental file 9). In 8 of these 12 minipools, sequences linked to SL1 and SL2 showed no difference in the downstream regions. For 3 of these 8 minipools, the corresponding transcript models were reportedly present in co-transcribed operons. R11D1.9 and C42C1.13 are the first genes in their respective operons, while Y56A3A.13b is downstream in its operon. In the remaining 4 of the 12 minipools, we captured alternative transcript variants. For C52D10.7, C52D10.9 and F26F12.7, SL1 was linked exclusively to isoforms extended at the 5' end (Table 6, Fig. 3). In F56H9.2, an extended 5' form was linked preferentially to SL1. C52D10.7 and C52D10.9 are highly homologous tandem repeat genes that displayed similar SL selection pattern (Fig. 3); neither is reported to be in an operon. While alternative promoter usage can explain alternative SL acceptor sites, alternative *trans*-splicing can be a competing mechanism leading to formation of these transcripts. In sum, alternative *trans*-spliced leaders are found in approximately 6% of the transcript

models tested, and alternative *trans*-spliced leaders are found in some cases to be preferentially associated with different transcript variants.

## Discussion

Proactively defining transcript or ORF structure with the large-scale RACE platform described here offers the opportunity to discover transcripts features not conforming to rules used in current *ab initio* gene predictions. *Proactivity* is critical where model-based *reactive* ORF verification (Reboul et al. 2003; Lamesch et al. 2004) fails. The high-throughput RACE approach relies only modestly on predicted transcript models. We have also demonstrated that our approach has excellent sensitivity and specificity. Most *C. elegans* transcripts have SL1 or SL2 leader sequences (~55% SL1, 15% SL2; (Blumenthal et al. 2002)). By including both SL1 and SL2 primers in the 5' RACE reactions, we could capture both SL1 and SL2 containing transcripts, targeting ~70% of protein coding transcript space. The SL mediated 5' RACE will clearly not work for transcripts without SL1 and SL2 sequences. The remaining transcripts, not having SL1 or SL2 leaders, can be captured by the ligation method (Schaefer 1995; Chenchik et al. 1996; Manichaikul et al. 2009), though we expect the efficiency of RNA ligation will be less than the SL1/SL2 approach. Transcripts with SL3-5 splice leader, though, form a minority of *trans*-spliced transcripts, and could be captured in future experiments by using primers corresponding to these leader sequences in the RACE PCR reaction.

Of the ORF models that we generated, ~31% of the touched but unverified models differed from the WormBase models. Strikingly, for untouched models whose cloning was attempted and failed previously, over ~73% of generated ORF models were novel relative to WS150 models. We also found alternative ORF structures for 13% of “well-annotated” positive control genes that we examined. Although a novel ORF model does not necessarily imply that the previous model is wrong, our results show that over one-fifth of *C. elegans* ORFeome can be alternatively annotated. Decisions to update models are left to annotation curators, who do so on the basis of their particular criteria for accepting alternative annotations or retiring existing ones. Genome-wide validation of

these annotations is urgently needed to gain a systems-level understanding of the protein-coding potential of the worm.

While technologies for whole-genome sequencing have rapidly evolved, the difficulties inherent in defining transcript and ORF structures within metazoan genomes have persisted. Several available large-scale “survey” methodologies, including CAGE (Shiraki et al. 2003), TEC-RED (Hwang et al. 2004), GIS-PET (Chiu et al. 2007; Ng et al. 2007), and MS-PET (Ng et al. 2006), may be employed to aid defining transcript structures and boundaries. These methods provide short tags that define transcript end(s), but importantly do not probe the internal exon structures. Transcriptome sequencing using parallel sequencing platforms (RNA-Seq) can define precise exon boundaries (Cloonan et al. 2008; Mortazavi et al. 2008; Nagalakshmi et al. 2008; Shin et al. 2008; Wilhelm et al. 2008). However, the wide dynamic range of gene expression in higher eukaryotes severely limits the ability of RNA-Seq approaches to determine transcript structure of rarely expressed genes. Recent transcriptome sequencing of L1 stage worms with the 454 sequencing platform, while informative, provided reads for only ~6,100 protein coding genes and partially verified only 200 untouched genes (Shin et al. 2008). Platforms with higher coverage might provide greater depth, but determining *cis*-connectivity of exons across the entire length of transcripts would still not be possible given the short read lengths of these platforms. The recent worm RNA-Seq effort using the Illumina Genome Analyzer (“Solexa”) platform successfully touched most transcript models (Hillier et al. 2009). However, the “Genelet” models generated from assembly of the obtained sequences are often truncated due to uneven coverage. In addition to *cis*-connectivity limitations, some transcriptome sequencing approaches lose strand information of the transcripts, limiting use for transcript annotation (Mortazavi et al. 2008).

