RESEARCH PAPER

RNA Biology 9:5, 1-14; May 2012; © 2012 Landes Bioscience

Multiple insert size paired-end sequencing for deconvolution of complex transcriptomes

Lisa M. Smith,[1] Lisa Hartmann,[2] Philipp Drewe,[3] Regina Bohnert,[3] André Kahles,[3] Christa Lanz[1] and Gunnar Rätsch[4,*]

[1]Department of Molecular Biology; Max Planck Institute for Developmental Biology; Tübingen, Germany; [2]Center for Plant Molecular Biology; University of Tübingen; Tübingen, Germany; [3]Friedrich Miescher Laboratory of the Max Planck Society; Tübingen, Germany; [4]Computational Biology Center; Sloan-Kettering Institute; New York, NY USA

Key words: RNA-seq, alternative splicing, antisense transcription, protocol, transcriptomes, *C. elegans*, human

Abbreviations: MISSSL, multiple insert size strand-specific library; PAGE, polyacrylamide gel electrophoresis; RMSE, root mean square error; SS-PE, strand-specific, paired-end; UTRs, untranslated regions

Deep sequencing of transcriptomes allows quantitative and qualitative analysis of many RNA species in a sample, with parallel comparison of expression levels, splicing variants, natural antisense transcripts, RNA editing and transcriptional start and stop sites the ideal goal. By computational modeling, we show how libraries of multiple insert sizes combined with strand-specific, paired-end (SS-PE) sequencing can increase the information gained on alternative splicing, especially in higher eukaryotes. Despite the benefits of gaining SS-PE data with paired ends of varying distance, the standard Illumina protocol allows only non-strand-specific, paired-end sequencing with a single insert size. Here, we modify the Illumina RNA ligation protocol to allow SS-PE sequencing by using a custom pre-adenylated 3' adaptor. We generate parallel libraries with differing insert sizes to aid deconvolution of alternative splicing events and to characterize the extent and distribution of natural antisense transcription in *C. elegans*. Despite stringent requirements for detection of alternative splicing, our data increases the number of intron retention and exon skipping events annotated in the Wormbase genome by 127% and 121%, respectively. We show that parallel libraries with a range of insert sizes increase transcriptomic information gained by sequencing and that by current established benchmarks our protocol gives competitive results with respect to library quality.

Introduction

Deep transcriptome sequencing (RNA-seq) is a powerful method that aims for parallel quantitative and qualitative analysis of all coding and many non-coding RNAs in a sample. Although relatively new, RNA-seq has already broadened our understanding of transcriptome complexity in both prokaryotes and eukaryotes.[1-4] RNA-seq has many advantages over its predecessor, microarrays, not least that it can be used for any organism regardless of the availability of a high-quality genome and detailed transcriptome annotation,[5-8] and has a wider dynamic range.[9]

Many, if not most, eukaryotic genes also give rise to natural antisense transcripts that overlap with the coding mRNA.[10] While some natural antisense transcripts are protein-coding, others are thought to have an important regulatory role. However, if sequencing data are not strand-specific, deconvolution of the data to assess coding mRNA vs. natural antisense transcript levels is difficult, if not impossible (e.g., in the case of antisense transcription vs. retained introns). To this end, strand-specific sequencing data are invaluable, for instance, for natural antisense transcript annotation.

While next generation sequencing platforms like Illumina can produce a high depth of coverage, the length of Illumina reads remains relatively short at around 80–150 bp. Therefore to elucidate information on splice variants, such as whether two variably spliced non-adjacent exons are coordinately retained, paired-end information, that is, two convergent reads from the same RNA molecule, can be very informative. This is especially important due to the high level of alternative splicing in many organisms (42% to 100% of multi-exon transcripts[3,11]). Extreme examples of alternative splicing are the 95–100% of human multi-exonic genes with more than one isoform,[12,13] and a Drosophila transcript with a remarkable 38,016 possible isoforms.[14] To identify specific isoforms in complex scenarios such as above, the spacing of the paired-end reads may be of critical importance.

With the rapidly decreasing cost of high-throughput sequencing, many methods have been developed for transcriptome sequencing over the past few years. These methods vary in their utility for single-end vs. paired-end sequencing and in their retention of strand-specific information. While the standard Illumina RNA-seq library preparation method is highly robust and allows for paired-end sequencing, it does not retain strand

*Correspondence to: Gunnar Rätsch; Email: ratschg@mskcc.org
Submitted: 09/29/11; Revised: 02/09/12; Accepted: 02/11/12
http://dx.doi.org/10.4161/rna.19683

information (based on the Illumina mRNA-seq kit protocol, catalog # RS-100-0801), so will not be the method of choice for all applications. Therefore a number of groups have developed methods for library preparation that retain strand-specificity, some of which are also suitable for paired-end sequencing.[15-17]

Levin et al.[18] recently compared several methods available for transcriptome library preparation and developed a series of criteria by which to assess library quality. Their key criteria include strand-specificity, evenness and continuity of coverage, accuracy compared with other methods of expression profiling and library complexity. According to the Levin comparison, the top two approaches for library construction are dUTP second-strand marking[15] and the Illumina RNA ligation protocol.[19] The dUTP second-strand marking method requires less handling of RNA, but is more expensive than the Illumina RNA ligation protocol according to the calculations of Levin et al.[18] A major limitation of the Illumina RNA ligation protocol is the inability to perform paired-end sequencing, although Levin et al.[18] note that modifications to overcome this drawback should be possible.

In this work, we first assessed the utility of differing library insert sizes (defined as the total fragment length between adaptors) for identification of alternative transcripts from *C. elegans* and human genes based on a computational model. Furthermore, we have adjusted the Illumina RNA ligation protocol to allow for paired-end sequencing by using a pre-adenylated DNA oligonucleotide as the 3' adaptor as opposed to an RNA adaptor, a modification recently confirmed to increase library quality for small RNAs.[20] Additional modifications reduce the number of gel purification steps to facilitate handling and allow for the parallel construction of multiple libraries with different insert sizes from the same sample. This latter modification allows for characterization of short transcripts as well as more complex alternative splicing events in longer transcripts from the same sample. We refer to our protocol as the Multiple Insert Size Strand-Specific Library (MISSSL) method. We created a test library with different insert sizes for *C. elegans*, compared the library quality with the libraries analyzed in Levin et al.[18] and analyzed alternative splicing and antisense transcription.

Results

Based on the results of computational analyses, we have developed a paired-end, strand-specific protocol for Illumina sequencing of transcriptomes, where we constructed multiple libraries, in parallel, with a range of insert sizes. We applied this protocol to the model nematode, *C. elegans*, and show that in practice this simultaneously increases the information gained on alternative splicing and antisense transcription. Additionally, the quality of data produced by this method compares favorably to current benchmark protocols.

Libraries with different insert sizes from the same sample enhance analysis of alternative splicing events. While a sufficient depth of single-end sequencing reads may be adequate in determining gene expression levels, only a fraction of the reads from genes with multiple isoforms will enable the read-generating isoform to be unambiguously identified.[21] Previous theoretical analysis has demonstrated that only paired-end as opposed to single-end sequencing of transcriptome libraries from short-read technologies (e.g., Illumina, Helicos, Solid and Ion Torrent) can be used to reliably reconstruct the majority of splice isoforms in an organism.[21] The utility of the sequencing reads will vary depending on the number of isoforms of a gene, the number of variant exons, read depth and the spacing between the paired reads. In particular, for paired sequencing reads to be associated with a specific splice variant, they must cover unique splice junctions or both ends must fall into variant exons that distinguish the isoform. Paired-end reads of a given spacing are especially important for identification of isoforms from genes with two or more variant exons.[21]

A previous analysis of paired-end sequencing with regards to multiple isoforms of a small set of human genes found that paired-end sequencing could identify all but a handful of the isoforms.[21] We developed a probabilistic model of the utility of paired-end reads with variable parameters for average library insert size (taken as the number of nucleotides between the adaptors), distribution of insert sizes, length of sequenced ends, coverage and whether paired ends or single ends were sequenced. We initially focused on annotated genes of the model nematode *C. elegans* and human with three or more isoforms, as these transcripts will be harder to distinguish with short reads from deep sequencing. To identify whether splicing of two alternatively used regions is coordinately regulated, paired-end reads that map to each of the alternatively spliced regions will typically be required.

To assess the utility of read pairs spaced varying distances apart, we calculated the minimum insert size between the paired-end reads that is required to uniquely identify a given isoform for genes with three or more annotated isoforms (**Fig. S1**). Among all *C. elegans* isoforms that are identifiable only by paired-end reads up to 1 kb apart, 15% (40 isoforms of 260 total) are identifiable by overlapping pairs of reads (based on a read length of 76 bp), while a further 28% of isoforms are identifiable by libraries with up to 148 bp between the paired reads (allowing for ±12.5% variability in library insert length; 73 of 260 isoforms; this corresponds to the most common total insert size of 300 bp including 2 x 76 bp reads). 43% of isoforms are not identifiable by reads up to 148 bp apart, but are identifiable with reads from 148 bp up to 648 bp apart (111 of 260 isoforms), which corresponds to the insert size range of 300–800 bp that we used in our libraries. Compared with higher eukaryotes such as humans, the level of alternative splicing in *C. elegans* is relative lowly (95–100% vs. 25% of genes having multiple isoforms[12,13,23]). Therefore, we applied our model in parallel to human genes with three or more isoforms to test the utility of parallel library construction in a more complex organism. For humans, 34% of isoforms (1,917 isoforms of 5,688 total) are identifiable by overlapping paired reads, a further 36% by reads up to 148 bp apart (2,037 of 5,688 isoforms), and an additional 26% with reads from 148 bp up to 648 bp apart (1,460 of 5,688 isoforms). Hence, a significant fraction of complex alternative isoforms can only be resolved with library fragment sizes larger than the typical 300 bp.

