Comprehensive comparative analysis of strand-specific RNA sequencing methods

Joshua Z Levin[1,6], Moran Yassour[1–3,6], Xian Adiconis[1], Chad Nusbaum[1], Dawn Anne Thompson[1], Nir Friedman[3,4], Andreas Gnirke[1] & Aviv Regev[1,2,5]

Strand-specific, massively parallel cDNA sequencing (RNA-seq) is a powerful tool for transcript discovery, genome annotation and expression profiling. There are multiple published methods for strand-specific RNA-seq, but no consensus exists as to how to choose between them. Here we developed a comprehensive computational pipeline to compare library quality metrics from any RNA-seq method. Using the well-annotated *Saccharomyces cerevisiae* transcriptome as a benchmark, we compared seven library-construction protocols, including both published and our own methods. We found marked differences in strand specificity, library complexity, evenness and continuity of coverage, agreement with known annotations and accuracy for expression profiling. Weighing each method's performance and ease, we identified the dUTP second-strand marking and the Illumina RNA ligation methods as the leading protocols, with the former benefitting from the current availability of paired-end sequencing. Our analysis provides a comprehensive benchmark, and our computational pipeline is applicable for assessment of future protocols in other organisms.

Recent advances in massively parallel cDNA sequencing (RNA-seq) have opened the way for comprehensive analysis of any transcriptome[1]. In principle, RNA-seq allows analysis of all expressed transcripts, with three key goals: (i) annotating the structures of all transcribed genes including their 5′ and 3′ ends and all splice junctions[2–4], (ii) quantifying expression of each transcript[5,6] and (iii) measuring the extent of alternative splicing[7–11].

Standard libraries for RNA-seq do not preserve information about which strand was originally transcribed. Synthesis of randomly primed double-stranded cDNA followed by addition of adaptors for next-generation sequencing leads to the loss of information about which strand was present in the original mRNA template. In some cases, strand information can be inferred by subsequent computational analyses using, for example, open reading frame (ORF) information in protein-coding genes, biases in coverage between 5′ and 3′ ends[4] or splice-site orientation in eukaryotic genomes[4,10,11].

Nevertheless, direct information on the originating strand can substantially enhance the value of an RNA-seq experiment. For example, such information would help to accurately identify antisense transcripts, with potential regulatory roles[12], determine the transcribed strand of other noncoding RNAs, demarcate the exact boundaries of adjacent genes transcribed on opposite strands and resolve the correct expression levels of coding or noncoding overlapping transcripts. These tasks are particularly challenging in small microbial genomes, prokaryotic and eukaryotic, in which genes are densely coded, with overlapping untranslated regions (UTRs) or ORFs and in which splice-site information is limited or nonexistent.

Many methods have been recently developed for strand-specific RNA-seq, and they fall into two main classes. One class relies on attaching different adaptors in a known orientation relative to the 5′ and 3′ ends of the RNA transcript (**Fig. 1a**). These protocols generate a cDNA library flanked by two distinct adaptor sequences, marking the 5′ end and the 3′ end of the original mRNA. A second class of methods relies on marking one strand by chemical modification, either on the RNA itself by bisulfite treatment or during second-strand cDNA synthesis followed by degradation of the unmarked strand (**Fig. 1b**). Both modification methods essentially follow the standard protocol for RNA-seq with the exception of these marking steps.

Although standard RNA-seq largely relies on one protocol, the great diversity of published protocols for strand-specific RNA-seq poses several challenges. First, when conducting an experiment, researchers are challenged to identify a suitable protocol. Furthermore, if protocols vary considerably in their performance, the chosen method can dramatically affect the conclusions drawn from an experiment, confounding interpretation and comparison across studies. There is therefore a substantial need for a systematic evaluation of the performance of different protocols for strand-specific RNA-seq.

Here we present a comprehensive comparison of seven protocols for strand-specific RNA-seq. Using *Saccharomyces cerevisiae* poly(A)$^+$ RNA, we built a compendium of libraries using these

[1]Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts, USA. [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]School of Engineering and Computer Science, Hebrew University, Jerusalem, Israel. [4]Alexander Silberman Institute of Life Sciences, Hebrew University, Jerusalem, Israel. [5]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [6]These authors contributed equally to this work. Correspondence should be addressed to J.Z.L. (jlevin@broadinstitute.org) or A.R. (aregev@broad.mit.edu).

RECEIVED 26 MARCH; ACCEPTED 20 JULY; PUBLISHED ONLINE 15 AUGUST 2010; DOI:10.1038/NMETH.1491

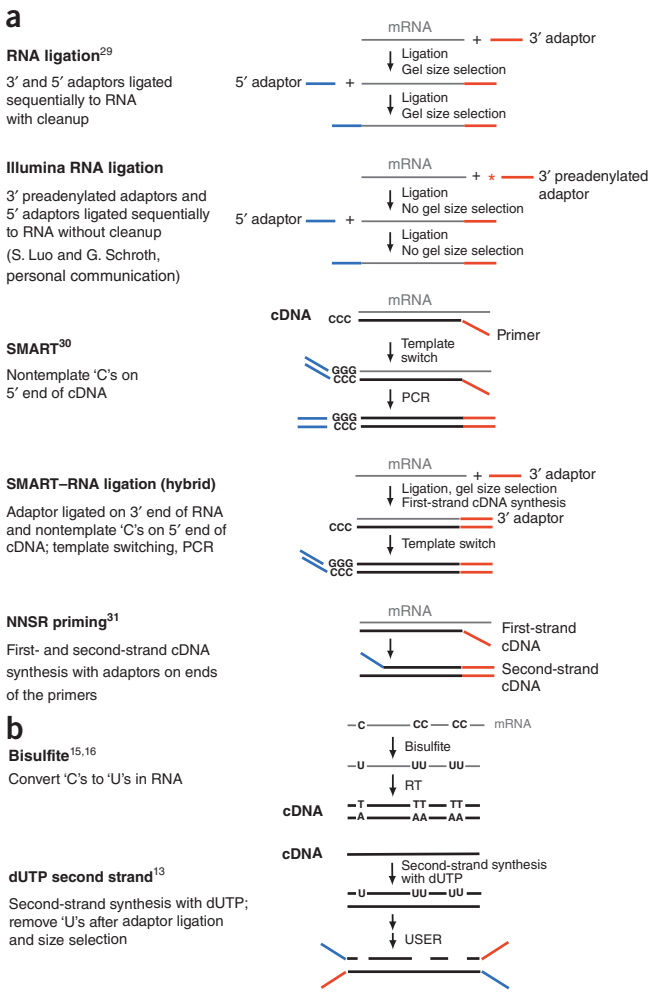

Figure 1 | Methods for strand-specific RNA-seq. (**a**,**b**) Salient details for differential adaptor methods including RNA ligation[29], SMART[30] and NNSR priming[31] (**a**) and differential marking methods (**b**). USER, uracil-specific excision reagent. mRNA is shown in gray and cDNA in black. For differential adaptor methods, 5′ adaptors are shown in blue, and 3′ adaptors are shown in red.

protocols and sequenced each of them on an Illumina Genome Analyzer instrument to deep coverage. We developed a computational pipeline to assess each library's quality according to library complexity, strand specificity, evenness and continuity of coverage, agreement with known genome annotation and quantitative accuracy for expression profiling, in addition to considering the ease of laboratory and computational manipulations. We identified the dUTP and Illumina RNA ligation methods as the leading protocols, with the dUTP library providing the added benefit of the ability to conduct paired-end sequencing.