How does targeted RACE compare with non-targeted, but highly parallel transcriptome sequencing? We carried out detailed analysis of the SL and exon structures determined by RACE with the mid-L2 transcriptome sequencing reported by Hillier et al. (Hillier et al. 2009) The SL1 and SL2 trans-splice acceptor sites identified in our

RACE were compared to the SL1 and SL2 trans-splice acceptors (based on release WS180) and Hillier et al. (Supplemental file 1). Out of 1,067 SL1 sites we identified, 310 sites overlap with Hillier et al., 564 sites matched to WS180 annotation, and 249 sites appear in all three datasets. For 207 SL2 sites we found, 61 were identified in all three approaches, while 68 and 123 matched to Hillier et al. (Supplemental file 1: Supplemental Fig. 3 and Supplemental Table 1) and WS180 annotation, respectively. In defining the position of SL1 trans-splicing, there is only 37% overlap within the tested space between findings reported here and the L2 stage reported by Hillier et al. Likewise, the overlap for SL2 sequence is ~33%. In summary, while there is overlap between the two data sets, the RACE approach provides new and alternative SL sites absent in the transcriptome sequencing.

We compared the 85 newly detected exons with exons found in Hillier et al. (Supplemental file 1: Supplemental Table 2). Approximately 42% (36) of exons do not overlap with Hillier et al. altogether; only ~29% (25) match exactly. We see similar trends for exons that we re-defined experimentally ((Supplemental file 1, Supplemental Table 3). Of the 313 exons that we modified, 88 (~28%) show no overlap, i.e., are not annotated by Hillier et al.; 172 (or ~55%) show different 5' and/or 3' boundaries; and only 46 (~15%) show exact matches.

The comparison of our RACE results with current RNA-Seq results shows that transcriptome sequencing generates distinct and potentially complementary results. Improvements in *cis*-connectivity determination of splicing events are expected with further technological innovations in RNA-Seq approaches, such as paired-end sequencing (Fullwood et al. 2009). Ambiguities are still likely to persist, as paired-end sequencing gives connectivity for transcript ends but not for splicing events in the middle of an ORF. The high-throughput RACE platform can correctly annotate protein coding genes of *C. elegans*. RACE, transcriptome sequencing, tiling array, PET, and others provide distinctive solutions to defining transcript structure, each with its own strengths and limitations. The final elucidation of genome-wide transcript annotation, in worm or any other eukaryote, rests on integration of multiple high throughput approaches, benefiting

from strengths that each platform offers.

## Methods

### RACE experiments

#### *General RACE primers*

The following primers were used for the first PCR of the 5' RACE experiments (sequences given 5' to 3'): SL1: GGTTTAATTACCCAAGTTTGAG; SL2: GGTTTAAACCCAGTTACTCAAG. The second 5' RACE PCR reactions used: GFSL1: GGGGACAACCTTTGTACAAAAAAGTTGGCGGTTTAATTACCCAAGTTTGAG, and GFSL2:

GGGGACAACCTTTGTACAAAAAAGTTGGCGGTTTAAACCCAGTTACTCAAG.

For the 3' RACE experiments, the following primers, derived from Invitrogen GeneRACER kit, were used: reverse transcription priming with the GR3 Primer, GCTGTCAACGATACGCTACGTAACGGCATGACAGTGTTTTTTTTTTTTTTTTTTTTTTT TTT TTT; the first PCR was done with GR3, GCTGTCAACGATACGCTACGTAACG; and the final amplification was done using GGRn3, GGGGACAACCTTTGTACAAGAAAGTTGGGCGCTACGTAACGGCATGACAGTG.

#### *Gene-specific RACE Primer design*

For the 5' RACE experiments, we designed two nested reverse primers antisense to the putative ORF region of the gene of interest. The distal primer was placed 100-500 bases 3' to the putative start of the ORF while the more proximal reverse primer, typically positioned in tandem to the distal primer, was tailed with the Gateway B2 sequence to allow recombinational cloning (Walhout et al. 2000b). For the forward primer of the 5' RACE we used a pool of SL1/SL2 sequences, each tailed with the B1.1 Gateway sequence at its 5' end. The nested 3' RACE primers had the same general design as the 5' RACE primers, except that they were in the forward orientation (sense relative to the mRNA) and the primer proximal to the poly(A) tail also contained a Gateway B1 tail. The distal primer was placed 100 to 400 bases upstream of the putative stop codon. The dT24 primer contained the Gateway B2.2 tail. All gene-specific primers were located in annotated exons and were designed to have a  $T_m$  between 55°C and 65°C.

#### *Generation of RACE amplicons*

To generate RACE fragments we reverse transcribed total *C. elegans* RNA, isolated from mixed-stage, asynchronously growing worm populations, using either dT<sub>16</sub> (for 5' RACE) or the tailed oligo dT GR3 primer for 3' RACE. For the first of the two nested PCRs we performed touchdown PCR (the annealing temperature of the first 10 cycles was 65°C, on average 5-10 degrees above the  $T_m$  of the gene-specific primers) using the distal gene-specific primers along with the appropriate universal primer (a 1:1 mixture of SL1 and SL2 primers for the 5' RACE reactions and a tailed dT16 GR3 Primer, for the 3' RACE reactions). For these PCR reactions we adjusted the amount of reverse transcribed material such that ~150 ng total RNA was present per reaction. We used less than 0.5  $\mu$ L of the first PCR reaction for the second stage PCR with the nested and tailed proximal gene-specific primers. Nested PCR step increases sensitivity and specificity of the experiment while providing Gateway tails for cloning.