We modeled the gain in alternative splicing data for (1) a single library with a given insert size (allowing for ±12.5% variability

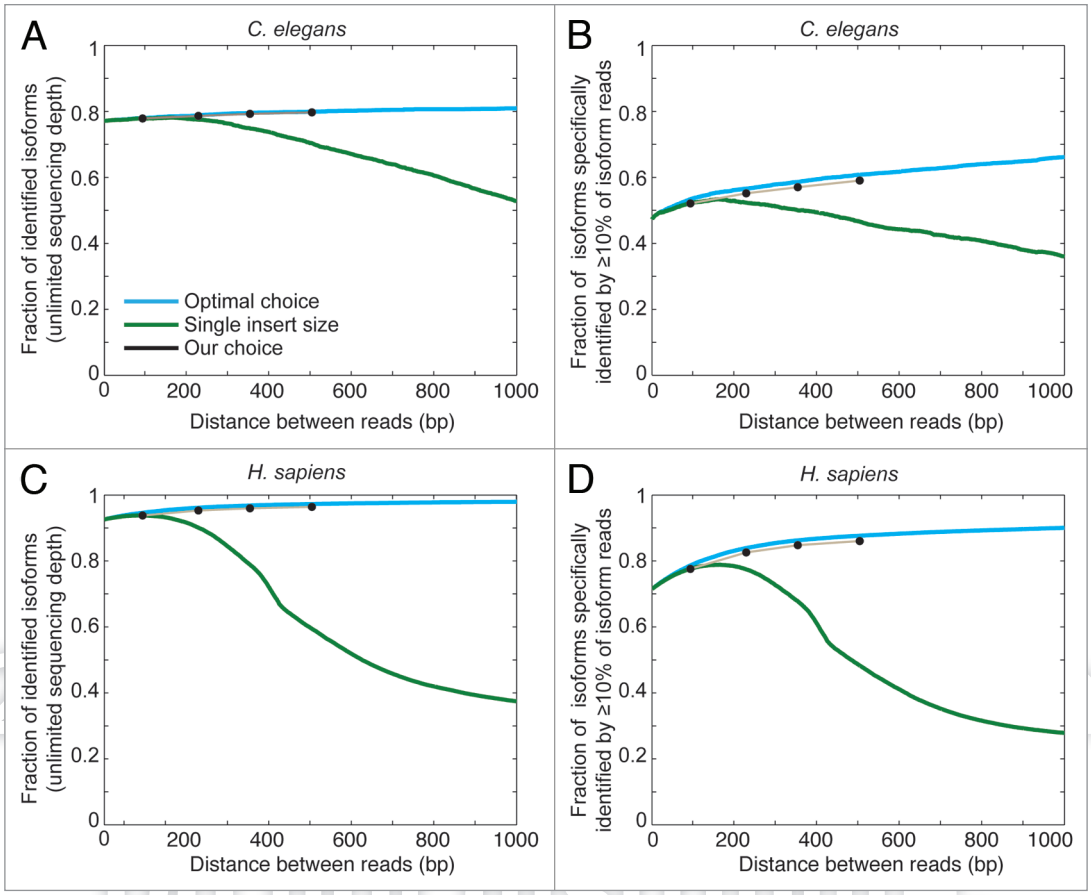

Figure 1. Theoretical insert size utility. All theoretically possible paired-end reads were generated and the ability of paired-end reads a given distance apart to distinguish gene isoforms was plotted for all genes with two or more isoforms. We examined libraries with (1) a single defined distance between the reads (±12.5%; green), (2) the combination of all possible read distances below a given maximum ("optimal choice"; blue) and (3) our selection of four evenly distributed read distances ("our choice"; black) for (A) *C. elegans* and (C) humans given an unlimited sequencing depth. We also modeled the fraction of isoforms that could be specifically identified by at least 10% of reads generated by that isoform, given the conditions above for (B) *C. elegans* and (D) humans. n = 5,718 for *C. elegans* and n = 66,654 for humans.

in library insert length, with a minimum insert size of 150 bp), (2) all possible insert sizes from 150 bp (~2 times read length) to the indicated length with the length of each insert known precisely ("optimal choice") and (3) our selection of four parallel libraries with insert sizes uniformly distributed around the means ±12.5% (means of 215 bp, 350 bp, 475 bp and 625 bp). The latter distribution of insert sizes was selected based on our experience of cluster generation with version 4 Illumina kits and an evenly-spaced insert sizes of the libraries. Our theoretical analysis of library insert sizes required to gain information on alternative splicing in genes with two or more isoforms demonstrated that no single insert size can be informative for all splicing events (**Fig. 1A and C**). We observed that all isoforms that require paired-end sequencing to be identified are derived from genes with three and more isoforms. A small insert size (i.e., 200–300 bp) is informative of a greater fraction of alternative isoforms, however, 147 isoforms from *C. elegans* and 1,734 isoforms from humans require a longer insert size to be identifiable. This is likely an underestimate of the information gained, as the overlap in predicted splice variants between RNA-seq data sets and transcriptome annotations is low (7.5% for Arabidopsis[22]), indicating that many splice variants are yet to be annotated. We checked whether our choice of the insert size variability made a difference and found that doubling the variability in insert size had a very minor impact on our results. Therefore, a small but significant increase in isoform information can be gained by generation of transcriptome libraries with differing insert lengths.

Our initial modeling determined if a specific isoform could be theoretically identified, given unlimited sequencing depth. However, efficiency of detection is also a practical issue since some genes and isoforms are lowly expressed. Hence, only a few relevant reads will be present in a transcriptome library. We therefore adjusted our model to determine whether more than 10% of reads originating from a transcript could be assigned uniquely to a particular gene isoform. Under these conditions, the fraction of identified isoforms was reduced and the information gain through using parallel libraries with differing insert sizes was more pronounced: (A) an 11% increase in information gain from a single library of optimal insert size compared with all possible read distances up to 1 kb for both humans (n = 66,654) and *C. elegans* (n = 5,718), and (B) an increase from 3,053 isoforms (53%) identified by a single library of optimum insert size

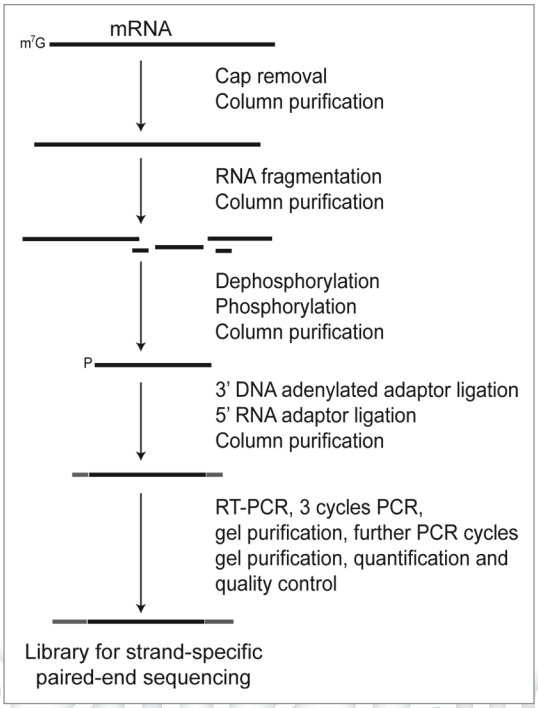

Figure 2. Strand-specific, paired-end mRNA sequencing. A brief overview of the steps involved in the MISSSL protocol. Full details can be found in Materials and Methods.

to 3,555 isoforms (62%) through use of our choice of four insert sizes for *C. elegans* (see **Fig. 1B**). The increase in information was similar for human isoforms with comparable conditions [**Fig. 1D**; 79% (52,571 of 66,654 isoforms from 12,179 genes) increasing to 87% (58,175 of 66,654 isoforms), with a total of 60,017 isoforms (90%) identifiable if using all possible insert sizes with paired reads up to 1 kb apart]. Moreover, while the optimal overall read distance is ~148 bp (insert size ~300 bp), we found 118 *C. elegans* isoforms and 1,194 human isoforms can be identified efficiently only with longer insert sizes. These isoforms would be not efficiently identifiable (less than 1% of reads) with an insert size of at most 300 bp, but would be efficiently identifiable (more than 10% of reads) with larger insert sizes (where the optimal insert size may be different for each transcript). This highlights the value of generating multiple libraries with different insert sizes, especially for transcriptomes with high levels of alternative splicing, as many analyses require a sufficiently high number of informative reads to produce high-confidence results.

Subsequent to test library production, we modeled how much variance in insert size can be tolerated while still allowing for unambiguous isoform identification in humans. There is a tradeoff between the total range of insert sizes covered and the information that can be gained through low insert size variance in a single library. We found that the number of efficiently identifiable transcripts drops significantly when the insert size variability is above ±50 bp (data not shown). The median exon length of about 130 bp can explain this observation. Some isoforms can tolerate no uncertainty in insert size; these include cases with closely spaced alternative splice junctions that will be best identified by reads that span the alternative region. For the remaining isoforms, most will be identified as long as insert size variance does not exceed 50 bp.

Modifications to the Illumina RNA ligation protocol. We modified the Illumina RNA ligation protocol in a number of ways to allow paired-end strand-specific sequencing and to improve handling of samples (**Fig. 2**). Our most significant modification to the Illumina RNA ligation protocol is use of a custom 3' adaptor, prepared by pre-adenylation[24] of a DNA oligonucleotide. This 3' adaptor is an inexpensive and versatile modification that allows fusion of the required sequences for paired-end sequencing[25] to RNA in a strand-specific manner. Using this modification, the MISSSL library can then be sequenced with the Illumina small RNA and second paired-end sequencing primers,[26] using published methods.[25] We designed the oligonucleotide based on the Illumina paired-end adaptor, however any sequence could theoretically be used for the 3' adaptor, with the complement as a sequencing primer, so long as the minimal sequence for annealing and amplification on the Illumina chip is added during PCR amplification.