RESULTS

A comparison of strand-specific RNA-seq

We evaluated 13 stand-specific libraries. We constructed 11 libraries based on seven strand-specific RNA-seq methods (**Fig. 1**), including two variations for four of the methods. We also compiled comparable data for two published libraries: a dUTP library[13] and a library based on another (eighth) method from the differential adaptor class[14] (3′ split adaptor; **Supplementary Fig. 1**). Finally, we prepared a standard, non–strand-specific cDNA library to use as a control in these comparisons.

We explored two different variations for four of the seven methods to improve our libraries (Online Methods). These variations were the addition of actinomycin D to the 'not not so random' (NNSR) library protocol, two published variations of the bisulfite library protocol ('H' and 'S'; Online Methods[15,16]), different size-selection methods for the Illumina RNA ligation libraries and different reverse transcription primers for the dUTP libraries. We present results only for the 'S' bisulfite library because we found no substantial differences between the two libraries in our analyses.

We used each method to prepare a cDNA library for Illumina sequencing from *S. cerevisiae* poly(A)$^+$ RNA. We chose *S. cerevisiae* because this eukaryotic model organism has an exceptionally well-annotated genome, facilitating quality evaluations. We used paired-end Illumina sequencing for each library (Online Methods), except for the RNA ligation and Illumina RNA ligation libraries, which we sequenced only from the 3′ end of each cDNA because of the RNA adaptors used in these protocols. These approaches could be modified in the future to accommodate paired-end sequencing by changing the RNA adaptor and PCR primer sequences.

An analysis framework for assessing RNA-seq libraries

To compare the quality of the different libraries, we defined six assessment criteria (**Fig. 2**) implemented in a computational pipeline (Online Methods). These criteria were library complexity, defined as the number of unique reads (**Fig. 2a**); strand specificity, defined as the number of reads mapping to known transcribed regions at the expected strand (**Fig. 2b**); evenness and continuity of coverage at annotated transcripts (**Fig. 2c**,**d**); performance at 5′ and 3′ ends, defined as agreement with known end annotation (**Fig. 2d**); and performance in expression profiling, defined by sensitivity, linearity and dynamic range. With the exception of strand specificity, we compared each criterion to that for the control library. We focused on only one variation per method unless there were substantial differences in performance between variations. We provide the full evaluation results in **Supplementary Tables 1–2** and **Supplementary Figures 2–4**.

Equal sampling of reads enables direct library comparisons

We mapped each library's reads to the *S. cerevisiae* genome using Arachne[17]. For paired-end libraries, we mapped unique pairs with opposite orientations and an appropriate separation; for single-end libraries, we identified unique mappings for individual reads[17] (Online Methods).

The libraries had a broad range of yields, measured by the total number of reads and by the number of reads or paired reads mapping to a unique location (**Supplementary Table 1**). In this initial comparison, the dUTP library had the highest percentage of paired-end mapped reads (**Supplementary Table 1**). The Illumina RNA ligation–solid-phase reversible immobilization (SPRI) library, which we prepared using SPRI-based size selection, had a smaller percentage of unique reads than the Illumina RNA ligation library, which we prepared using gel-based size selection (35% versus 59%; **Supplementary Table 1**). This was likely due to the difficulty in physically removing cDNAs shorter than 76 base pairs with the SPRI method, resulting in the ends

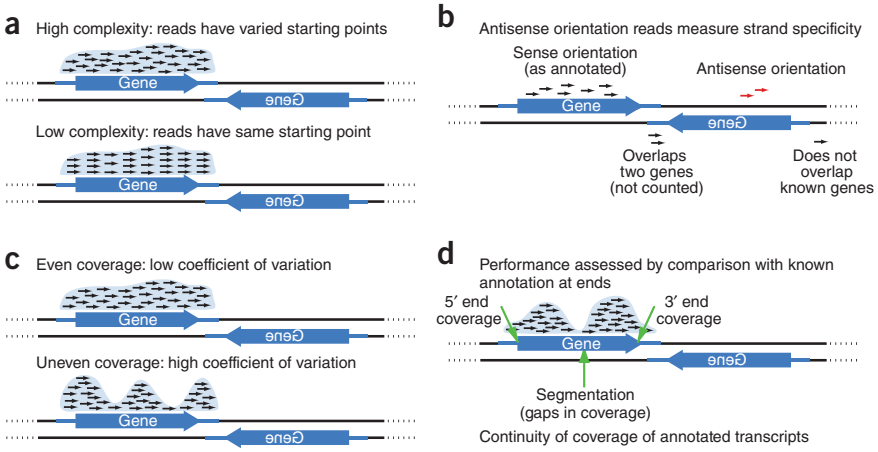

Figure 2 | Key criteria for evaluation of strand-specific RNA-seq libraries. (**a**–**d**) Categories of quality assessment were complexity (**a**), strand specificity (**b**), evenness of coverage (**c**) and comparison to known transcript structure (**d**). Double-stranded genome with gene ORF orientation (blue arrows) and UTRs (blue lines) are shown along with mapped reads (black and red arrows, reads mapped to sense and antisense strands, respectively).

of sequencing reads containing an Illumina adaptor sequence that could not be aligned to the yeast genome. Indeed, when we trimmed these reads to 51 bases, the percentage of aligned reads improved dramatically (data not shown). Below, we report results only for the Illumina RNA ligation library, which we prepared using gel-based size selection.

Some of this variation in performance may reflect variation in sequencing yields between sequencing runs and lanes (**Supplementary Table 1**), unrelated to the library protocol. As many of our measures were sensitive to read quantity and length, we used sampling to obtain the same number of reads from each library (Online Methods). Unless specifically noted, we conducted all subsequent comparisons with 2.5 million sampled reads from each library. The 'switching mechanism at 5′ end of RNA template' (SMART) library had only 930,686 reads because of repeated poor yields, but with the exception of complexity, we obtained overall similar results when using the SMART reads 'as is' (without any compensatory calculations for there being fewer than 2.5 million reads) or when randomly resampling the same reads more than once to reach 2.5 million (data not shown). To compare libraries with different read lengths (51 or 76 bases in our libraries and 36 bases in published data), we sampled the first 36 bases of every read.

Complexity of single- and paired-end libraries

We next assessed the complexity of each library, defined as the number of distinct (unique) read start positions (**Fig. 2a**). A high complexity library, with many different start positions, is preferable as it does not suffer from 'jackpot' effects in fragment amplification or a strong bias in selection of fragment ends. Using single-end mapping (**Fig. 3a** and **Supplementary Table 2**), we observed the best complexity for the control library (42% unique) followed closely by the 3′ split adaptor method (42% unique), SMART (41% unique) and the published dUTP method (40% unique).

Single-read complexity calculations may overestimate the number of redundant cDNAs in a library. For paired-end libraries, we also estimated complexity as unique pairs of start and end positions (**Fig. 3b**), because cDNAs that have the same start site for one read can be distinguished based on a different start site for the other read in the pair. Comparing paired-end libraries by this measure, we found that the control and dUTP libraries performed best, with 88% and 84% unique paired reads, respectively. This demonstrates that paired-end sequencing substantially improves estimates of library complexity relative to estimates using only single reads.

Strand specificity across libraries

We measured the strand specificity of each library by comparing the mapped reads to the expected transcribed strand based on the known *S. cerevisiae* annotation (Online Methods). Based on recent studies[18], we conservatively assumed that most of the *S. cerevisiae* genes are not transcribed from the antisense strand and used the fraction of reads mapped to the opposite (antisense) strand of known transcripts as a measure of strand specificity (**Fig. 2b**, **Supplementary Table 2** and Online Methods).