### **Gateway cloning of amplicons and sequencing**

PCR products generated in RACE or in ORF verification experiments were recombinationally cloned by a BP reaction into pDONR223 to generate Gateway Entry clones (Rual et al. 2004a). The products from the BP reactions were used to transform chemically competent DH5a *E. coli*, in 96-well microtiter plates containing spectinomycin, for growth and selection of cells bearing Entry clones. Following growth in liquid media, the transformed bacteria were used as a source of template in PCR reactions, using vector primers to generate the final DNA template for sequencing. PCR products were sequenced using conventional automated cycle sequencing to generate RACE sequence tags (RSTs) or ORF sequence tags (OSTs (Reboul et al. 2001)). Sequencing was carried out by Agencourt Bioscience Corp. (Beverly, Mass).

For the benchmark sets, forward and reverse reads were obtained for the cloned 5' RACE products. For the rest of the RACE experiments, including all 3' RACE, only a 5' forward read was generated. For ORF verification, 5' and 3' reads were obtained.

For ORF verification experiments, Gateway-tailed primers were designed to amplify complete ORFs as described (Reboul et al. 2003). PCR products were Gateway-cloned and sequenced from both ends to generate ORF Sequence Tags (OSTs). Vector and quality trimmed OSTs from both ends were assembled. If there was overlap between the 5' and 3' OSTs, a contig was assembled and further analyzed to find the reading frame.

### **Reconstruction of transcripts from RACE sequence**

#### *Manual Analysis of the benchmark sets*

All sequence traces were assigned a unique ID and stored in a MySQL database. Each RST was initially aligned to its corresponding annotated WS150 CDS using the bl2seq program. Only RSTs with an alignment length greater than 100 bp and with high sequence quality (having 200 or more consecutive bases with a PHRED  $\geq 20$ ) were retained for further analysis. After clipping vector sequences, low quality sequences, SL1/SL2, and poly(A) sequences from the original traces, RSTs were aligned to the *C. elegans* genome, WormBase sequence version WS150, using the assembly program of the AceDB (a *C. elegans* DataBase). For each set of RSTs that aligned to the expected genomic region, an ORF model was generated and compared with the existing WS150 model.

#### *Large-scale RACE analysis*

We generated Perl scripts to process and analyze the RACE data. The computational pipeline: RST sequences were base-called and vector trimmed with PHRED. Quality trimming excluded sequences with average PHRED score below 15 in a sliding window of 20 nucleotides. Vector and quality-trimmed RSTs were aligned by BLAT against WormBase WS150 genomic sequence. We enforced two filters. First, for the 5' RSTs, no gap on RST was allowed between the 5' SL primer and the continuing sequences. Similarly, in the 3' RACE, no gap was allowed between the gene-specific primer and the rest of the RST. Such gaps were occasionally observed as short stretches of sequences on the RST that could not be aligned to the genome. Second, for any RST with a gap between RST-exons, the exons following the gap were not further processed. Those RSTs

passing the imposed filters were broken into hit blocks (RST-exons) based on their best BLAT hits, to avoid inclusion of homologous gene segments. To eliminate errors (PCR-induced or otherwise), genomic sequences corresponding to RST-exons were identified and used in place of the actual RST-exons in all subsequent steps.

Processed and filtered 5' and 3' RSTs of the same targeted gene were merged to generate a transcript model. Usually the RSTs did not span the entire transcript model; for these (~90% of the transcripts) we used the existing WS150 transcript model sequences to fill in the gap between the 5' and 3' RSTs, generating RST-model-hybrids.

### **ORF and UTR prediction from RD-transcripts**

The first ATG that gave rise to the longest open reading frame in each RD-transcript was used to define the start of the open reading frame. The ORF is considered “complete” if a stop codon was found (only rarely was a stop codon not found). Once the ORF region was defined, 5' and 3' UTRs were assigned. The first base of 5' UTR was defined by the end of the SL sequence, and the end of the 5' UTR was defined by the first ATG of the ORF. For 3' UTR assignment, the ORF stop codon delineated the start of the UTR and the poly(A) stretch marked the end. We aligned the sequence to the genome, and a complete 3' UTR was characterized as exons with no gaps in between and followed by a poly(A) tract. Sometimes, due to low complexity or extended length, the complete 3' UTR could not be defined.

### **Databases**

We used WormBase WS150 (released Oct 2005) for primer design, sequence analysis and comparison. The UCSC Genome Browser (<http://genome.ucsc.edu/cgi-bin/hgBlat>) was used to analyze and display individual sequences, with the March 2004 assembly of the *C. elegans* genome sequence as the reference genome sequence.

## **Acknowledgments**

This work was funded by a grant from the Ellison Foundation (awarded to M.V.), and by Institute Sponsored Research funds from the Dana-Farber Cancer Institute Strategic Initiative in support of the Center for Cancer Systems Biology (CCSB). M.V. is a ‘Chercheur Qualifié Honoraire’ of the Fonds de la Recherche Scientifique (FRS-FNRS, French Community of Belgium). F.P.R. was supported in part by NIH grants HG004233, HG0017115, HG003224 and by the Canadian Institute for Advanced Research. We thank Adnan Derti for helpful discussions and comments on the manuscript.