Additional modifications include the use of Ambion RNAqueous Micro columns for library clean-up steps due to their small elution volume and high efficiency of RNA recovery. Using column-based purification allows for the removal of excess adaptor that has not been ligated to the RNA and facilitates elution in small volumes for immediate use in downstream preparation steps. This reduces handling time by eliminating multiple precipitation, centrifugation and resuspension steps.

To construct multiple libraries of varying insert size in parallel, we conducted three rounds of PCR after RT-PCR to convert the ssDNA to dsDNA. This DNA was then separated on an agarose gel and the DNA extracted from gel slices with the desired insert sizes. Further rounds of PCR were then used to amplify the library before quantification and sequencing. The intermediate size selection step is necessary due to the bias of PCR amplification toward lower insert sizes, but we feel this is a minor increase in handling time compared with the potential information gain.

Analysis of alternative splicing and antisense transcription in *C. elegans*. In a test of our protocol with mRNA from the nematode *C. elegans*, we used a short RNA fragmentation time to generate a range of fragment sizes and constructed, in parallel, four transcriptome libraries with insert sizes averaging 215, 350, 475 and 625 bp as measured by Agilent Bioanalyzer and 155, 320, 480 and 595 bp as estimated from read alignments. The standard deviation in library insert size was approximately 50 bp as estimated from read alignments. Insert size range is largely influenced by the width of the excised gel slice, and therefore could be further reduced.

The utility of this library construction approach for more comprehensive analysis of splice variants, with different insert sizes benefiting the interpretation of alternative splicing, is further demonstrated in **Figure 3**. In the case shown in **Figure 3**, there are five annotated isoforms of gene *mdt-28* listed in WormBase. In both cases where the longer exon 3 is included in the isoform, it is paired with the short exon 7 rather than longer variant exons 8 or 9 (**Fig. 3A**). Plotting read coverage from all

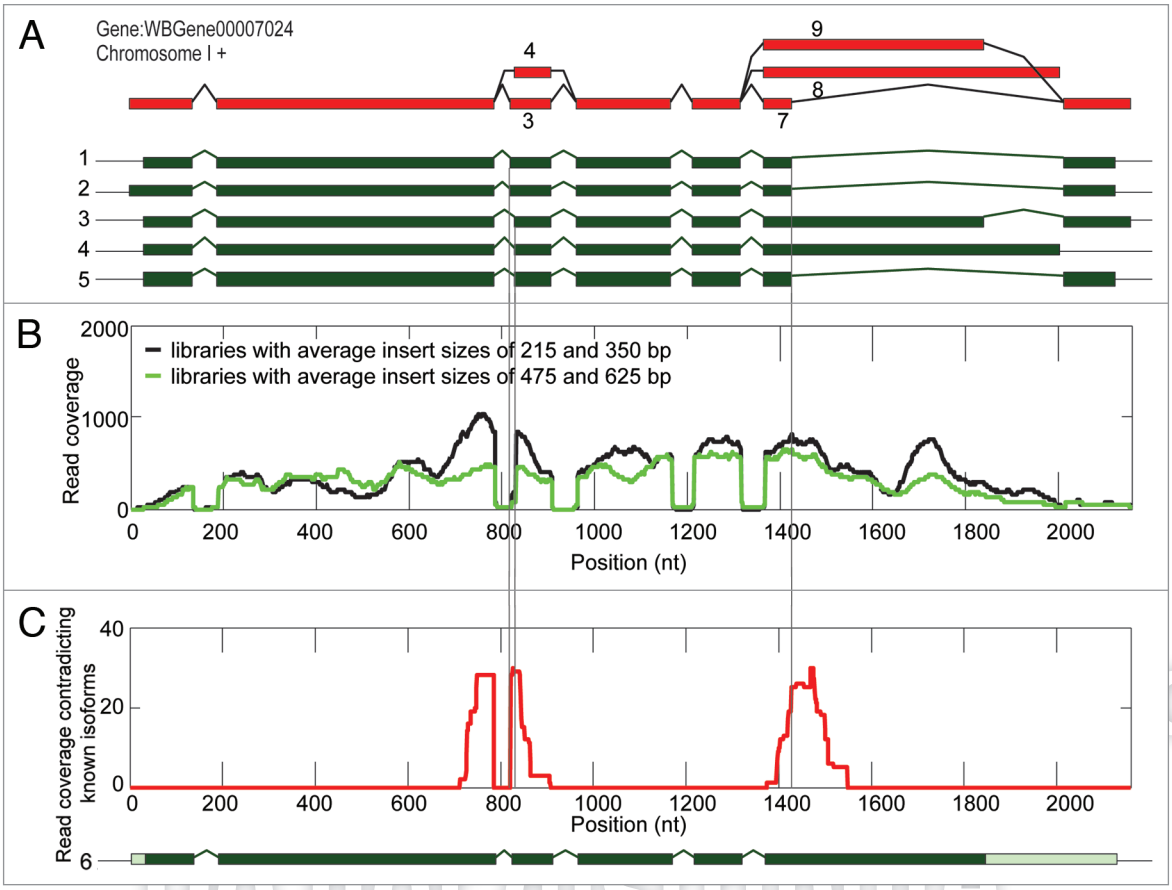

Figure 3. Example of insert size utility. (A) The WormBase splicing model and the five annotated isoforms of gene WBgene00007024 (*mdt-28*) are shown with numbering for selected variant exons. (B) Read coverage of *mdt-28* exons and introns from libraries with average insert sizes of 215 and 350 bp (black) and 475 and 625 bp (green). (C) Reads from the 475 and 625 bp insert libraries where one pair of the read covers exon 3 and the second read extends beyond exon 7, indicating that it must come from alternative exons 8 or 9, are shown in red. The new isoform splicing pattern is indicated below with ambiguous exons indicated in light green.

the libraries, it is clear that while isoforms with exon 4 are predominant, isoforms containing exon 3 are also present (**Fig. 3B**). Using read coverage from the libraries with an average insert size of 475 and 625 bp, we can show that approximately 6.5% of the transcripts contain exon 3 paired with either exon 8 or exon 9 (peak of 30 transcripts at 827 bp in **Fig. 3C**/peak of 460 transcripts at 827 bp for green line in **Fig. 3B and C**). Therefore, in this case the use of a longer insert size was able to confirm presence of a new isoform of the gene where shorter insert sizes would not have been informative.

To investigate the genome-wide contribution of a multiple library approach to the identification of novel transcript isoforms, we calculated the number of paired read alignments in each library specifically mapping to each annotated isoform. We also counted a novel isoform generated as a path through a splicing graph as more consistent with our paired-end coverage if it explained at least 10 read-pairs more than any annotated isoform. We found 993 novel isoforms with paired-end read support. The overlap between each combination of library insert sizes was calculated (**Fig. S2**) and demonstrates that while 441 of 993 (44%) novel isoforms are supported by all libraries, a substantial fraction of isoforms were only supported by a subset of the libraries with different insert sizes. This illustrates that multiple insert size libraries facilitate the detection of novel isoforms.

We examined the proportion of annotated alternative splicing events, as defined by single exon skipping and intron retention events, which could be verified by our libraries. The *C. elegans* genome annotation (WS199) contains a total of 802 intron retention and 517 exon skipping events. We detected 201 of these known intron retention events (25%) and 343 of the annotated exon skipping events (66%) in our RNA-seq data, while we identified 1,021 novel intron retention and 630 novel exon skipping events (**Fig. 4**; annotations available from **Sup. website**). A similarly small overlap of RNA-seq derived alternative splicing events compared with annotated events was previously seen for Arabidopsis tiling arrays and RNA-seq data from a broad panel of tissues.[22,27] The low level of overlap between annotated and detected alternative splicing events is likely due to a number of factors: splicing events may be annotated in databases regardless of how rarely they occur (or they may have been a mis-splicing event[28]); splicing may be tissue-, life stage- or cell line-specific and below detection limits in the whole-organism libraries prepared here; and splicing may be environmentally regulated such that a given event is not present under our growth conditions.

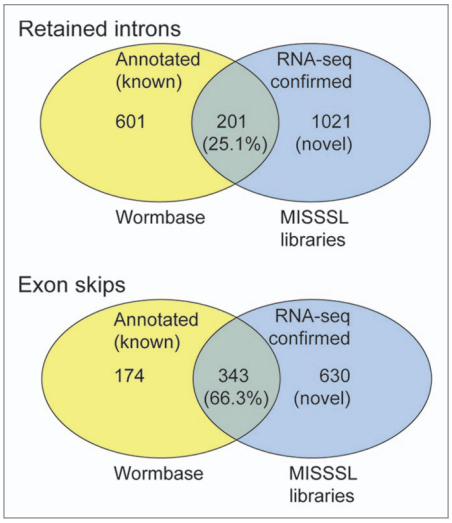

Figure 4. Overlap in annotated and detected alternative splicing events. The overlap between alternative splicing events annotated in WormBase (yellow) and those verified by our libraries (blue) is shown for single retained introns and skipped exons. The percentage of previously annotated splicing events that we were able to confirm is indicated. From an RNA-seq-based splicing graph extension and analysis, we were able to identify many novel alternative splicing events.

We further compared the exon skipping events identified from our data set to recent deep-sequencing data sets that combined encompass approximately 14x as many RNA-seq reads as we generated.[23,29,30] These analyses were performed with the exon skipping data since intron retention events are indistinguishable from incomplete mRNA processing with this data. While the majority of exon skipping events we detected were also annotated in these data sets (**Fig. S3 and Table S1**), 8% of our exon skipping events were either not present or matched to unconfirmed events. From the latter subset of 79 genes, we tested 15 exon skipping events by RT-PCR and were able to confirm all exon retention events and 13 of the exon skipping events (**Table S1**).