Four of the protocols, RNA ligation, Illumina RNA ligation, dUTP and NNSR (with actinomycin D), performed best, whereas the SMART approach was the least strand-specific method, by a wide margin (**Fig. 4** and **Supplementary Fig. 5**). Only 0.47–0.63% of the reads mapped to the antisense strand for the four best performing methods. Notably, addition of actinomycin D dramatically improved the strand specificity of the NNSR method (**Supplementary Table 2**). Actinomycin D treatment cannot be used to improve the strand specificity of SMART because it inhibits both second-strand synthesis and template switching[19] (X.A. and J.Z.L.; data not shown).

Evenness and continuity of annotated transcript coverage

Using RNA-seq for effective transcriptome annotation, which includes transcript assembly[3,4], separating neighboring genes correctly and identifying full-length transcripts with correct 5′ and 3′ ends requires even, continuous and complete coverage along each transcript's length.

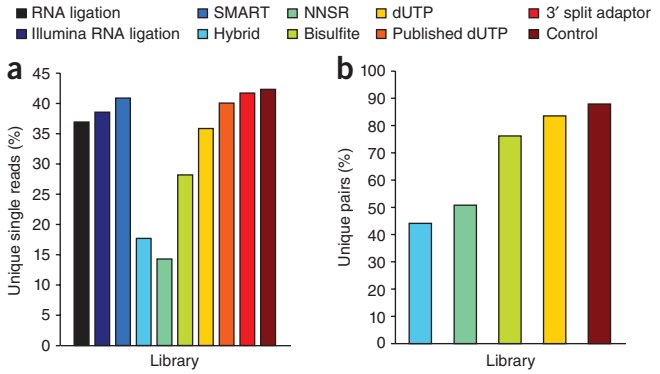

Figure 3 | Complexity of single- and paired-end libraries. (**a**,**b**) Percentage of unique reads mapping out of the total number of mapped reads, when considering only single-mapped reads (**a**; all libraries) or uniquely mapped pairs (**b**; only paired-end libraries).

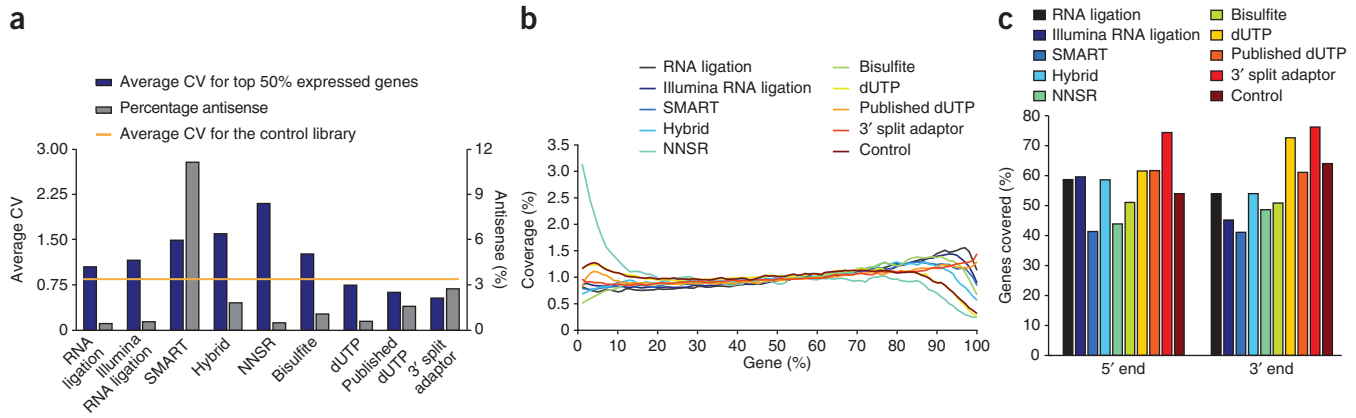

Figure 4 | Strand specificity and evenness of transcript coverage. (**a**) Strand specificity (percentage antisense) and evenness of coverage (average coefficient of variation (CV)) for all libraries. (**b**) Relative gene coverage at each percentile of a gene's length, averaged across all genes in each library. The 5′ end is on the left. (**c**) Percentage of genes with 5′-end and 3′-end coverage in each library.

To measure evenness of coverage for each library, we calculated the average of the coefficient of variation of gene coverage for the top 50% expressed genes (**Figs. 2c** and **4a**, **Supplementary Fig. 5** and **Supplementary Table 2**). We found the most even coverage for the 3′ split adaptor method[14] (average coefficient of variation, 0.54), closely followed by that for the dUTP approach (average coefficient of variation of 0.64 in the original dataset[13] and 0.76 in our hands).

We defined two measures of continuity of coverage. First, we counted the number of segments into which each known transcript was broken, where we defined a break as a stretch of at least five bases without read coverage (**Figs. 2d** and **5a** and **Supplementary Table 2**). We then averaged this measure across all genes, weighting by the relative expression of each gene (we expected low-expressed genes to be less covered and more segmented). The best performing methods by this measure were the 3′ split adaptor method[14] (2.29 segments per gene), the dUTP libraries (2.41 and 2.48 segments per gene with published data[13] and in our hands, respectively) and the Illumina RNA ligation libraries (2.61 segments per gene).

Second, we calculated the fraction of bases without coverage in each transcript (**Figs. 2d** and **5b–e** and **Supplementary Fig. 2**) and examined the distribution of this fraction at different expression levels, as defined by pooling data across libraries (Online Methods). As expected, in all libraries, the fraction of uncovered bases decreased as expression increased (**Fig. 5b–e** and **Supplementary Fig. 2**). However, both the rate of decrease and the coverage per transcript at higher expression levels were variable between better performing libraries (**Fig. 5c**,**d**) and poorly performing ones (**Fig. 5e**). To systematically assess this difference, we compared the Lowess fits of each of the distributions (**Fig. 5b** and **Supplementary Fig. 2**). We found that the dUTP (both in our hands (**Fig. 5c**) and in published data[13]) and 3′ split adaptor (**Fig. 5d**) methods performed best.

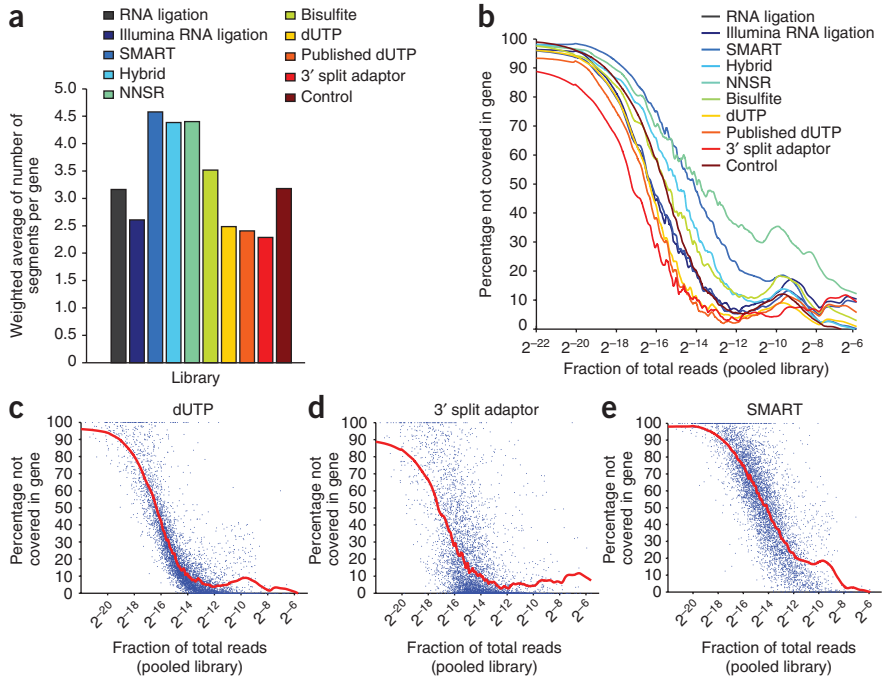

Figure 5 | Continuity of transcript coverage. (**a**) Average number of segments (separated by at least five bases of zero coverage) weighted by the average expression of each gene, in each library. (**b**) Lowess fit for each library. (**c–e**) Plots for the dUTP method (**c**), the 3′ split adaptor method (**d**) and the SMART method (**e**). In **c–e**, a Lowess fit is shown as a red curve, and each gene is represented by a blue dot.