## **Authors' contributions**

M.V., K.S.A., D.E.H., and F.P.R. conceived the project, K.S.A., D.E.H. and M.V. directed the execution of the project. K.S.A., D.Z., H.J.L., R.R.M., X.Y., S.M., N.S. carried out the RACE and cloning. C.L., T.H., C.F., and K.S.A., established the computational RACE pipeline, C.L., T.H., L.G., Y.S., and K.S.A. carried out the analyses. K.S.A., C.L., D.E.H., M.E.C., and M.V. wrote the manuscript.

## **Figure Legends**

**Figure 1.** Experimental and computational RACE pipelines. (A) Experimental strategy begins with reverse transcription of messenger RNA (with a tailed dT<sub>24</sub> for 3' RACE) to generate a complementary DNA strand used as template to generate the first set of RACE products. These amplicons are then reamplified with nested internal primers to generate the final RACE products, which are cloned, then sequenced. (B) Computational pipeline begins with vector and quality trimming of the reads. RACE sequences are aligned by BLAT against the *C. elegans* genome then parsed as genome (exon) blocks. RACE sequences are then replaced by matching genomic sequences to correct sporadic sequencing or PCR errors. If an exon is of low complexity, or has regions that can not be aligned to the genome, the exon, together with subsequent exons (if present), is trimmed. For short transcripts, 5'-RSTs and 3'-RSTs can be readily assembled to generate full transcripts. For longer transcripts the gaps of non-overlapping 5'-RST and 3'-RST are filled by sequences from Wormbase transcript models.

**Figure 2.** Generated transcript and ORF models compared with the corresponding WormBase annotation. (A) The percentage of new ORF models that differ from the corresponding WormBase models. If an ORF model had internal changes co-existing with changes in boundary (redefined 5', 3' or both ends), the ORF model was counted as having a change in boundary. (B) Chromosomal span change of novel ORF models in kilobases (kb). A change can be either an increase or a decrease in span. (C, D) Changes in annotated UTRs. UTRs in the RD-transcript models are divided into several categories: confirmatory (WormBase), new variant (different from WormBase) and never previously defined (no UTR indicated in WormBase). Many transcript models in WormBase do not have UTRs defined, and a model can be annotated as having multiple UTRs. "Incomplete UTRs" are those 3' UTRs that could not be defined with certainty.

**Figure 3.** Examples of alternative *trans*-spliced leader usage. Single colony sequencing was done on RSTs that showed evidence of mixed *trans*-splicing. (A-C) Alignment of reads obtained from individual 5' RACE clones shows association of SL1 and SL2 to different transcript variants. (D) Alternative usages of SL1 and SL2 were observed for the longer transcript variant, while SL2 *trans*-splicing is observed for the shorter variant.

## Tables

**Table 1. Summary of “benchmark set” annotation and confirmation results**

Transcripts	Positive control set	Experimental reference set
Models tested	94	94
WS150 models only	50	25
New ORF isoforms only	12	35
WS150 and new ORF isoforms	25	18
Total 87		78
ORF models attempted/models verified	30/30	64/60

**Table 2. Distribution of the benchmark ORFs (PCS plus ERS)**

Description	Total ORFs	ORFs in PCS	ORFs in ERS
Match model	118	75	43
Internal difference only	12	7	5
Modified 3' end	22	6	16
Frame shift	1	1	0
Modified 5' end	76	37	39
Modified 5' and 3' ends	13	2	11
Match model manually	10	7	3
Total	252	135	117

**Table 3. Cloning, sequencing and analysis of RACE products<sup>†</sup>**

	5' RACE			3' RACE		
	Overall	Touched by EST	Untouched by EST	Overall	Touched by EST	Untouched by EST
Attempted	2,039	1,569	470	2,039	1,569	470
Cloned	1,850	1,457	393	1,773	1,425	348
Passed minipool filters	1,346	1,164	182	1,574	1,219	283

<sup>†</sup> Positive cloning of RACE products were deduced from the presence of recombined Gateway tags, rather than the presence of high PHRED score insert sequences since multiple splice forms in minipools could generate mixed unreadable sequence reads. Sequences eliminated by minipool filters were either unreadable in the supposed insert region, or had gaps between the Gateway tag and insert sequence.

**Table 4. Summary of RST and ORF annotation success rates**

	Total	Touched models <sup>†</sup>	Untouched models <sup>†</sup>
Attempted	2,039	1,569	470
RD-transcripts <sup>††</sup> defined	1,090	961	129
Full length ORF models generated	973	869	104

<sup>†</sup> “Touched” and “untouched” refers to annotated models with and without EST evidence.

<sup>††</sup> RACE-defined transcripts.

**Table 5. Confirmation of ORF models**

	Tested	Confirmed <sup>†</sup>
Total	143	134
WS150 models	31	31
New ORF models	112	103
Touched ORF models <sup>††</sup>	110	104
Untouched ORF models <sup>††</sup>	33	30

<sup>†</sup> Confirmed by cloning and sequencing.

<sup>††</sup> ‘Touched’ and ‘untouched’ refers to ORF models with and without EST evidence.