Antisense transcription, at least in humans and mice, is most common in the 250 bp upstream of transcriptional initiation and within 1.5 kb downstream of stop codons.[10,31-33] However, given the unusually low level of antisense transcription in nematodes,[34] these generalities may not apply. We examined the distribution of sense and antisense reads in relation to protein-coding genes. As expected we found enrichment for sense transcription in exons (**Fig. 5A**). We also found above-background levels of transcriptional activity in flanking intergenic regions, which we attribute to incomplete annotations of 5' and 3' untranslated regions (UTRs; data not shown). Moreover, we detected enrichment for antisense transcription in the last exon (~2-fold), 3' untranslated regions (~3-fold) and 250 bp downstream (~6-fold) of *C. elegans* protein coding genes (**Fig. 5B**). We observed that most antisense reads (~97%) originate from regions around protein coding genes, with coverage at least two to three times higher than in pseudogenes, transposable elements, non-coding genes and structural genes (e.g., tRNAs and rRNA; data not shown). Manual inspection of gene-flanking regions with the highest levels of antisense read coverage showed that in some cases this may be due to incomplete annotation of neighboring genes, with read coverage sometimes extending over 1 kb beyond the annotated gene model. While our comparison was to WS199/200, some of these antisense regions have since been correctly annotated by the integrated transcripts of modENCODE,[30] however other transcribed regions have either been only partly extended, or may have been incorrectly extended due to the lack of strand-specific data (e.g., WBGene00018833, WBGene00021602, WBGene00008032 and WBGene00012713; **Table S2**). Moreover, we found a low correlation between the number of sense and antisense reads for protein coding genes (Spearman correlation 26%), as previously observed for yeast.[35]

We were interested in whether genes associated with particular biological functions or processes were more frequently associated with antisense transcripts. An analysis of gene ontology categories for genes with antisense transcription showed highly significant enrichment of several categories. By function, nucleotide binding ability (p-value = 4×10^{-5}) and catalytic activity (p-value = 4×10^{-8}), along with its subcategory of hydrolase activity (p-value = 3.4×10^{-6}), were most significantly enriched. Enrichment of catalytic activity and nucleotide binding was also previously observed in a data set of convergent, overlapping transcripts from human, fly and mouse.[36] When analyzed by process, genes for several biological processes were highly significantly enriched for antisense transcripts including; metabolic processes (p-value = 2×10^{-6}), localization establishment (p-value = 2.4×10^{-7}), cellular processes [p-value = 3.3×10^{-19}; with sub-category transport (p-value = 1.2×10^{-6})], developmental processes (p-value = 1.2×10^{-23}), reproduction (p-value = 1×10^{-10}) and locomotion (p-value = 3.8×10^{-7}).

Quality assessment of the MISSSL method. We analyzed library quality using the pipeline developed by Levin et al.[18] for combined reads from the two libraries with insert sizes averaging 215 bp and 350 bp. We also adjusted the number of reads used in the pipeline to account for the larger *C. elegans* genome and transcriptome compared with the *S. cerevisiae* samples used by Levin et al. (see Materials and methods). Compared with the Levin data, we find that the MISSSL protocol is comparable to the top performing methods of Illumina RNA ligation and dUTP second-strand marking in terms of percentage unique single and paired reads [for instance, unique first strand reads 38.2% (MISSSL), vs. 35.8% (dUTP) and 38.5% (Illumina RNA ligation); **Fig. 6A and Table 1**]. These parameters are important as they indicate the library complexity and will be lower if there is significant clonal sequencing due to the use of too many PCR cycles during library amplification. The average coefficient of variation, a measure of evenness of read coverage across the genes, was also comparable between MISSSL, dUTP second-strand marking and the Illumina RNA ligation methods (**Fig. 6B and Table 1**). In addition, we observed a lower weighted average number of segments per gene (**Fig. 6B**) hinting at a more uniform read coverage.

Our libraries have a lower number of antisense strand reads (0.15% vs. 0.55/0.58% for the Illumina RNA ligation and dUTP

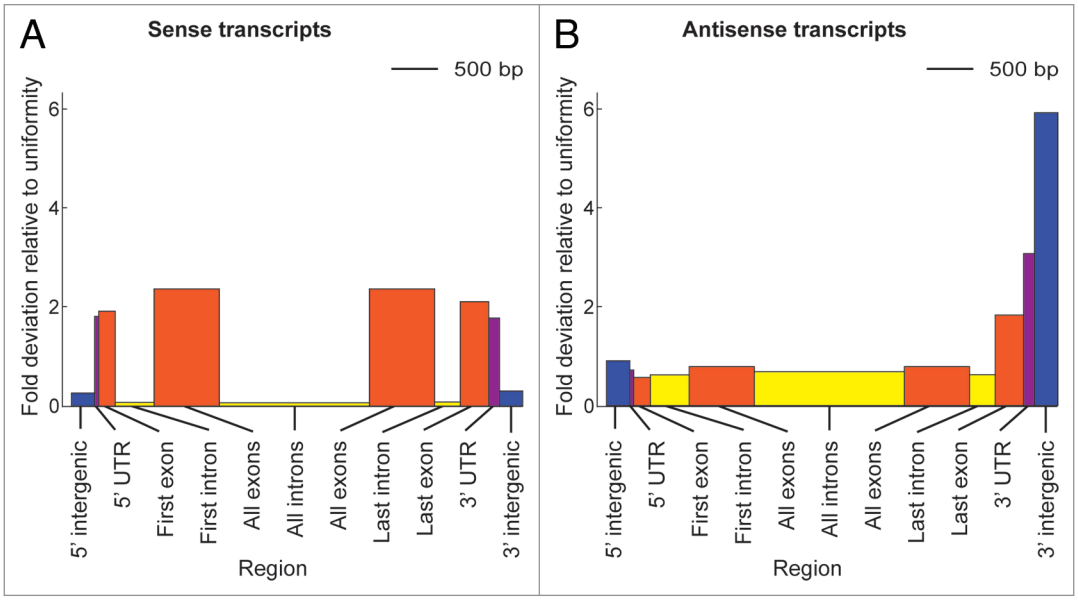

Figure 5. Location of sense and antisense transcription in relation to protein-coding genes. The abundance of sense and antisense reads in relation to protein-coding gene structure was calculated for regions: up to 250 bp upstream of annotated transcriptional start sites; within 5' UTRs; within first, last or all exons; within first, last or all introns; within 3' UTRs; and up to 250 bp downstream of annotated genes. The distribution of (A) sense reads over each described region is shown for all libraries in comparison to the distribution of (B) antisense reads.

second-strand marking methods respectively **Fig. 6C**), however this is likely due to the lower degree of antisense transcription in *C. elegans* compared with *S. cerevisiae*, previously reported as 0.5% and 8% respectively.[34,37] We believe the higher percentage of reads in non-exonic regions in *C. elegans* (2.81%) as compared with *S. cervisiae* (0.57–0.85%; **Fig. 6D**) to be due to less complete annotation of the more complex *C. elegans* transcriptome,[38] as was observed for the antisense reads. We further compared the reads to the modENCODE transcribed regions,[30] which, in conjunction with the WS200 annotation, reduced the non-exonic reads to 2.73% of the data set. Inclusion of pseudogenes in the definition of exonic reads reduced the percentage of exonic reads to 2.57% and 2.51%, respectively, for the WS200 annotation and modENCODE plus WS200.

We also compared gene expression estimates based on our library to those from tiling microarray data[39] to assess the quantitative ability of MISSSL. Our sequenced libraries were prepared with RNA from nematodes at mixed developmental stages, however no tiling array data was available from a comparable RNA sample. Although the microarray data are not from the same RNA sample, and incorporates only a subset of the *C. elegans* developmental stages that we used in the construction of our library, we found a respectable agreement between the data sets (R^2 = 0.76, see **Fig. 6E**, also showing the evaluation by Levin et al. on *S. cerevisiae* data for comparison). The lower correlation, compared with Levin et al. for yeast, is likely due to expression differences in the different developmental stages that were used to generate the microarray and RNA-seq data sets. In summary, using the parameters determined by Levin et al. data generated using our method of paired-end, strand-specific sequencing is overall of comparable quality to current benchmark methods.[18]

Discussion

In this paper, we model the utility of libraries with multiple insert sizes from the same mRNA sample in the identification of splice variants and develop a method for paired-end, strand-specific sequencing with multiple parallel libraries. We assess library quality and identify new alternative splicing and natural antisense transcription events. Our method is based on the Illumina RNA ligation method, but extends the practicality of this method to paired-end reads. As calculated by Levin et al.[18] there were quality parameters where the Illumina RNA ligation method out-performed the dUTP second-strand marking method; the dUTP method produced fewer mapped reads (63% vs. 70% for the dUTP and Illumina RNA ligation methods respectively, Levin et al. **Sup. Table 1**), and fewer uniquely mapping first reads (42% vs. 59% respectively; Levin et al. **Sup. Table 1**). The number of segments per kb were significantly lower with our method compared with the dUTP method (1.48 vs. 1.85 segments per kb) indicating more even read coverage. The preparation time for the Illumina RNA ligation and dUTP second-strand marking methods are comparable,[18] however with limited gel purifications steps and workflow optimization, library preparation with the MISSSL method can be completed in 2–3 d. Furthermore, our results demonstrate that the MISSSL method can generate strand-specific, paired-end libraries of comparable quality to the current benchmarks of the dUTP second-strand marking and Illumina RNA ligation methods.