Coverage at 5′ and 3′ ends

Coverage at 5′ and 3′ ends is crucial for correctly identifying full-length transcripts. To estimate this, we computed for each library the average coverage at each percentile of length from the annotated 5′ end to the annotated 3′ end of known transcripts[18] (**Figs. 2d** and **4b**), as well as the number of genes with complete coverage of their 5′ and 3′ ends (**Fig. 4c**). For paired-end libraries, we computed 5′ and 3′ end coverage based on both read pairs, thus estimating coverage of each end based on the relevant read.

We found substantial variation in the average coverage along a gene's length, with

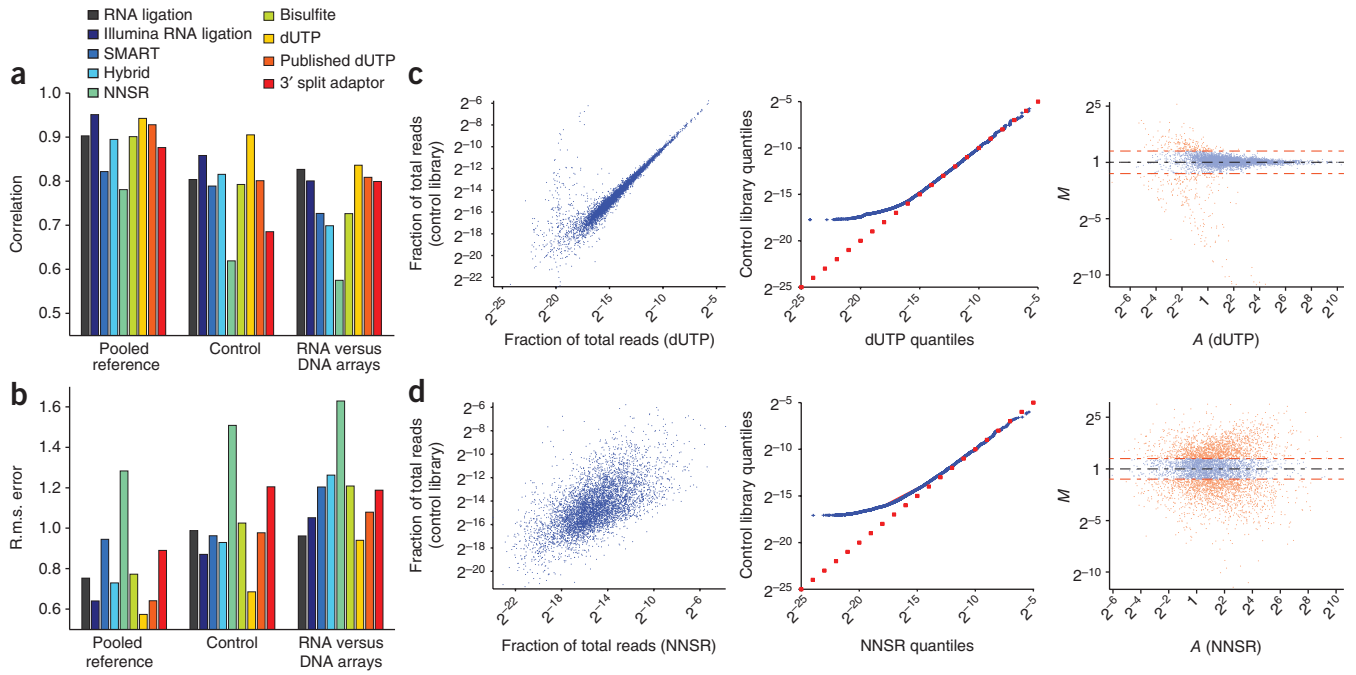

Figure 6 | Digital expression profiling using strand-specific RNA-seq. (**a**,**b**) Pearson correlation coefficient (**a**) and r.m.s. error (**b**) for each library when compared to a pooled reference, the control library and Agilent microarrays (right). (**c**,**d**) Scatter (left), Q-Q (middle) and MA (right) plots for the best performing (dUTP; **c**) and worst performing (NNSR; **d**) libraries, in comparison to the control library. The scatter plots show the fraction of total reads for each gene (blue dot) in the control library against a strand-specific library. The Q-Q plot shows the level at each quantile (rank) of expression in the control library against the strand-specific library. A slope = 1 line is shown for reference (red). The MA plot shows for each gene (dot) the difference in expression levels between the control and strand-specific libraries (*M*; *y* axis) compared to their mean expression level (*A*; *x* axis). Red and blue dashed lines indicate twofold and onefold difference in expression, respectively.

specific biases in 5′ and 3′ coverage (**Fig. 4b**,**c**, **Supplementary Fig. 3** and **Supplementary Table 2**). The NNSR library data had more coverage at the 5′ ends of transcripts, whereas the remaining libraries had modestly increased coverage of the 3′ ends (**Fig. 4b** and **Supplementary Fig. 3**). Consistent with its evenness and continuity, the 3′ split adaptor method had the best coverage of both 5′ and 3′ ends (75% and 77% of genes covered completely at each end, respectively), followed by the dUTP method (62% and 73%) (**Fig. 4c** and **Supplementary Table 2**). The addition of oligo(dT) primers for reverse transcription for the dUTP method, both in our results and in the published data[13], did not increase the coverage at the 3′ ends (**Supplementary Table 2**), although more lenient read mapping may assist with this task in reads that contain portions of the poly(A) tail.

Performance for digital expression profiling

We compared the performance of each library in digital expression profiling relative to reference expression measurements estimated from three 'standard' sources: the control (non–strand-specific) library; a pooled estimate generated from the sampled reads of nine of the strand-specific libraries (Online Methods); and expression profiles measured by competitive hybridization of a mid-log phase RNA sample versus genomic DNA using Agilent arrays (Online Methods). We calculated the expression of each gene as its length-normalized read coverage and normalized all values for the total number of reads.

We used several standard quality measures[20] to estimate each library's performance. These included the Pearson correlation coefficient of expression levels across all genes (**Fig. 6a** and **Supplementary Table 2**); the root mean squared (r.m.s.) error of the expression measurements in each library using the reference measurement as the expected level (**Fig. 6b** and **Supplementary Table 2**); and scatter, quantile-quantile (Q-Q) and MA[21] plots—the last of which compare for each gene the difference in expression between two libraries to the mean expression of that gene in the two libraries (Online Methods, **Fig. 6c**,**d** and **Supplementary Fig. 4**) that help compare differences in expression levels across the dynamic range.