**Table 6. Summary of “mix” SL1/SL2 usages from deconvoluted 5’ RST minipools**

	Total Sequences	SL1 form	SL2 form	Extended 5' end	
				SL1	SL2
C52D10.9	12	5	7	4	0
C52D10.7	11	4	7	4	0
F26F12.7	12	7	5	7	0
F56H9.2	11	3	8	4	2

## **Supplemental files**

### **Supplemental File 1 – Supplemental figures and tables**

### **Supplemental File 2 – RACE and RD-transcripts of the benchmark sets**

5' RACE “out” and “in” primers are nested reverse primers for 5' RACE. Likewise 3' RACE “out” and “in” primers are nested forward primers for 3' RACE. RSTs (5' and 3' reads) are vector and quality trimmed. The field is blank if there is no good quality sequence after vector and quality trimming. RACE defined transcripts (RD transcripts) are generated by merging corrected 5' and 3' RSTs. The field is blank if either 5' or 3' RST did not pass minipool filter.

### **Supplemental File 3 – ORF models of the benchmark sets**

Sequences of ORF models built for the benchmark sets including the positive control set (PCS) and the experimental reference set (ERS).

### **Supplemental File 4 – RACE and RD-transcripts of the main experimental set**

RACE “out” and “in” primers are nested reverse primers for 5' RACE. Likewise 3' RACE “out” and “in” primers are nested forward primers for 3' RACE. RSTs are vector and quality trimmed. The field is blank if there is no good quality sequence after vector and quality trimming. RACE defined transcripts (RD transcripts) are generated by merging corrected 5' and 3' RSTs. The field is blank if either 5' or 3' RST did not pass the minipool filter.

### **Supplemental File 5 – Sequences of ORF models generated for the main experimental set**

Complete sequences of ORF models generated for the main experimental set.

### **Supplemental File 6 – Novel exon summary**

Target transcript ID and chromosomal coordinates of the novel exons (position and length).

### **Supplemental File 7 – Modified exon summary**

Target transcript ID and chromosomal coordinates of the modified exons (position and length).

### **Supplemental File 8 – Splicing signal summary**

Target transcript ID, chromosomal coordinates of the modified/new exon boundary, splice signal, and type of exon modification.

## Supplemental File 9 – Sequences of alternatively trans-spliced SL1 and SL2

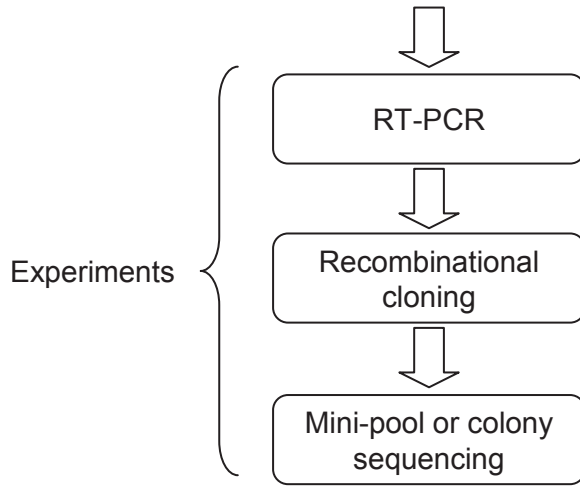
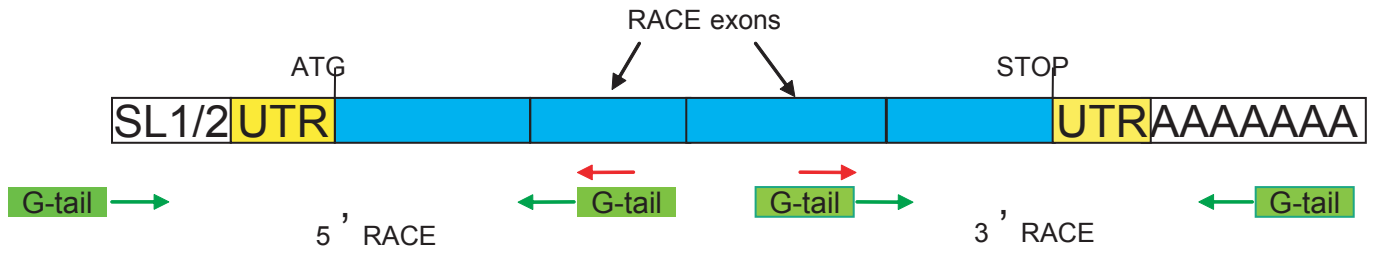
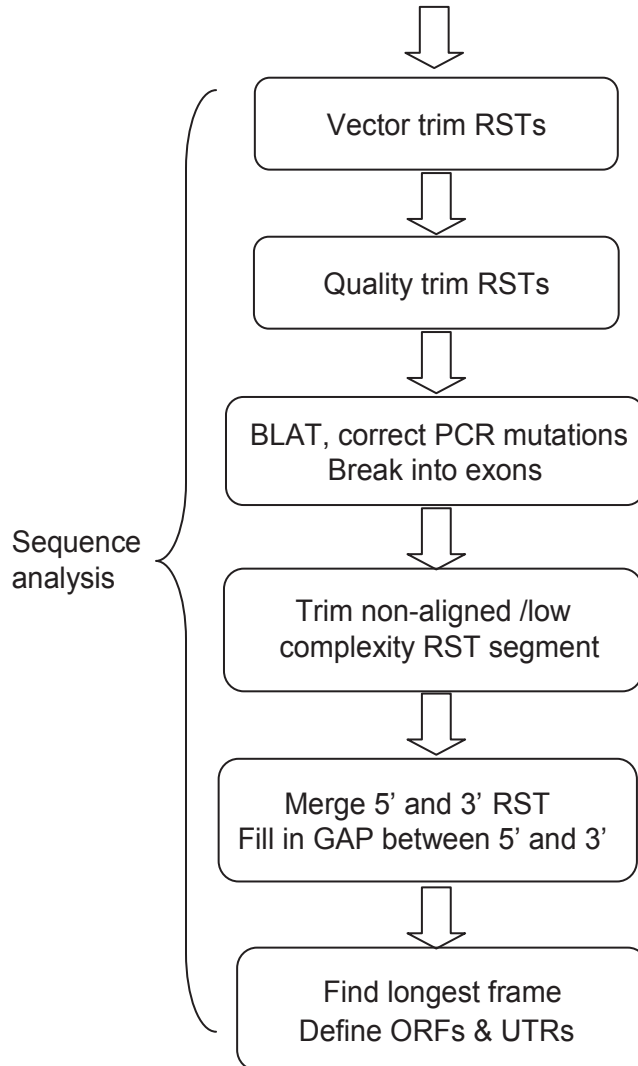
Twelve colonies were sequenced from both ends in each case. RSTs were vector and quality-trimmed then assembled.