Importantly, we have incorporated into our protocol the ability to make multiple libraries in parallel, which differ by insert size, from a single RNA sample. Our probabilistic models show that in a complex organism some genes with two or more isoforms can only be identified by paired-end reads with an insert

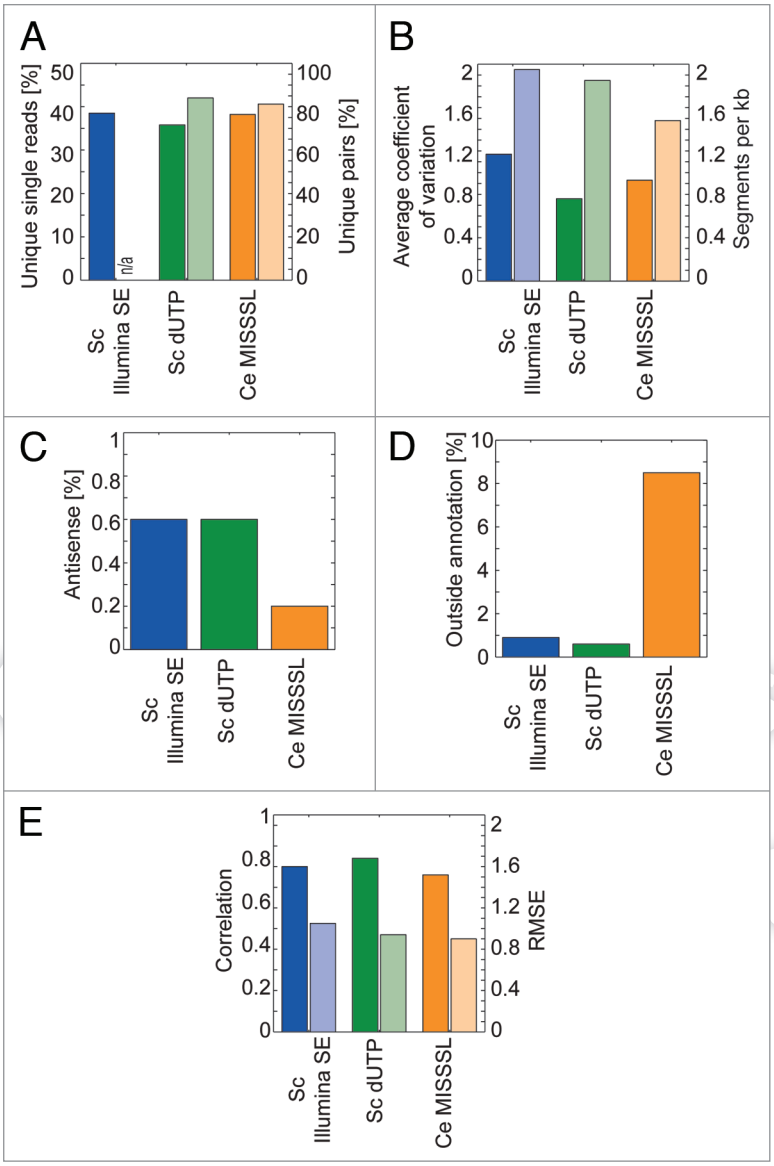

Figure 6. Comparison of MISSSL to the Illumina RNA ligation and dUTP methods. (A-E) Comparison of the results of our strand-specific, paired-end RNA sequencing protocol to the published analyses of the Illumina RNA ligation and dUTP methods [see details in Methods and Levin et al.[18] for definitions of the criteria]. The parameter on the left y-axis is shown in deep shading, while the parameter on the right y-axis is shown in light shading. Sc Illumina SE and Sc dUTP indicate the Illumina RNA ligation single-end and dUTP paired-end library data respectively from *S. cerevisiae* as presented by Levin et al.[18] while Ce MISSSL denotes the paired-end data from *C. elegans* that we present here. n/a = not applicable.

size greater than 300 bp (*C. elegans*: 147 isoforms; human: 1,734 isoforms). This is likely a large underestimate of information gain since many splice variants remained unannotated in WS199/200 (overlaps between RNA-seq data sets and transcriptome annotations have been reported as low as 7.5% for Arabidopsis;[22] see further discussion below). We demonstrate the practical utility of combining information from multiple insert sizes with an example from *C. elegans*, moreover our model demonstrates that the gain in information from making multiple parallel libraries differing in insert size will be greater for organisms with a higher level of alternative splicing. Although we sequenced our libraries in individual lanes to ensure deep coverage and facilitate analysis, barcoding and mixing of libraries after quantification for sequencing in a single lane would also be feasible. We conclude from our analyses that insert size information will be of most use when there is a low level of insert size variability (not more than 50 bp). From computational modeling on human data (not shown), we can recommend libraries with 0 bp, 150 bp, 400 bp and 800 bp between the paired reads, based on insert size variability of 50 bp. (For this recommendation, we give insert sizes between the reads as Illumina read length capabilities are steadily increasing). We also observe that the dUTP method could similarly be adjusted to allow for parallel construction of multiple libraries of varying insert size.

Although Vivancos et al. developed a similar paired-end sequencing protocol, a significant advantage of the MISSSL method is that it involves no PAGE gel purification steps (eliminating time-consuming RNA/DNA recovery steps), since it is based on the Illumina RNA ligation protocol[19] rather than the original RNA ligation protocol. The elimination of PAGE gel purification steps means shorter nucleic acid recovery times. Since PAGE allows greater resolution of nucleic acids, any group requiring reduced insert size standard deviation could adapt MISSSL for PAGE purification. Library insert size can also be modified as needed with empirical adjustment of RNA fragmentation time (conditions here give a range of 130–730 bp). If only smaller insert sizes are desired, fragmentation time could be increased and starting RNA reduced. Differences in ligation efficiency due to RNA sequence and/or secondary structures are the major factor in library bias for small RNAs.[20] This may partly explain why the Illumina RNA ligation protocol generally performed better than the RNA ligation protocol in the analysis by Levin et al.[18]

For short RNA species, where fragmentation is not required to generate small lengths for cloning, current read lengths may allow for complete sequencing of transcripts with a paired-end protocol. Since the dUTP second-strand marking method partly relies on random priming to generate a relatively even read coverage across transcripts, a ligation-based protocol that preserves 3' end information may be more appropriate for sequencing short non-polyadenylated transcripts, such as those produced by RNA polymerase V in plants.[40] The MISSSL method preserves 3' end information.

Antisense transcription plays an important role in gene regulation.[10] A recent report using Tag-seq data from *C. elegans* confirmed previous reports of abundant antisense transcription from five of the 12 mitochondrial genes but only less abundant antisense signal from a small number of nuclear genes.[41,42] Here our strand-specific data shows that antisense transcripts are most commonly located in the 250 bp downstream of genes, the 3' UTRs and in the last exon, similar to the previously noted distribution

Table 1. Comparison of three library protocols

Library	Sc Illumina SE	Sc dUTP	Ce Illumina PE
total number of reads	2,500,018	2,500,019	6,773,929[a]
unique read starts	962,917	895,698	2,584,467
unique read starts [%]	38.5	35.8	38.2
unique pairs [%]	n/a	84.0	81.2
number of reads on expected strand	2,293,081	2,319,635	6,178,675
number of reads on antisense strand	13,837	14,609	10,247
antisense reads [%]	0.55	0.58	0.15
number of reads outside exonic annotations	21,374	14,320	190,052
reads outside of exonic regions [%]	0.85	0.57	2.81
total number of reads in single feature regions	2,328,292	2,348,564	6,766,427
average coefficient of variation (CV) for top 50% expressed genes	1.17	0.76	0.93
segments per kb	1.95	1.85	1.48
weighted average of segments per gene	2.61	2.48	2.10
number of segments normalized by mean transcript length [%]	0.19	0.19	0.15
correlation to microarrays	0.80	0.84	0.76
RMSE to microarrays	1.05	0.94	0.9

Illumina RNA Ligation with single end reads in *S. cerevisiae* (Sc Illumina SE), dUTP in *S. cerevisiae* (Sc dUTP), and our proposed paired-end strand-specific library protocol applied to *C. elegans* (Ce Illumina PE). [a]The number of reads used in our analysis was scaled according to transcriptome size for *C. elegans* vs *S. cerevisiae*.

in flanking intergenic regions for humans and mice.[10,31-33] We observe enrichment for antisense transcription in specific gene ontologies. To our knowledge, this is the first genome-wide analysis of antisense read distribution for *C. elegans*.

Two recent studies of the *C. elegans* transcriptome have characterized a large number of novel alternative splicing events, many of which are developmentally regulated and rare.[23,29] Our study would not have detected many of these events due to use of mixed-stage nematodes and the stringent requirement for five supporting reads or 90% retained intron coverage (indeed, we failed to detect 1,465 exon skipping events described in ref. 23). Nonetheless, our data identified 127% and 121% more intron retention and exon skipping events, respectively, compared with WS199/200. While these studies[23,29,30] utilized extensive data sets, and in one case also incorporated strand-specific sequencing, the deep sequencing reads were short single-end reads [36 and 39 bp respectively for ref. 29 and 23]. Comparison of the MISSSL data set with two recent alternative splicing data sets[23,29,30] showed a higher level of overlap in exon skipping events compared with WS199/200, indicating the recently improved annotation of alternative splicing in *C. elegans*. Although our RNA-seq data set was approximately 14 times smaller and probably did not reach saturation level for annotation of alternative splicing events,[23] 8% of the exon skipping events we annotated had not previously been confirmed. With improvement of the Illumina platform and the development of transcriptome sequencing methodology, paired reads of up 150 bp are now possible. Longer, paired sequencing reads will facilitate transcriptome characterization as longer reads increase the probability of spanning splice junctions while paired-ends allow identification of coordination for more distant alternative splicing events. Together these developments will increase annotation accuracy of higher genomes, especially those with extensive alternative splicing and antisense transcription.