We found that the dUTP library had the best correlation and lowest r.m.s. error relative to all three references (**Fig. 6b** and **Supplementary Table 2**). The only exception was that the Illumina RNA ligation method had a slightly better (0.95 versus 0.94) correlation to the pooled library (**Supplementary Table 2**). Furthermore, visual inspection of the scatter, Q-Q and MA plots showed an excellent linear relation between the dUTP library and the control library across a broad range of values, with weaker performance only for genes with very low expression (**Fig. 6c**). The Illumina RNA ligation protocol also performed reasonably well based on the correlation and r.m.s. error measures but with noticeably broader scatter across the expression range (**Supplementary Fig. 4**). The worst performing methods were the SMART, NNSR and 3′ split adaptor libraries (**Fig. 6d** and **Supplementary Fig. 4**), by all measures.

DISCUSSION

The evaluated RNA-seq protocols broadly represent existing approaches (for a summary of their relative merits, see **Supplementary Table 3**), and we excluded some protocols because of well-known technical limitations, incomplete method development

or high similarity to tested methods. These excluded protocols comprise single-stranded cDNA library synthesis[22] (owing to chimeric cDNAs created); deep sequencing of ribosome-protected mRNA fragments[14] (because cDNA lengths are too short, and the original method involves a complex procedure for RNA preparation; we included published data from the nonprotected library designated as the 3′ split adaptor method; **Supplementary Fig. 1**); Helicos single-molecule digital gene expression[23] and direct RNA sequencing[24] (coverage heavily biased to the 5′ or 3′ ends of transcripts, respectively; the latter is currently still under development); and ligation of adaptor to 5′ end and C-tailing at 3′ end of RNA[25] and the double-random priming method[26] (similar to NNSR). We did not include FRT-seq[27] and SOLiD Whole Transcriptome Analysis kit (Applied Biosystems)[28] because they are similar to the two RNA ligation methods we tested, and it would be difficult to distinguish differences owing to library construction methods from those because of the different sequencing methods.

In addition to the formal criteria we evaluated, there is substantial variation in the experimental complexity of different protocols (**Supplementary Table 4**). The original RNA ligation method is the most labor intensive and requires substantial amounts of starting material. The NNSR protocol is the simplest. It is unclear how well the original RNA ligation method can be adapted to larger fragments (greater than 152 base pairs) needed for paired-end sequencing with 76-base reads as it requires the adaptor-ligated RNA to be separated on a gel from unligated RNA, an increasing challenge as the length of the RNA increases.

The libraries also vary in the facility of computational analysis, in particular at early processing steps. The bisulfite method is the most computationally challenging, as reads must be aligned to two reference 'genomes' that have all the cytosine bases converted to thymine bases on one of the two strands. This alignment is complicated both by the imperfect efficiency of the bisulfite treatment and by inherent sequencing errors.

Our analysis allowed us to assess the tradeoff between different protocol modifications. For example, we found that actinomycin D improved the strand specificity of the NNSR protocol (**Supplementary Table 2**) but had the opposite effect on the coefficient of variation, 5′ and 3′ end coverage and correlation of expression levels (**Supplementary Table 2**). For the Illumina RNA ligation libraries, it is preferable to use gel size selection rather than SPRI because removing the shorter cDNAs increased the fraction of reads aligning to the yeast genome. If read length is reduced below 76 bases, this may be less of an issue, but such a choice would also impact other sequencing outputs. Notably, SPRI is amenable to liquid handling automation and can increase the throughput and convenience of any of the other methods, except for RNA ligation. Although these modifications impacted library quality for the NNSR and Illumina RNA ligation methods, most of the variations tested did not alter the performance characteristics of the libraries (**Supplementary Table 2** and **Supplementary Figs. 2–4**), an indication of the reproducibility of the methods. We did not directly evaluate the experimental features, such as PCR conditions or adaptor sequences, that contributed to each method's success (or lack thereof) because these may be complex. We note, however, that the amount of starting material did not correlate with library complexity (**Supplementary Tables 2** and **4**).

The dUTP protocol provided the most compelling overall balance across criteria, followed closely by the Illumina RNA ligation protocol (**Supplementary Note 1**). Currently, the dUTP protocol is compatible with paired-end sequencing, whereas the present Illumina RNA ligation protocol is not. Paired-end sequencing increases the number of mappable reads (unique as pairs), and in higher eukaryotes provides substantial power in transcriptome reconstruction[10,11]. The 3′ split adaptor method[14] excelled in measures critical for genome annotation, but was less well suited for expression profiling. Finally, our compendium and analysis pipeline, which is available online (http://www.broadinstitute.org/regev/rnaseqmethods/) and will be provided as a GenePattern module (http://www.broadinstitute.org/cancer/software/genepattern/), are important resources and include a general benchmarking dataset and tools for testing the quality of future libraries.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Accession code. Gene Expression Omnibus: GSE21739 (sequence and microarray data).

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank members of the Broad Genome Sequencing Platform for sequencing work, J. Meldrim for advice on monotemplate sequencing issues, T. Fennell for help with read processing, S. Luo and G. Schroth (Illumina) for sharing their Illumina RNA ligation protocol, L. Gaffney for assistance with figure graphics, J. Weissman for discussions and T. Liefeld and M. Reich for assistance with the GenePattern module. This work was supported by a US National Institutes of Health Director's Pioneer award, a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, the Human Frontiers Science Program, a Sloan Fellowship, the Merkin Foundation for Stem Cell Research at the Broad Institute, and Howard Hughes Medical Institute (A.R.), by the US-Israel Binational Science Foundation (N.F. and A.R.), by the Canadian friends of the Hebrew University (M.Y.) and by US National Human Genome Research Institute grant HG03067-05 (C.N.).

AUTHOR CONTRIBUTIONS

J.Z.L., M.Y., X.A., D.A.T., N.F. and A.R. wrote the paper. J.Z.L., M.Y., X.A., C.N., D.A.T., N.F., A.G. and A.R. assisted in editing the paper. D.A.T. prepared the poly(A)$^+$ RNA. J.Z.L. and X.A. prepared the cDNA libraries. M.Y., N.F. and A.R. developed and performed the computational analysis. J.Z.L., X.A., M.Y., N.F. and A.R. conceived the research.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
2. Wilhelm, B.T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239–1243 (2008).
3. Denoeud, F. *et al.* Annotating genomes with massive-scale RNA sequencing. *Genome Biol.* **9**, R175 (2008).
4. Yassour, M. *et al.* Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc. Natl. Acad. Sci. USA* **106**, 3264–3269 (2009).
5. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517 (2008).
6. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
7. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).

8. Wang, E.T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
9. Sultan, M. *et al.* A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**, 956–960 (2008).
10. Guttman, M. *et al. Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28**, 503–510 (2010).
11. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
12. Core, L.J., Waterfall, J.J. & Lis, J.T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).
13. Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* **37**, e123 (2009).
14. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. & Weissman, J.S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
15. He, Y., Vogelstein, B., Velculescu, V.E., Papadopoulos, N. & Kinzler, K.W. The antisense transcriptomes of human cells. *Science* **322**, 1855–1857 (2008).
16. Schaefer, M., Pollex, T., Hanna, K. & Lyko, F. RNA cytosine methylation analysis by bisulfite sequencing. *Nucleic Acids Res.* **37**, e12 (2009).
17. Jaffe, D.B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91–96 (2003).
18. Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033–1037 (2009).
19. Guo, J., Wu, T., Bess, J., Henderson, L.E. & Levin, J.G. Actinomycin D inhibits human immunodeficiency virus type 1 minus-strand transfer in *in vitro* and endogenous reverse transcriptase assays. *J. Virol.* **72**, 6716–6724 (1998).
20. Gentleman, R., Carey, V., Huber, W., Irizarry, R. & Dudoit, S. (eds.). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, 473 (Springer, Secaucus, NJ, 2005).
21. Yang, Y.H. *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, e15 (2002).
22. Croucher, N.J. *et al.* A simple method for directional transcriptome sequencing using Illumina technology. *Nucleic Acids Res.* **37**, e148 (2009).
23. Lipson, D. *et al.* Quantification of the yeast transcriptome by single-molecule sequencing. *Nat. Biotechnol.* **27**, 652–658 (2009).
24. Ozsolak, F. *et al.* Direct RNA sequencing. *Nature* **461**, 814–818 (2009).
25. Affymetrix / Cold Spring Harbor Laboratory ENCODE Transcriptome Project. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* **457**, 1028–1032 (2009).
26. Li, H. *et al.* Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. *Proc. Natl. Acad. Sci. USA* **105**, 20179–20184 (2008).
27. Mamanova, L. *et al.* FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat. Methods* **7**, 130–132 (2010).
28. Linsen, S.E. *et al.* Limitations and possibilities of small RNA digital gene expression profiling. *Nat. Methods* **6**, 474–476 (2009).
29. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
30. Zhu, Y.Y., Machleder, E.M., Chenchik, A., Li, R. & Siebert, P.D. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* **30**, 892–897 (2001).
31. Armour, C.D. *et al.* Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat. Methods* **6**, 647–649 (2009).