## References

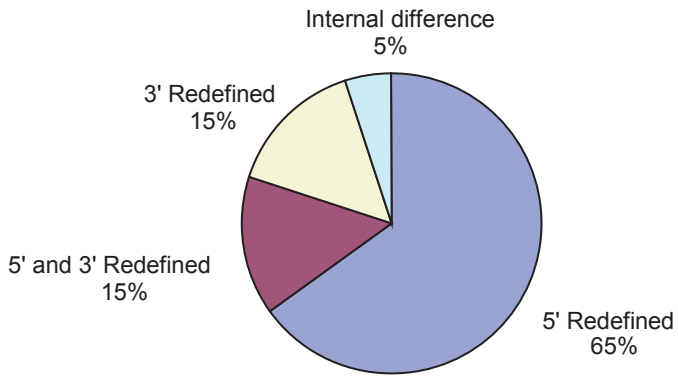
- Bao, Y. and Hull, R., 1993. Mapping the 5'-terminus of rice tungro bacilliform viral genomic RNA. *Virology* **197**: 445-448.
- Blumenthal, T., 2005. Trans-splicing and operons. *WormBook*: 1-9.
- Blumenthal, T., Evans, D., Link, C.D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W.L., Duke, K., Kiraly, M. et al., 2002. A global analysis of *Caenorhabditis elegans* operons. *Nature* **417**: 851-854.
- Chenchik, A., Diachenko, L., Moqadam, F., Tarabykin, V., Lukyanov, S., and Siebert, P.D., 1996. Full-length cDNA cloning and determination of mRNA 5' and 3' ends by amplification of adaptor-ligated cDNA. *Biotechniques* **21**: 526-534.
- Chiu, K.P., Ariyaratne, P., Xu, H., Tan, A., Ng, P., Liu, E.T., Ruan, Y., Wei, C.L., and Sung, W.K., 2007. Pathway aberrations of murine melanoma cells observed in Paired-End diTag transcriptomes. *BMC Cancer* **7**: 109.
- Cloonan, N., Forrest, A.R., Kolle, G., Gardiner, B.B., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G. et al., 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5**: 613-619.
- Conrad, R., Lea, K., and Blumenthal, T., 1995. SL1 trans-splicing specified by AU-rich synthetic RNA inserted at the 5' end of *Caenorhabditis elegans* pre-mRNA. *RNA* **1**: 164-170.
- Fullwood, M.J., Wei, C.L., Liu, E.T., and Ruan, Y., 2009. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res* **19**: 521-532.
- Graber, J.H., Salisbury, J., Hutchins, L.N., and Blumenthal, T., 2007. *C. elegans* sequences that control trans-splicing and operon pre-mRNA processing. *RNA* **13**: 1409-1426.
- He, H., Wang, J., Liu, T., Liu, X.S., Li, T., Wang, Y., Qian, Z., Zheng, H., Zhu, X., Wu, T. et al., 2007. Mapping the *C. elegans* noncoding transcriptome with a whole-genome tiling microarray. *Genome Res* **17**: 1471-1477.
- 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.
- Hillier, L.W., Reinke, V., Green, P., Hirst, M., Marra, M.A., and Waterston, R.H., 2009. Massively parallel sequencing of the polyadenylated transcriptome of *C. elegans*. *Genome Res* **19**: 657-666.
- Huang, X.Y. and Hirsh, D., 1989. A second trans-spliced RNA leader sequence in the nematode *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* **86**: 8640-8644.
- Hwang, B.J., Muller, H.M., and Sternberg, P.W., 2004. Genome annotation by high-throughput 5' RNA end determination. *Proc. Natl. Acad. Sci. USA* **101**: 1650-1655.