In this paper we have shown through modeling and data from *C. elegans* that there is a significant advantage, when annotating alternative splicing events, to inclusion of a broader range of insert size lengths than is currently used for transcriptome library preparation. A range of insert sizes, in combination with strand-specific reads, allows analysis of mRNA levels, alternative splicing events and antisense transcripts from a single sample. In conclusion, we believe the MISSSL method for construction of multiple insert size, strand-specific libraries for paired-end RNA sequencing provides a competitive alternative to the currently published protocols and will aid in the further annotation of complex transcriptomes.

Materials and Methods

Insert size utility modeling. We created a probabilistic model to study the influence of various parameters of library construction on the resolution of alternative splicing events from mRNA-seq data. The model was used to calculate the probability of identifying different isoforms of a gene given the insert size, the band width that had been cut out of the gel, the length of the sequenced ends, read coverage and whether paired-end or single-end reads were used. The model for the paired-end reads was created as follows. For a given gene with multiple isoforms, we first determined all possible read pair matches independent of the parameters. We then established which of those read pairs could be mapped back to one and only one isoform. For example, these could be reads that cover an exon or splice junction unique to one isoform. Next, we checked which of the reads that uniquely identified a

particular isoform were in concordance with the parameters of our model. If no such read existed, we reported that this isoform could not be identified with the given parameters.

To estimate how many isoform-specific reads were expected for a given read coverage, we assumed that all paired reads remaining after previous filtering from an isoform were equally likely. We then calculated the fraction of reads which distinguished this specific isoform from other isoforms given all the reads potentially generated from the isoform under the chosen parameters. The reads were distributed according to a binomial distribution $B(n,p)$ with n being the total number of reads from the isoform and p being the fraction described above. This gives $n \cdot p$ reads as the expected number of informative reads. For our models we used the human HG19 annotation[43,44] and the *C. elegans* WormBase (WS199/200) annotation.[45] The program can be obtained from the **Supplemental website**.

Primer and adaptor sequences. (A) 3' adaptor[25,26] P-GAT CGG AAG AGC GGT TCA GCA GGA (C7 amino blocked on 3' end)

(B) 5' adaptor[24] GUU CAG AGU UCU ACA GUC CGA CGA UC

(C) Antisense primer TCC TGC TGA ACC GCT CTT CCG ATC TAT AGT GCA GT for 3' adaptor adenylation

(D) RT primer CTC GGC ATT CCT GCT GAA CCG CTC TTC CGA TCT

(E) PCR primer 1 [25,26] CAA GCA GAA GAC GGC ATA CGA GAT CGG TCT CGG CAT TCC TGC TGA ACC GCT CTT CCG ATC T

(F) PCR primer 2 [25,26] AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CGA CAG GTT CAG AGT TCT ACA GTC CGA.

3' adaptor pre-adenylation. We performed 3' adaptor adenylation largely as described in Vigneault et al. Five μl of 3' adaptor (A; 500 μM) was mixed with 75 μl of antisense primer (C; 100 μM), 0.5 μl 1 M Tris (pH 7.5), 0.3 μl 0.5 M NaCl and 10 μl 1 mM EDTA in a total volume of 115 μl. The adaptors were annealed by heating to 95°C for 3 min then cooling immediately on ice. 40 μl 0.2 M MOPS, 2 μl 1 M $MgCl_2$, 20 μl 0.1 M DTT, 20 μl 0.1 M ATP and 18 μl of T4 DNA ligase (NEB; #M0202M; 2,000 U/μl) were added and the mix incubated at 25°C for 24 h to adenylate the adaptor. The 3' adaptor and antisense oligonucleotide were separated on a 1.5 mM thick 17.5% TBE-Urea polyacrylamide gel at 300 V for 60 min. The 3' adaptor band was excised, and the gel slice fragmented by centrifuging though needle holes in the bottom of a 0.5 ml tube into a 1.5 ml tube. Two volumes of 0.3 M NaCl were added and the adaptor eluted at 4°C overnight with shaking. The mixture was transferred to a SpinX Cellulose Acetate filter (Costar; #CLS8163) to recover the eluate. After ethanol precipitation (-80°C for at least 1 h), the adaptor was resuspended in 20 μl water, quality checked on a 17.5% TBE-Urea polyacrylamide gel, and measured to be at a concentration of 89 μM based on absorbance at 260 nm on a Nanodrop (Thermo Scientific). The 3' adaptor adenylation protocol should provide sufficient adaptor for at least 15–20 libraries. Adaptor should be used at a concentration around 100 μM.

Library construction. Total RNA (150 μg) was isolated with Trizol (Invitrogen; #15596-018) from *C. elegans* (strain N2) of mixed developmental stages and then mRNA purified with the μMACS mRNA Isolation kit following the manufacturer's instructions (Miltenyi Biotec; #130-090-276). After mRNA integrity was confirmed on a Bioanalyzer RNA 6000 Nano Chip (Agilent technologies; #5067-1511), the 5' cap was removed with 10 U of Tobacco Acid Pyrophosphatase (Epicenter Biotechnologies; #T19050) in the presence of 40 U Ribonuclease inhibitor (Fermentas; #EO0381) at 37°C for 90 min. RNA was purified using RNAqueous Micro columns (Ambion; #AM1931) and eluted in 2 x 9 μl, following the manufacturer's LCM protocol with a few modifications: the sample volume was brought to 100 μl using lysis buffer; 1.25 volumes of ethanol were added to the sample unless otherwise noted; and no LCM additive was used.

An amount of 2 μl of fragmentation buffer (Ambion; #AM8740) was used to fragment the RNA for 90 sec at 70°C, followed by addition of 2 μl stop buffer and brief incubation on ice. RNA was purified using RNAqueous Micro columns (Ambion; #AM1931) and eluted in 2 x 8 μl. RNA was dephosphorylated then 5' phosphorylated as in Levin et al.[18] with the exception that we added Antarctic phosphatase buffer instead of PNK buffer for rephosphorylation. RNA was purified using RNAqueous Micro columns (Ambion; #AM1931), eluted in 2 x 10 μl, ethanol precipitated (-80°C overnight) and resuspended in 5.2 μl of DEPC-treated water. The 3' and 5' adaptors (adenylated-A and B respectively) were ligated as in Levin et al.[18] but using 10 U T4 RNA ligase 2 truncated (Fermentas; #M0242) and T4 RNA ligase respectively (Fermentas; #EL0021). RNA was purified using RNAqueous Micro columns (Ambion; # AM1931), precipitating with 0.8 volumes of ethanol to remove un-ligated adaptors, eluted in 2 x 10 μl, ethanol precipitated (-80°C overnight) and resuspended in 5 μl of DEPC-treated water.

The RevertAid FirstStrand cDNA synthesis kit was used for reverse transcription (Fermentas; #K1621). 0.5 μl small RNA RT primer (D, 100 μM) was added to the RNA and incubated at 65°C for 10 min. Two μl 5x First Strand Buffer, 1 μl 10 mM dNTPs and 0.5 μl RiboLock RNase inhibitor were added. After incubation at 48°C for 3 min, 1 μl RevertAid was added and the reaction incubated at 42°C for 1 h before deactivating the enzyme at 70°C for 10 min. Three cycles of PCR were used to convert the ssDNA template into dsDNA (200 μl reaction volume, 1x GC buffer, 125 nM for each PCR primer (E and F), 250 μM dNTPs, 3% DMSO and 4 U Phusion (Finnzymes; #F-530) with thermocycling conditions of 30 sec at 98°C, 3 cycles of 10 sec at 98°C, 45 sec at 72°C, then an extension of 5 min at 72°C). Following ethanol precipitation (-80°C for at least 1 h), DNA was resuspended in 10 μl water and run on a 1% agarose gel in 1x TAE buffer. Regions with the desired insert sizes were excised, DNA extracted using the Qiaquick gel elution kit (Qiagen; #28704) and eluted in 30 μl EB. Test PCRs were run to determine the optimum DNA dilution for each sample. Final PCRs were run (500–1,500 μl reaction volume depending on test PCR results, 10 μl RT reaction, 1x GC buffer, 125 nM for each PCR primer (E and F), 250 μM dNTPs, 3% DMSO and 20 U/mL Phusion

(Finnzymes; #F-530) with thermocycling conditions of 30 sec at 98°C, 17 cycles of 10 sec at 98°C, 45 sec at 72°C, then an extension of 5 min at 72°C). PCR products were ethanol precipitated (-80°C for at least 1 h), resuspended in 10 μL of water and run on a 1% agarose gel in 1x TAE buffer. Regions with the desired insert sizes were excised, DNA extracted using the Qiaquick gel elution kit (Qiagen; #28704) and eluted in 30 μl water.

Library quality and concentration were evaluated on a Bioanalyzer DNA 1000 Chip (Agilent technologies: #5067-1504). Library concentrations were brought to 10 nM with EB supplemented with 0.1% Tween20. Notes: (1) Combined adaptor size after PCR is approximately 120 bp. In our experience, the libraries run 50 bp higher on the second agarose gel, so we adjusted band excision accordingly. (2) Number of PCR cycles could be further optimized to reduce clonal sequencing.

Illumina GA-II sequencing. The libraries were sequenced on an Illumina Genome Analyzer IIx with 76 bp paired-end reads using version 4 sequencing reagent kits (Illumina) with the Illumina small RNA and PE 2 sequencing primers. Sequencing primers were ordered from Eurofins MWG and prepared as detailed in Quail et al.

On average per library we obtained 26.7 million read pairs (24.4, 26.0, 32.4 and 23.8 million reads for the libraries with average insert sizes of 215, 350, 475 and 625, respectively), for a total of 106.6 million read pairs. The reads were not filtered by read quality (specifically the Illumina quality filter was turned off).