ONLINE METHODS

Yeast RNA preparation. We grew *S. cerevisiae* strains Bb32 and BY4741 to mid-log phase. We used mid-log phase RNA from Bb32 for the original RNA ligation and SMART libraries; other libraries were made from a single sample of BY4741 RNA (the two strains are closely related and interchangeable for this study). We made one library (hybrid) from post–diauxic shift BY4741 RNA (slightly impacting its performance in expression profiling and not otherwise). We isolated total and poly(A)$^+$ RNA and treated it with Turbo DNA-free (Ambion) as described[4].

RNA ligation library. We created the library using a previously described method[29] starting from 1.2 µg of poly(A)$^+$ RNA with these modifications. We fragmented RNA by incubation at 70 °C for 8 min in 1× fragmentation buffer (Ambion) and isolated 65–80-nucleotide RNA fragments from a gel. We reverse-transcribed RNA with SuperScript III (Invitrogen) at 55 °C and amplified the cDNA with Herculase (Stratagene) in the presence of 5% DMSO for 16 cycles of PCR followed by a cleanup with 1.8 volumes of AMPure beads (Agencourt) rather than gel purification.

Illumina RNA ligation library. The Illumina method used a preadenylated 3′ adaptor, which enables the subsequent ligation of the 5′ adaptor without an intermediate purification step. Our method has been modified from the version provided by Illumina. We created our libraries starting from 100 ng of poly(A)$^+$ RNA as follows. We decapped RNA by adding 10 U of tobacco acid pyrophosphatase (Epicentre), 1 µl of 10× buffer, 40 U of RNaseOut (Invitrogen) and water in a 10-µl reaction, and incubated it at 37 °C for 90 min, followed by extraction with 25:24:1 phenol:chloroform:isoamyl alcohol (PCIA; Invitrogen), ethanol precipitation and resuspension in 16 µl of H_2O. We fragmented decapped RNA by heating at 94 °C for 6 min in 1× fragmentation buffer (Affymetrix), followed by ethanol precipitation and resuspension in 16 µl of H_2O. We 3′ dephosphorylated fragmented RNA by adding 2 µl of 10× phosphatase buffer, 5 U of Antarctic phosphatase (New England Biolabs (NEB)) and 40 U of RNaseOut and incubating at 37 °C for 30 min followed by 5 min at 65 °C before chilling on ice. We 5′ phosphorylated the RNA by adding 5 µl of 10× PNK buffer, 20 U of T4 polynucleotide kinase (NEB), 5 µl of 10 mM ATP (Epicentre), 40 U of RNaseOut, 17 µl of water and incubating at 37 °C for 60 min. We adjusted the reaction volume to 100 µl with water and cleaned up with the RNeasy MinElute kit (Qiagen) following the instructions of the manufacturer except 400 µl of 100% ethanol were used in step two. We concentrated RNA to 6 µl by Vacufuge (Eppendorf), followed by mixing with 1 µl 1× v1.5 sRNA 3′ adaptor (Illumina), incubating at 70 °C for 2 min and chilling on ice for 2 min. We prepared the 3′ ligation with this RNA adaptor mix, 1 µl 10× T4 RNA ligase 2 truncated reaction buffer, 0.8 µl of 100 mM $MgCl_2$ (Sigma), 20 U of RNaseOut, 300 U of T4 RNA ligase 2, truncated (NEB) and incubated at 22 °C for 1 h. We denatured 1 µl of SRA 5′ adaptor (Illumina) at 70 °C for 2 min and chilled it on ice before combining it with the 3′ adaptor–ligated RNA, 1 µl of 10 mM ATP and 1 µl of T4 RNA ligase (Illumina) and incubating at 20 °C for 1 h. We combined 12 µl of this doubly adaptor-ligated RNA with 3 µl of 0.2× SRA reverse transcription (RT) primer (Illumina), followed by incubation at 70 °C for 2 min, and chilling on ice. We synthesized single-stranded cDNA with this RNA primer mix by adding 6 µl 5× first-strand buffer, 6 µl 100 mM DTT, 1.5 µl 12.5 mM dNTPs, 600 U SuperScript III and 30 U SUPERase-In (Ambion) and incubating for 1 h at 55 °C. We divided the cDNA into two aliquots that we processed with different size selection methods yielding libraries with differing insert lengths. In the first method, we mixed two-thirds of the cDNA with 5 U RNase H (NEB), incubated at 37 °C for 1 h and 75 °C for 15 min, PCIA extracted, ethanol precipitated and resuspended in 10 µl H_2O. We selected single-stranded cDNA ranging in size from 175 to 225 nt on a Criterion 10% TBE-urea gel (Bio-Rad). We crushed the gel slice and eluted with 250 µl 0.3 M NaCl by rotating at room temperature (20–23 °C) for over 4 h. We filtered the crushed gel slice and buffer mixture through a Spin-X cellulose acetate filter (Corning) by centrifugation at 16,000*g* for 3 min. We ethanol-precipitated the eluate and resuspended it in 10 µl RNase-free water. We prepared a 50 µl PCR with 5 µl water, 25 µl 2× Phusion High-Fidelity Master Mix with GC buffer (NEB), 13 µl 5 M betaine (Sigma), 1 µl each primer GX1.0 and 2.0 (Illumina) and 5 µl size-selected cDNA. Thermocycling conditions were: 30 s at 98 °C, 14 cycles of 98 °C for 10 s, 60 °C for 30 s, and 72 °C for 15 s, followed by 10 min at 72 °C. We removed PCR primers using 1.8 volumes of AMPure beads. This generated a cDNA library ranging in size from 180 to 240 base pairs (bp) (insert size of 110–170 bp). In the second method (SPRI), we used one-sixth of the cDNA without size selection in a 50 µl PCR prepared as in the first method. We purified the PCR product twice with 1.3 volumes of AMPure beads to generate a library ranging in size from 120 to 250 bp (insert size of 50–180 bp).