- Krause, M. and Hirsh, D., 1987. A *trans*-spliced leader sequence on actin mRNA in *C. elegans*. *Cell* **49**: 753-761.
- Lamesch, P., Milstein, S., Hao, T., Rosenberg, J., Li, N., Sequerra, R., Bosak, S., Doucette-Stamm, L., Vandenhoute, J., Hill, D. et al., 2004. *C. elegans* ORFeome version 3.1: increasing the coverage of ORFeome resources with improved gene predictions. *Genome Res* **14**: 2064-2069.
- Manichaikul, A., Ghamsari, L., Hom, E.F.Y., Lin, C., Murray, R.R., Chang, R.L., Balaji, S., Hao, T., Shen, Y.C., Arvind K., Thiele, I. et al., 2009. Metabolic network analysis integrated with genome-wide transcript verification. *Nat Methods* **In press**.
- Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**: 621-628.
- Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M., 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344-1349.
- Ng, P., Tan, J.J., Ooi, H.S., Lee, Y.L., Chiu, K.P., Fullwood, M.J., Srinivasan, K.G., Perbost, C., Du, L., Sung, W.K. et al., 2006. Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic Acids Res* **34**: e84.
- Ng, P., Wei, C.L., and Ruan, Y., 2007. Paired-end diTagging for transcriptome and genome analysis. *Curr Protoc Mol Biol* **Chapter 21**: Unit 21 12.
- Page, B.D., Zhang, W., Steward, K., Blumenthal, T., and Priess, J.R., 1997. ELT-1, a GATA-like transcription factor, is required for epidermal cell fates in *Caenorhabditis elegans* embryos. *Genes Dev* **11**: 1651-1661.
- Reboul, J., Vaglio, P., Rual, J.F., Lamesch, P., Martinez, M., Armstrong, C.M., Li, S., Jacotot, L., Bertin, N., Janky, R. et al., 2003. *C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.* **34**: 35-41.
- Reboul, J., Vaglio, P., Tzellas, N., Thierry-Mieg, N., Moore, T., Jackson, C., Shin-i, T., Kohara, Y., Thierry-Mieg, D., Thierry-Mieg, J. et al., 2001. Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans*. *Nat. Genet.* **27**: 332-336.
- Rual, J.F., Hill, D.E., and Vidal, M., 2004a. ORFeome projects: gateway between genomics and omics. *Curr. Opin. Chem. Biol.* **8**: 20-25.
- Rual, J.F., Hirozane-Kishikawa, T., Hao, T., Bertin, N., Li, S., Dricot, A., Li, N., Rosenberg, J., Lamesch, P., Vidalain, P.O. et al., 2004b. Human ORFeome version 1.1: a platform for reverse proteomics. *Genome Res* **14**: 2128-2135.
- Schaefer, B.C., 1995. Revolutions in rapid amplification of cDNA ends: new strategies for polymerase chain reaction cloning of full-length cDNA ends. *Anal Biochem* **227**: 255-273.
- Sheth, N., Roca, X., Hastings, M.L., Roeder, T., Krainer, A.R., and Sachidanandam, R., 2006. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res.* **34**: 3955-3967.

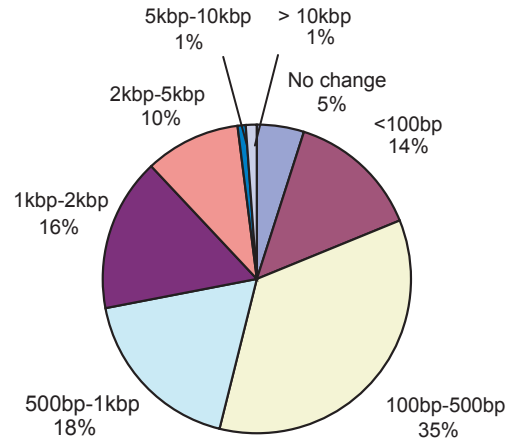
- Shin, H., Hirst, M., Bainbridge, M.N., Magrini, V., Mardis, E., Moerman, D.G., Marra, M.A., Baillie, D.L., and Jones, S.J., 2008. Transcriptome analysis for *Caenorhabditis elegans* based on novel expressed sequence tags. *BMC Biol* **6**: 30.
- Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T. et al., 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. USA* **100**: 15776-15781.
- The *C. elegans* Sequencing Consortium., 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**: 2012-2018.
- Thierry-Mieg, D. and Thierry-Mieg, J., 2006. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol* **7 Suppl 1**: S12 11-14.
- Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A., Thierry-Mieg, N., and Vidal, M., 2000a. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**: 116-122.
- Walhout, A.J., Temple, G.F., Brasch, M.A., Hartley, J.L., Lorson, M.A., van den Heuvel, S., and Vidal, M., 2000b. GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes. *Methods Enzymol* **328**: 575-592.
- Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., and Bahler, J., 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**: 1239-1243.