Read mapping. The reads of the four libraries were aligned with PALMapper[46] version 0.4a to the *C. elegans* genome[45] allowing at most six mismatches and two indels, a maximal intron size of 25 kb and at most two introns per alignment.[38] This resulted in a total of 95.3 million aligned fragments (89.3%) where at least one of the two reads could be aligned (21.2, 22.8, 29.4 and 21.9 million fragments for libraries with average insert sizes of 215 bp, 350 bp, 475 bp and 625 bp, respectively).

For library quality evaluation (see below) we considered only a subset of the reads for alignment. We aligned reads from the libraries with smaller insert sizes (average 215 and 350 bp) and added paired-end information by using the tool Fixmate included in the SAMTools release.[47] At most one alignment per read was considered in further analyses: we took the highest sum of alignment scores of consistent pairs of alignments or highest score for singleton alignments.

Library quality evaluation. We adapted the scripts used for the analyses by Levin et al.[18] to the *C. elegans* genome and our RNA-seq data (scripts obtained from http://broadinstitute.org/regev/rnaseqmethods/on August 27, 2010). All analyses were based on comparisons to the protein-coding genes in the WormBase annotation.[48] The set of alignments was randomly down-sampled to 6,775 M reads to correct for the different transcriptome sizes of *C. elegans* (28.3 Mb) and *S. cerevisiae* (10.4 Mb) where 2.5 M reads were used for library quality evaluation. After this adjustment, the median expression of the top 50% expressed genes of our sample was similar to that of the yeast dUTP set (8.4 vs. 7.8). To account for the different average transcript lengths in the two organisms, we reported a normalized version of the weighted number of segments per gene, where we used the mean transcript length of the respective organism for normalization. Finally, we compared gene expression estimates based on our RNA-seq sequences to estimates from a tiling array experiment undertaken with an RNA sample from young adult worm cells. From the RNA-seq read alignments, we estimated gene expression by considering the binary logarithm of the number of aligned reads that fall within the annotated genic region. We used expression estimates (binary logarithm-based) provided by Stefan R. Henz[39] based on tiling array measurements that were prepared using a combination of different analysis tools including RMA.[49] We determined the Pearson's correlation coefficient between the RNA-seq and the tiling array-based estimates, and calculated the root mean square error (RMSE) after normalizing the estimates by the mean absolute difference of the two sets and fitting a linear model of the RNA-seq data set to the tiling array data set. Results were comparable when using Spearman's correlation coefficients. The modified scripts and the reads used for the analyses are available from the **Supplemental website**.

Intergenic read mapping. To assess read mapping in intergenic regions, we created a logical map of the *C. elegans genome*. Based on the wormbase annotation (WS199), the modENCODE data set (see below) and the transcriptionally active regions (TARs) from Spencer et al.,[39] all genic regions were marked with a 1 (the rest with 0). We counted the percentage of uniquely aligned reads that overlapped at least 10 bp with an intergenic region (defined as regions of 0 within the logical genome array). For read counting, the alignments from the two shortest insert size libraries were merged and filtered to retain only the best mapping alignments (identified by alignment quality score). From this alignment set we randomly chose a subset of 6.77 M read alignments for further processing.

Detection of alternative splicing events. We identified intron retention and exon skipping events for each library using splicing graphs, as implemented in the mGene-Toolbox[38,50] with extensions described in reference 22 and 27. For each gene, we constructed a splicing graph[50] representing the set of all possible isoforms, initialized with the exons of annotated isoforms as nodes, and edges connecting nodes whenever an annotated intron connects two exons. For each library as well as for their union, we generated a splicing graph that was then modified to reflect any library-specific splicing event supported by the RNA-seq alignments (specific settings chosen by manual inspection):

To identify intron retention events. We checked for each annotated intron whether at least 90% of the intron was covered with RNA-seq reads, and that the average alignment coverage was at least 5 and between 20% and 120% of the coverage of flanking exons (flanking exons were required to have less than 4-fold difference in their alignment coverage). If all conditions were satisfied, we extended the splicing graph by an additional exon starting at the 5' position and ending at the 3' end of the two flanking exons, respectively. Vice versa, if spliced alignments, with at most one mismatch, of at least five RNA-seq reads confirmed an intron that was completely contained in an exon represented in the splicing graph, we extended the splicing graph by

introducing two new exons connected by this intron and otherwise inheriting 5' and 3' edges of the exon.

To detect exon skipping events. For each intron confirmed by at least five spliced RNA-seq alignments, with at most one mismatch, we checked whether it connected two exons in the splicing graph that were not connected by an edge. If so, the graph was extended by a new edge connecting the two exons.

Using the modified splicing graphs, we then identified candidate intron retention and exon skipping events. For intron retentions we searched for exons that completely covered the intron. For exon skips, we found all triplets of exons, where the first was connected to the second and third, and the second connected to the third (i.e., the second exon can be spliced in and out). We then combined all candidate events from all libraries and merged overlapping candidates.

We next tested every candidate for RNA-seq evidence of inclusion as well as exclusion of the exon or intron: For each intron retention candidate, we checked whether over 90% of the intron was covered with RNA-seq reads and the intron was spliced out, as above. Similarly, for exon skips we checked for evidence of exon inclusion as well as skipping. We only report alternative splicing events for which these conditions could be verified.

Detection by PCR was performed through reverse transcription of total RNA from *C. elegans* (strain N2) of mixed developmental stages using a RevertAid First Strand cDNA Synthesis kit (Fermentas; #K1622) followed by PCR using Phusion DNA polymerase (Finnzymes; #F-530). Primers to detect exon skipping were designed to span the exon-exon boundary.

Comparison to further data sets. We downloaded the coordinates of transcripts and confirmed introns for all provided modENCODE experiments from ftp://data.modencode.org/all_files/cele-interpreted-1 (December 5th, 2011). All coordinates were transformed from WS220 to WS200 using the UCSC liftOver tool (June 2011 version). Transcript information from Ramani et al.[23] was downloaded from the public GBrowse instance at splicebrowse.ccbr.utoronto.ca/cgi-bin/gb2/gbrowse/elegans/on December 5th, 2011, and filtered to retain alternative splice variants.

To identify novel alternative splicing events, we queried all introns present in the modENCODE transcripts and in the data set from Ramani et al.[23] As a correction for possible coordinate offsets caused by the liftover, we allowed for soft matching at the borders within a 10 bp window. For verification of novel exon skips, we extracted events not present in either one or both of the data sets.

Antisense transcript characterization. To assess antisense transcription, we examined all reads that mapped in an antisense orientation to protein-coding genes and within 250 bp upstream and downstream. We considered only the best alignment for each read and allowed at most four mismatches and one gap. For each gene, the associated antisense read coverage was summed over the following regions: 5' intergenic (up to 250 bp upstream), 3' intergenic (up to 250 bp downstream), 5' UTR, 3' UTR, first exon, last exon, first intron, last intron, all exons, all introns (note: regions are not mutually exclusive). We excluded positions overlapping with other genes or within 500 bp up or downstream of other genes. We also recorded the cumulative length for each of these regions for all considered genes.

We calculated the antisense read distribution in the following way. We first normalized the read coverage of genes with at least one antisense read by dividing the summed coverage of each region in a particular gene by the mean antisense coverage of that gene. This step was done to limit the influence of single, highly covered regions. We then averaged the distributions over all genes by summing these normalized coverages over all genes and dividing by the cumulative lengths over all genes of each region. The resulting numbers correlate with the fold-deviation from uniformity: a coverage of one for each region would indicate equal distribution of the antisense reads along the length of the gene. We repeated the same analysis for sense reads.

Enrichment for specific gene ontology categories in the list of genes with antisense transcription was assessed by GOrilla (cbl-gorilla.cs.technion.ac.il/;[51]). p-values were corrected for multiple testing by Bonferroni correction.

Novel transcript support. To calculate the number of paired read alignments specifically supporting single transcript isoforms, we counted how many pairs are compatible with each isoform, that is having a consistent overlap of at least 5 bp per mate. The counts were computed for all four different insert size libraries. We counted a novel isoform as more consistent with our paired-end coverage if it explained at least more 10 read-pairs than any annotated isoform.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgements

We thank Jonathan Perot at IMBC, Strasbourg for sharing a detailed pre-adenylation protocol with us, Daniel Turner at The Wellcome Trust Sanger Institute for assistance in trouble-shooting and Detlef Weigel at the Max Planck Institute for Developmental Biology for support and discussion. Additionally, we thank Vipin Sreedharan for visualization of novel splicing events and help with conversion of coordinates between Wormbase versions, Valerie Reinke for providing tiling microarray measurements pre-publication and Stefan R. Henz for providing expression estimates from these measurements. Part of this work was done while G.R. was at the Friedrich Miescher Laboratory of the Max Planck Society (Tübingen, Germany).

Author's Contributions

L.M.S. and G.R. conceptualized the research; L.M.S. and L.H. developed and tested the MISSSL protocol; C.L. sequenced the libraries; P.D. performed computational analyses of isoform identification and analysis of antisense transcription; R.B. assessed the quality of the library and performed comparisons with Levin et al.; A.K. aligned the RNA-seq reads, calculated insert sizes and compared exon skipping events to the modENCODE integrated transcripts; and G.R. aligned the reads and performed analysis of alternative splicing and antisense transcription. L.M.S. wrote the manuscript with contributions from all other authors.

Funding

This work was supported by a European Community FP7 Marie Curie Fellowship (PIEF-GA-2008-221553) and an EMBO Long-Term fellowship (L.M.S.) in addition to core funding by the Max Planck Society (G.R.), by the German Research Foundation under grant RA1894/2-1 (G.R.), and the Sloan-Kettering Institute (to G.R.).