SMART library. We adapted the SMART method[30] developed for SOLiD[32] to Illumina Genome Analzyer sequencing. In our method, reverse transcriptase–primed cDNA synthesis with an oligonucleotide comprised of an Illumina adaptor sequence 5′ of a random hexamer, added three nontemplate cytosine nucleotides at the 3′ end of the cDNA, followed by template switching to a second oligonucleotide containing a second Illumina adaptor sequence 5′ of three guanine ribonucleotides. Specifically, we created the SMART library starting from 100 ng of poly(A)$^+$ RNA as follows. We fragmented RNA by heating at 98 °C for 40 min in 0.2 mM sodium citrate, pH 6.4 (Ambion), followed by concentrating it to 3.5 µl, mixing with 1 µl 2 µM SMART tagged random primer, incubating at 70 °C for 10 min and chilling on ice for 2 min. (Sequences of all custom primers used in this study are listed in **Supplementary Table 5**.) We synthesized first-strand cDNA from this RNA primer mix by adding 2 µl 5× buffer, 1 µl 20 mM DTT, 0.5 µl 10 mM dNTPs, 50 U SMARTScribe reverse transcriptase (Clontech), and 10 U SUPERase-In and incubating at room temperature for 10 min followed by 45 min at 42 °C. We denatured 1 µl 10 µM 5′ SMART oligo at 70 °C for 5 min and added it to the cDNA synthesis reaction, which we then incubated at 42 °C for another 15 min and chilled on ice. We cleaned up the cDNA using 1× volume of AMPure beads and eluted with 20 µl of elution buffer (Qiagen). We prepared a 160 µl PCR with 96 µl water, 16 µl 10× HF 2 PCR buffer, 16 µl 10× HF 2 dNTP mix, 6.4 µl 25 µM primer PE 1.0 (Illumina), 6.4 µl 5µM SMART reverse primer, 3.2 µl 50× Advantage-HF 2 polymerase mix (Clontech) and 16 µl cDNA. Thermocycling conditions were: 5 min at 94 °C, 19 cycles of 94 °C for 15 s and 68 °C for 30 s. We PCIA extracted, ethanol precipitated and resuspended the PCR products in 10 µl

doi:10.1038/nmeth.1491

H_2O. We selected PCR products ranging in size from 220 to 420 bp on a 4% NuSieve 3:1 agarose (Lonza) TAE gel and purified them with the MinElute Gel Extraction kit (Qiagen).

SMART-RNA ligation 'hybrid' library. The SMART–RNA ligation ('hybrid') library combined ligation of an RNA adaptor to the 3′ end of fragmented RNA with SMART's template switching to attach a second adaptor at the 3′ end of the cDNA. We created the library starting from 500 ng poly(A)$^+$ RNA as follows. We fragmented RNA as described for the SMART library and dephosphorylated it with 1.5 µl 10× buffer 3 (NEB), 15 U calf intestinal alkaline phosphatase (NEB), 40 U RNaseOut and water in a final volume of 15 µl for 1 h at 37 °C and then chilled it on ice. We PCIA extracted, ethanol precipitated and resuspended this RNA in 5 µl H_2O. We denatured this RNA and 1 µl 4 µM 3′ RNA adaptor oligo at 70 °C for 2 min, chilled them on ice, combined them with 40 U RNaseOut, 1 µl 100% DMSO (NEB), 10 U T4 RNA ligase (Promega), and 1 µl 10× T4 RNA ligase buffer, and incubated for 6 h at 20 °C and then 4 h at 4 °C. We cleaned up adaptor-ligated RNA using 1.8 volumes of RNAClean beads (Agencourt) and eluted with 10 µl water. We repeated this process to minimize the amount of unincorporated RNA adaptor oligos. We used half of this RNA for cDNA synthesis as described for the SMART library, except we used 1 µl 10 µM Hybrid reverse transcription primer in the reverse transcription reaction for 45 min at 42 °C before adding the 5′ Hybrid oligo. We degraded RNA by adding 2.5 U RNase H, 1.5 µl 10× RNase H buffer, 3 µl water and incubating at 37 °C for 1 h. We PCIA extracted, ethanol precipitated and resuspended the cDNA in 6 µl H_2O. We selected single stranded cDNA ranging in size from 300 to 500 nt on a Criterion 5% TBE-Urea gel and eluted it as described for the Illumina RNA ligation library. We prepared a 125 µl PCR with 2.5 µl water, 62.5 µl 2× Phusion High-Fidelity Master Mix with GC buffer, 50 µl 5 M betaine, 2.5 µl each 25 µM Hybrid forward and Hybrid reverse primers and 5 µl size-selected cDNA. Thermo-cycling conditions were: 30 s at 98 °C, 5 cycles of 98 °C for 10 s, 50 °C for 30 s and 72 °C for 30 s, 13 cycles of 98 °C for 10 s, 65 °C for 30 s and 72 °C for 30 s, followed by 5 min at 72 °C. We removed PCR primers using 1.8 volumes of AMPure beads.

NNSR library. We modified the original NSR method[31], which creates a strand-specific library, by replacing the 'not so random' primers for cDNA synthesis with random (or 'not not so random') primers. The NNSR method used two different primers, each comprised of a different adaptor sequence and random hexamers, for first- and second-strand cDNA synthesis. We created the NNSR library starting from 250 ng of poly(A)$^+$ RNA. We concentrated RNA to 5 µl, mixed it with 2 µl of 100 µM tagged first-strand NNSR primers, incubated them at 65 °C for 5 min and placed them on ice. We synthesized first-strand cDNA with this RNA primer mix by adding 4 µl of 5× first-strand buffer, 2 µl of 100 mM DTT, 1 µl of 10 mM dNTPs, 4 µg actinomycin D (USB), 200 U SuperScript III and 20 U SUPERase-In and incubating at 45 °C for 30 min followed by 15 min at 70 °C. We PCIA extracted twice, ethanol precipitated and resuspended first-strand cDNA in 10 µl H_2O. We treated it with RNase H in 1× RNase H buffer at 37 °C for 20 min followed by 15 min at 75 °C, clean up using 1.8 volumes of RNAClean beads and elution with 30 µl water. We synthesized second-strand cDNA in a 100 µl reaction by adding 10 µl 10× buffer 2 (NEB), 5 µl 10 mM dNTPs, 20 U Klenow Fragment (3′ to 5′ exo$^-$; NEB), 10 µl of 100 µM tagged second-strand NNSR primers and water and incubating at 37 °C for 30 min. We purified the cDNA with 1.8 volumes of AMPure beads. We prepared a 50 µl PCR with 9.5 µl water, 10 µl of 5× reaction buffer 2, 2.5 µl of 10 mM dNTP mix, 5 µl of 25 mM $MgCl_2$, 5 µl of each 10 µM NNSR forward and NNSR reverse primers, 0.5 µl of Expand^{PLUS} enzyme (Roche) and 12.5 µl cleaned up cDNA. Thermo-cycling conditions were: 2 min at 94 °C, two cycles of 94 °C for 10 s, 40 °C for 2 min and 72 °C for 1 min; eight cycles of 94 °C for 10 s, 60 °C for 30 s and 72 °C for 1 min; four cycles of 94 °C for 15 s, 60 °C for 30 s and 72 °C for 1 min with an additional 10 s added at each cycle; 72 °C for 5 min. We purified PCR products using 1.8 volumes of AMPure beads. We selected PCR products ranging in size from 325 to 525 bp on a Criterion 10% TBE gel and eluted them as described for the Illumina RNA ligation library.

We made a second NNSR library in parallel without actinomycin D.