**A****B**

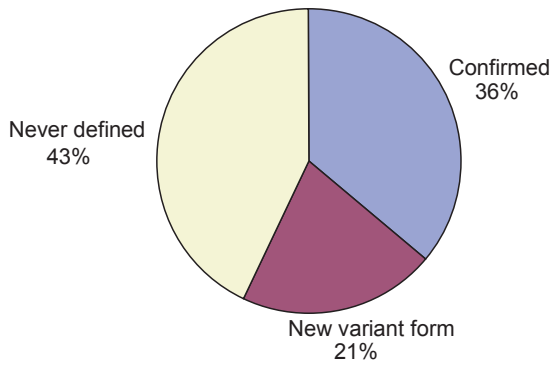
**A**



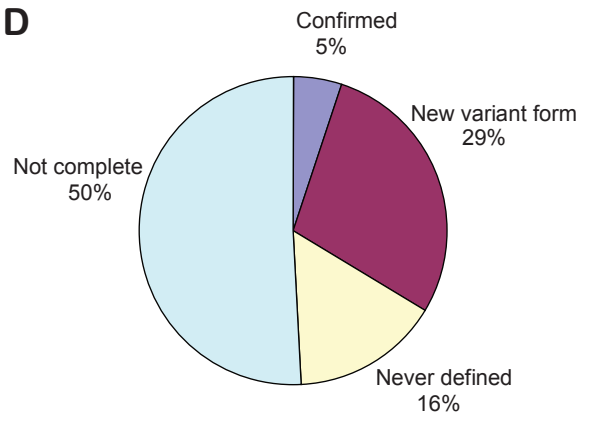
**B**



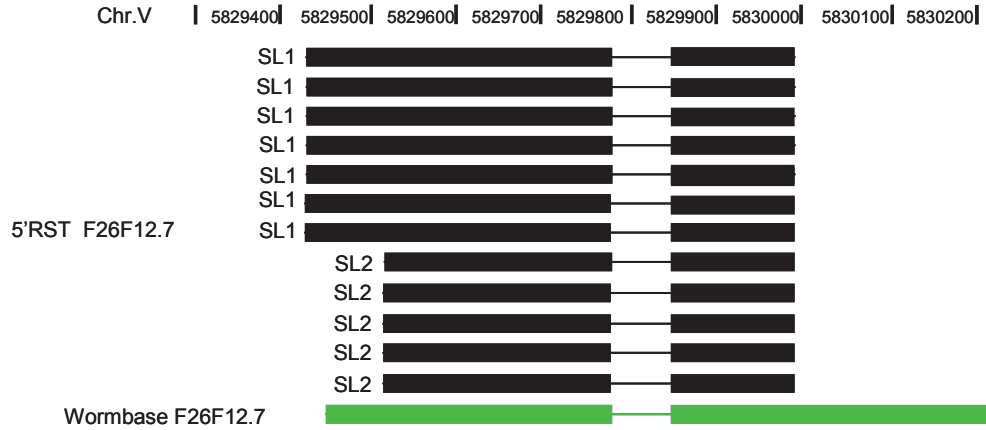
**C**



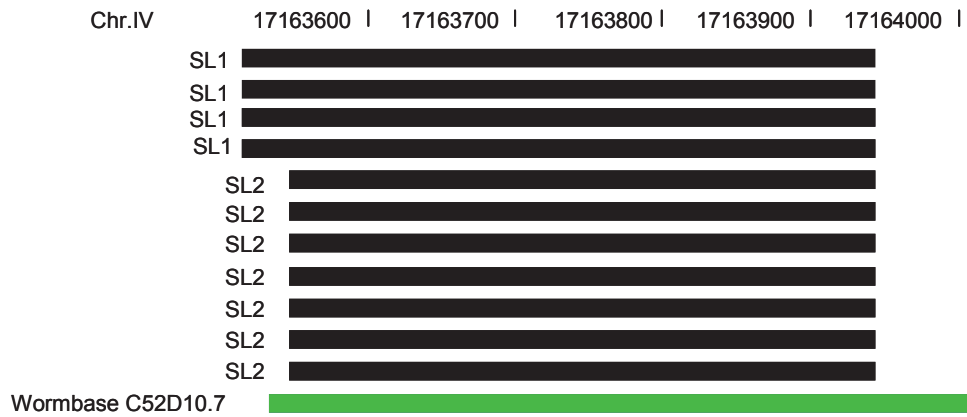
**D**



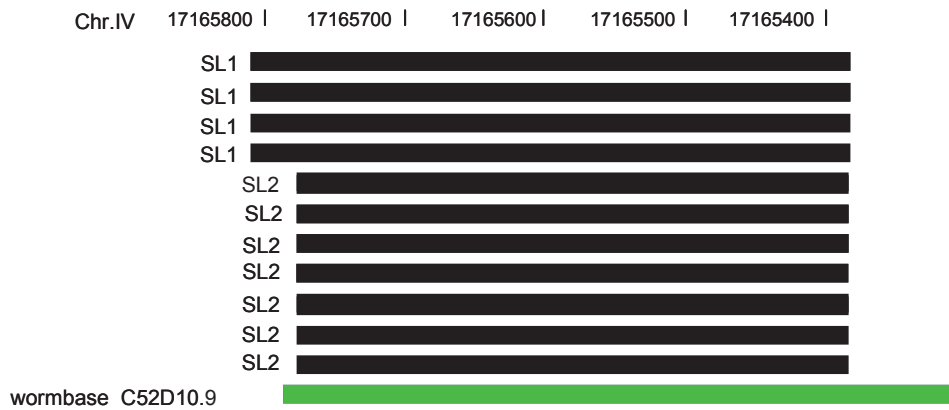
### A F26F12.7



### B C52D10.7



### C C52D10.9



### D F56H9.2