Supplemental Materials

Supplemental material can be found at:
www.landesbioscience.com/journal/rnabiology/articles/19683

References

1. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, et al.; ENCODE Project Consortium; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007; 447:799-816; PMID:17571346; http://dx.doi.org/10.1038/nature05874.
2. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 2007; 316:1484-8; PMID:17510325; http://dx.doi.org/10.1126/science.1138341.
3. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 2008; 321:956-60; PMID:18599741; http://dx.doi.org/10.1126/science.1160342.
4. Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, Yu L, et al. A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi.* PLoS Genet 2009; 5:1000569; PMID:19609351; http://dx.doi.org/10.1371/journal.pgen.1000569.
5. http://www.ebi.ac.uk/~zerbino/oases/
6. Hahn DA, Ragland GJ, Shoemaker DD, Denlinger DL. Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophaga crassipalpis*. BMC Genomics 2009; 10:234; PMID:19454017; http://dx.doi.org/10.1186/1471-2164-10-234.
7. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, et al. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol Ecol 2008; 17:1636-47; http://dx.doi.org/10.1111/j.1365-294X.2008.03666.x; PMID:18266620.
8. Surget-Groba Y, Montoya-Burgos JI. Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Res 2010; 20:1432-40; PMID:20693479; http://dx.doi.org/10.1101/gr.103846.109.
9. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009; 10:57-63; PMID:19015660; http://dx.doi.org/10.1038/nrg2484.
10. Faghihi MA, Wahlestedt C. Regulatory roles of natural antisense transcripts. Nat Rev Mol Cell Biol 2009; 10:637-43; PMID:19638999; http://dx.doi.org/10.1038/nrm2738.
11. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, et al. Genome-wide mapping of alternative splicing in *Arabidopsis thaliana.* Genome Res 2010; 20:45-58; PMID:19858364; http://dx.doi.org/10.1101/gr.093302.109.
12. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 2008; 40:1413-5; PMID:18978789; http://dx.doi.org/10.1038/ng.259.
13. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature 2008; 456:470-6; PMID:18978772; http://dx.doi.org/10.1038/nature07509.
14. Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, et al. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell 2000; 101:671-84; PMID:10892653; http://dx.doi.org/10.1016/S0092-8674(00)80878-8.
15. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, et al. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res 2009; 37:123; PMID:19620212; http://dx.doi.org/10.1093/nar/gkp596.
16. Vivancos AP, Güell M, Dohm JC, Serrano L, Himmelbauer H. Strand-specific deep sequencing of the transcriptome. Genome Res 2010; 20:989-99; PMID:20519413; http://dx.doi.org/10.1101/gr.094318.109.
17. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 2008; 133:523-36; PMID:18423832; http://dx.doi.org/10.1016/j.cell.2008.03.029.
18. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods 2010; 7:709-15; PMID:20711195; http://dx.doi.org/10.1038/nmeth.1491.
19. Filiatrault MJ, Stodghill PV, Bronstein PA, Moll S, Lindeberg M, Grills G, et al. Transcriptome analysis of *Pseudomonas syringae* identifies new genes, noncoding RNAs and antisense activity. J Bacteriol 2010; 192:2359-72; PMID:20190049; http://dx.doi.org/10.1128/JB.01445-09.
20. Hafner M, Renwick N, Brown M, Mihailović A, Holoch D, Lin C, et al. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. RNA 2011; 17:1697-712; PMID:21775473; http://dx.doi.org/10.1261/rna.2799511.
21. Lacroix V, Sammeth M, Guigo R, Bergeron A. Exact Transcriptome Reconstruction from Short Sequence Reads. Algorithms in Bioinformatics 2008; 5251:50-63; http://dx.doi.org/10.1007/978-3-540-87361-7_5.
22. Eichner J, Zeller G, Laubinger S, Rätsch G. Support vector machines-based identification of alternative splicing in *Arabidopsis thaliana* from whole-genome tiling arrays. BMC Bioinformatics 2011; 12:55; PMID:21324185; http://dx.doi.org/10.1186/1471-2105-12-55.
23. Ramani AK, Calarco JA, Pan Q, Mavandadi S, Wang Y, Nelson AC, et al. Genome-wide analysis of alternative splicing in *Caenorhabditis elegans.* Genome Res 2011; 21:342-8; PMID:21177968; http://dx.doi.org/10.1101/gr.114645.110.
24. Vigneault F, Sismour AM, Church GM. Efficient microRNA capture and bar-coding via enzymatic oligonucleotide adenylation. Nat Methods 2008; 5:777-9; PMID:19160512; http://dx.doi.org/10.1038/nmeth.1244.
25. Quail MA, Swerdlow H, Turner DJ. Improved protocols for the illumina genome analyzer sequencing system. Curr Protoc Hum Genet 2009; 18:2; PMID:19582764; http://dx.doi.org/10.1002/0471142905.hg1802s62.
26. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008; 456:53-9; PMID:18987734; http://dx.doi.org/10.1038/nature07517.
27. Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, et al. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana.* Nature 2011; 477:419-23; PMID:21874022; http://dx.doi.org/10.1038/nature10414.
28. Pickrell JK, Pai AA, Gilad Y, Pritchard JK. Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genet 2010; 6:1001236; PMID:21151575; http://dx.doi.org/10.1371/journal.pgen.1001236.
29. Lamm AT, Stadler MR, Zhang H, Gent JI, Fire AZ. Multimodal RNA-seq using single-strand, double-strand and CircLigase-based capture yields a refined and extended description of the *C. elegans* transcriptome. Genome Res 2011; 21:265-75; PMID:21177965; http://dx.doi.org/10.1101/gr.108845.110.
30. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, et al.; modENCODE Consortium. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. Science 2010; 330:1775-87; PMID:21177976; http://dx.doi.org/10.1126/science.1196914.
31. Sun M, Hurst LD, Carmichael GG, Chen J. Evidence for a preferential targeting of 3'-UTRs by cis-encoded natural antisense transcripts. Nucleic Acids Res 2005; 33:5533-43; PMID:16204454; http://dx.doi.org/10.1093/nar/gki852.
32. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 2008; 322:1845-8; PMID:19056941; http://dx.doi.org/10.1126/science.1162228.
33. Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, et al. Divergent transcription from active promoters. Science 2008; 322:1849-51; PMID:19056940; http://dx.doi.org/10.1126/science.1162253.
34. Sun M, Hurst LD, Carmichael GG, Chen J. Evidence for variation in abundance of antisense transcripts between multicellular animals but no relationship between antisense transcriptionand organismic complexity. Genome Res 2006; 16:922-33; PMID:16769979; http://dx.doi.org/10.1101/gr.5210006.
35. Yassour M, Pfiffner J, Levin JZ, Adiconis X, Gnirke A, Nusbaum C, et al. Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species. Genome Biol 2010; 11:87; PMID:20796282; http://dx.doi.org/10.1186/gb-2010-11-8-r87.
36. Numata K, Okada Y, Saito R, Kiyosawa H, Kanai A, Tomita M. Comparative analysis of cis-encoded antisense RNAs in eukaryotes. Gene 2007; 392:134-41; PMID:17250976; http://dx.doi.org/10.1016/j.gene.2006.12.005.
37. David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, et al. A high-resolution map of transcription in the yeast genome. Proc Natl Acad Sci USA 2006; 103:5320-5; PMID:16569694; http://dx.doi.org/10.1073/pnas.0601091103.
38. Schweikert G, Zien A, Zeller G, Behr J, Dieterich C, Ong CS, et al. mGene: accurate SVM-based gene finding with an application to nematode genomes. Genome Res 2009; 19:2133-43; PMID:19564452; http://dx.doi.org/10.1101/gr.090597.108.

39. Spencer WC, Zeller G, Watson JD, Henz SR, Watkins KL, McWhirter RD, et al. A spatial and temporal map of *C. elegans* gene expression. Genome Res 2011; 21:325-41; PMID:21177967; http://dx.doi.org/10.1101/gr.114595.110.
40. Wierzbicki AT, Haag JR, Pikaard CS. Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. Cell 2008; 135:635-48; PMID:19013275; http://dx.doi.org/10.1016/j.cell.2008.09.035.
41. Ruzanov P, Riddle DL. Deep SAGE analysis of the *Caenorhabditis elegans* transcriptome. Nucleic Acids Res 2010; 38:3252-62; PMID:20129939; http://dx.doi.org/10.1093/nar/gkq035.
42. Jones SJ, Riddle DL, Pouzyrev AT, Velculescu VE, Hillier L, Eddy SR, et al. Changes in gene expression associated with developmental arrest and longevity in *Caenorhabditis elegans.* Genome Res 2001; 11:1346-52; PMID:11483575; http://dx.doi.org/10.1101/gr.184401.
43. ftp://ftp.ensembl.org/pub/release-59/gtf/homo_sapiens/
44. http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/
45. ftp://ftp.wormbase.org/pub/wormbase/genomes/c_elegans/sequences/dna/c_elegans.WS200.dna.fa.gz
46. Jean G, Kahles A, Sreedharan VT, De Bona F, Rätsch G. RNA-Seq read alignments with PALMapper. Curr Protoc Bioinformatics 2010; 11:6; PMID:21154708.
47. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al.; 1,000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. Bioinformatics 2009; 25:2078-9; PMID:19505943; http://dx.doi.org/10.1093/bioinformatics/btp352.
48. Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, et al. WormBase: a comprehensive resource for nematode research. Nucleic Acids Res 2010; 38:463-7; PMID:19910365; http://dx.doi.org/10.1093/nar/gkp952.
49. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization and summaries of high density oligonucleotide array probe level data. Biostatistics 2003; 4:249-64; PMID:12925520; http://dx.doi.org/10.1093/biostatistics/4.2.249.
50. Heber S, Alekseyev M, Sze SH, Tang H, Pevzner PA. Splicing graphs and EST assembly problem. Bioinformatics 2002; 18:181-8; PMID:12169546; http://dx.doi.org/10.1093/bioinformatics/18.suppl_1.S181.
51. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 2009; 10:48; PMID:19192299; http://dx.doi.org/10.1186/1471-2105-10-48.