Bisulfite libraries. We created the 'H' and 'S' bisulfite libraries using two previously described methods[15,16], respectively, starting from 1 µg of poly(A)$^+$ RNA with the following modifications. The S library bisulfite reaction followed the 6× cycles for human 28S RNA treatment[16] and was ethanol precipitated before and after desulfonation. We cleaned up the H library bisulfite reaction with an Amicon Ultra-15 3k MWCO filter (Millipore) centrifuged at 4,000*g* at 25 °C for 50 min. In subsequent steps we followed a previously published procedure[15], except as noted. We synthesized first-strand cDNAs from 100 ng of bisulfite-treated poly(A)$^+$ RNA with 1.5 µg 'random octamer' mixture, prepared as described[15], in a 40 µl reaction for 10 min at 25 °C followed by 60 min at 55 °C. We synthesized second-strand cDNA with 5× second-strand buffer (Invitrogen) in a 300 µl reaction. Because bisulfite treatment fragmented the RNA (data not shown), it was not necessary to fragment the cDNA. We prepared a paired-end library for Illumina sequencing as for the dUTP library, except that we gel-purified the final PCR products with an insert size of 160–300 bp.

dUTP library. We created the dUTP second strand library starting from 200 ng of poly(A)$^+$ RNA using a previously described method[13] with the following modifications. All reagents were from Invitrogen except as noted. We fragmented RNA as described for the SMART library, concentrated it to 5 µl, mixing with 3 µg random hexamers, followed by incubation at 70 °C for 10 min and chilling on ice. We synthesized first-strand cDNA with this RNA primer mix by adding 4 µl 5× first-strand buffer, 2 µl 100 mM DTT, 1 µl 10 mM dNTPs, 4 µg of actinomycin D, 200 U SuperScript III and 20 U SUPERase-In, incubating at room temperature for 10 min followed by 1 h at 55 °C. We cleaned up first-strand cDNA by PCIA extraction twice, ethanol precipitation with 0.1 volumes 5 M ammonium acetate to remove dNTPs and resuspension in 104 µl H_2O. We synthesized second-strand cDNA by adding 4 µl of 5× first-strand buffer, 2 µl of 100 mM DTT, 4 µl of 10 mM dNTPs with dTTP replaced by dUTP (Sigma), 30 µl of 5× second-strand buffer, 40 U of *Escherichia coli* DNA polymerase, 10 U of *E. coli* DNA ligase and 2 U of *E. coli* RNase H, and incubating at 16 °C for 2 h. We prepared a paired-end library for Illumina sequencing according to the instructions provided, with the following modifications. First, we ligated five times less adaptor mix to the cDNAs. Second, we incubated 1 U USER (NEB) with

180 to 480 bp size-selected, adaptor-ligated cDNA at 37 °C for 15 min followed by 5 min at 95 °C before PCR. Third, we performed PCR with Phusion High-Fidelity DNA polymerase with GC buffer and 2 M betaine. Fourth, we removed PCR primers using 1.8 volumes of AMPure beads.

In addition, we made a second cDNA library in parallel with 2.7 μg random hexamers plus 1.1 μg anchored oligo$(dT)_{20}$ (Invitrogen) in the first-strand synthesis.

'Control' (non–strand-specific) library. We prepared a control library that used dTTP instead of dUTP for second-strand cDNA synthesis at the same time as the dUTP library. In addition, we made a second control cDNA library in parallel with 2.7 μg of random hexamers plus 1.1 μg of anchored oligo$(dT)_{20}$ in the first-strand synthesis.

Illumina sequencing. We sequenced each of the cDNA libraries with an Illumina Genome Analyzer II (one or two lanes of 76 base reads) using the standard SBS3 and SBS8 sequencing primers (Illumina), except as noted below. We sequenced the SMART library with the standard SBS3 primer for the first read and the custom SBS11 primer for the second read; both reads were 51 bases. We sequenced the RNA ligation and Illumina RNA ligation libraries with the small RNA sequencing primer (Illumina). The NNSR, SMART and Hybrid libraries have a short, identical sequence at the start of every read that leads to 'monotemplate' issues during cluster image processing (**Supplementary Note 2**).

Library read mapping. For SMART, Hybrid and NNSR libraries, we trimmed reads before mapping, to remove specific adaptor-derived bases expected at the start of the read. We mapped reads using Arachne[17]. We mapped reads in single end libraries uniquely, allowing up to four mismatches. We first mapped reads in paired-end libraries non-uniquely allowing up to four mismatches and then searched for unique pairing of the non-unique read mappings (a single pair of mappings on the same chromosome, up to 500 bp apart, with reads on opposite strands). For the bisulfite libraries, we first converted each 'C' in the genome to 'T', resulting in two pseudo-genomes (one per strand), to which the reads were mapped (a unique read mapped to a single location in exactly one of those pseudo genomes).

Read sampling and trimming. We sampled 2.5 million mapped read 'starts' from the aligned reads of each library, with the exception of the SMART and Bisulfite 'H' libraries where we used all reads (~0.9 million and 2.1 million reads, respectively), owing to their repeated low yields. (Resampling these libraries to 2.5 million did not change the results substantially, data not shown.) As the libraries have various read lengths, we used only the first 36 bases of each mapped read (the shortest fragment length in our compendium). We used the sampled 36 base extended coverage for all subsequent method comparison.

Library complexity. We calculated the fraction of reads starting at a distinct (unique) genomic location. In paired libraries we measured the fraction of pairs whose combination of start and end locations was unique, as a proxy for the number of unique cDNAs loaded on the sequencer.

Strand specificity. We used the known annotation from (*Saccharomyces* Genome Database (SGD), http://www.yeastgenome.org/; downloaded in November 2007), and published estimates of UTR lengths[18], or when absent an estimation of 100 bp for each of the UTRs. We considered only high-quality annotations ('verified' or 'uncharacterized'; SGD) and excluded all regions with annotated overlapping transcripts (at UTRs or ORFs) and all genes designated as 'dubious'. We calculated the number of reads that map to the sense and opposite strand of known transcripts.

Evenness of coverage. We used the known annotation from SGD, divided the length of each gene into 100 bins of equal length and calculated the relative coverage in each bin compared to the entire gene. We averaged across all 'verified' and 'uncharacterized' annotated genes.

Continuity of coverage. We measured for each gene the fraction of the gene's total length that had no read coverage. We plotted these values against the relative expression of the gene based on a 'pooled' library (below) and calculated in each plot the Lowess fit of these data (Matlab version 2009b; MathWorks). For each gene, we also counted the number of segments of length 5 bp or longer that had no read coverage. We averaged these measurements across all genes, weighting by the relative expression of each gene.

Comparison to *S. cerevisiae* annotation of 5′ and 3′ ends. Conservatively, we used known annotation of verified and uncharacterized genes (SGD). For each end, we measured the number of genes where a window of ten bp around the translation start and end sites was fully covered by aligned reads.

Expression. We used three standards: microarray data, the 'control' library and a 'pooled' library with 2.5 million sampled mapped reads from each of nine strand-specific libraries (RNA ligation, Illumina RNA ligation, SMART, Hybrid, NNSR, bisulfite, our dUTP, published dUTP and 3′ split adaptor). For each library, we calculated the relative expression level of known genes (SGD) by calculating the mean coverage over the coding region length, and normalizing it to a distribution over all genes[4]. We compared each library to each reference using the Pearson correlation coefficient and the r.m.s. error measures. We also generated scatter, Q-Q and MA plots for each library-reference pair.

MA and Q-Q plots. Both plots compare two sets of data (D_1, D_2). An MA plot displays the $\log_2(D_1) + \log_2(D_2)$ versus $\log_2(D_1) - \log_2(D_2)$. If the samples are very similar, they should be close to the $y = 0$ axis regardless of the x-axis position. A Q-Q plot displays a quantile-quantile plot of D_1 (x axis) and D_2 (y axis). If the samples were drawn from the same distribution, the plot should be a straight line.

Microarray data. Microarray data preparation methods are described in **Supplementary Note 3**.

32. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5**, 613–619 (2008).

doi:10.1038/nmeth.1491