Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2011 March 01.

Published in final edited form as:

Nat Methods. 2010 September ; 7(9): 709–715. doi:10.1038/nmeth.1491.

Comprehensive comparative analysis of strand-specific RNA sequencing methods

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

[1]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142 USA

[2]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139 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, MA 02140 USA

Abstract

Strand-specific, massively-parallel cDNA sequencing (RNA-Seq) is a powerful tool for novel transcript discovery, genome annotation, and expression profiling. Despite multiple published methods for strand-specific RNA-Seq, 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 novel 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 identify 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

Users may view, print, copy, download and text and data- mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use: http://www.nature.com/authors/editorial_policies/license.html#terms

Correspondence should be addressed to J.Z.L. (jlevin@broadinstitute.org) and A.R. (aregev@broad.mit.edu).
[6]These authors contributed equally to this work

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

AUTHOR CONTRIBUTIONS

J.Z.L., M.Y., X.A., D.A.T., N.F., and A.R. wrote the paper. All the authors assisted in editing the paper. D.A.T. prepared the polyA$^+$ 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.

AOP:

Authors compare quality metrics of libraries from seven strand-specific RNA-Seq methods in terms of complexity, strand specificity, evenness and continuity of coverage, and expression profiling. They provide a computational pipeline to compare these metrics from any RNA-Seq protocol.

Issue

Authors compare quality metrics of libraries from seven strand-specific RNA-Seq methods in terms of complexity, strand specificity, evenness and continuity of coverage, and expression profiling. They provide a computational pipeline to compare these metrics from any RNA-Seq protocol.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

comprehensive benchmark, and our computational pipeline is applicable for assessment of future protocols in other organisms.

INTRODUCTION

Recent advances in massively-parallel cDNA sequencing (RNA-Seq) have opened the way for comprehensive analysis of any transcriptome1. In principle, RNA-Seq allows us to study all expressed transcripts, with three key goals: first, annotating the structures of all transcribed genes including their 5' and 3' ends and all splice junctions2–4; second, quantifying the level of expression of each transcript5,6; and third, measuring the level of alternative splicing7–11.

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' ends4, or splice site orientation in eukaryotic genomes4,10,11.

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 roles12, determine the transcribed strand of other non-coding RNAs, demarcate the exact boundaries of adjacent genes transcribed on opposite strands, and resolve the correct expression levels of coding or non-coding overlapping transcripts. These tasks are particularly challenging in small microbial genomes, prokaryotic and eukaryotic, where genes are densely coded, with overlapping UTRs (untranslated regions) or ORFs, and where splice site information is limited or non-existent.

A host of methods has been recently developed for strand-specific RNA-Seq (Fig. 1), that 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 respectively. A second class of methods relies on marking one strand by chemical modification, either on the RNA itself by bisulfite treatment (Fig. 1b) 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.

While 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 *S. cerevisiae* polyA$^+$ RNA, we built a compendium of libraries using these protocols (Fig. 1) and Illumina sequenced each of them 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 identify 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 comparative compendium of strand-specific RNA-Seq

We evaluated a compendium of 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 . In addition to data from our own libraries, we also compiled comparable data for two published published libraries, a dUTP library13, and a library based on an eighth method from the differential adaptor class14 (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 NNSR library protocol, two published variations of the bisulfite library protocol ("H" and "S" Online Methods15,16), 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 no substantial differences between the two libraries were observed in our analyses.

We used each method to prepare a cDNA library for Illumina sequencing from *S. cerevisiae* polyA$^+$ RNA. We chose *S. cerevisiae* since 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 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 (Fig.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2c) and continuity of coverage at annotated transcripts(Fig. 2d), 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 the control library. We focus on only one variation per method unless there are substantial differences in performance between variations. The full evaluation results are provided in the Supplementary Material (Supplementary Tables 1–2 and Supplementary Figs. 2–4).

Equal sampling of reads enables direct library comparisons

We mapped each library's reads to the *S. cerevisiae* genome using Arachne17. 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 reads17 (Online Methods).

The libraries displayed a broad range of yields, measured by the total number of reads, and 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 fraction of paired-end mapped reads (% Paired, Supplementary Table 1). The Illumina RNA ligation – SPRI library, which was prepared using Solid Phase Reversible Immobilization (SPRI)-based size selection, had a smaller percentage of unique reads than the Illumina RNA ligation library, which was prepared using gel-based size selection (35% *vs*. 59%, Supplementary Table 1). This was likely due to the difficulty in physically removing cDNAs shorter than 76 bp with the SPRI method resulting in the ends of sequencing reads containing an Illumina adaptor sequence that could not align to the yeast genome. Indeed, when these reads were trimmed to 51 bases, the percentage of aligning reads improved dramatically (data not shown). In the text and figures below, we report results only for the Illumina RNA ligation library, which was 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. Since many of our measures are sensitive to read quantity and length, we used sampling to obtain the same number of reads from each library (Online Methods). Unless specifically noted, all subsequent comparisons were conducted with 2.5 million sampled reads from each library. The SMART library had only 930,686 reads, because of repeated poor yields, but with the exception of complexity, we obtain 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 re-sampling 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, 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Supplementary Table 2), the best complexity was obtained by 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 over-estimate 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), since cDNAs that share 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). Consistent with recent studies18, we conservatively assumed that most of the *S. cerevisiae* genes do not have transcription from the antisense strand, and used the fraction of reads mapped to the opposite (antisense) strand of known transcripts as a measure for strand specificity (Fig. 2b, Supplementary Table 2 and 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. 4a and Supplementary Fig. 5). Only 0.47–0.63% of the reads mapped to the antisense strand in the four best performers. 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 since it inhibits both second-strand synthesis and template switching19 (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 assembly3,4, 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. .

To measure evenness of coverage for each library, we calculated the average of the coefficient of variation (CV) of gene coverage for the top 50% expressed genes (Fig. 2c, Fig.4a, Supplementary Fig. 5 and Supplementary Table 2). The most even coverage was found for the 3' split adaptor method14 (average CV 0.54), closely followed by the dUTP approach (average CV 0.64 in the original dataset13, 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 is broken, where we define a break as a stretch of at least five bases without read coverage (Fig. 2d, Fig. 5a and Supplementary Table 2). We then averaged this measure across all genes, weighting by the relative expression of each

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

gene (low expressed genes are expected to be less covered and more segmented). The best performing methods by this measure were the 3' split adaptor method14 (2.29 segments/gene), the dUTP libraries (2.41 and 2.48 segments/gene, with published data13 and in our hands, respectively), and the Illumina RNA ligation libraries (2.61 segments/gene).

Second, we calculated the fraction of bases without coverage in each transcript (Fig. 2d, Fig. 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 level 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 the published data13) and 3' split adaptor (Fig. 5d) methods performed best.

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 (Fig. 2d and Fig. 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 specific biases in 5' and 3' coverage (Fig. 4b,c, Supplementary Fig. 3 and Supplementary Table 2). The NNSR library showed 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, respectively), followed by the dUTP method (62% and 73%) (Fig. 4c and Supplementary Table 2). Surprisingly, the addition of oligo dT primers for reverse transcription for the dUTP method, both in our results and in the published data13, 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 polyA 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 RNA sample *vs*. genomic DNA using Agilent arrays (Online Methods). We calculated the expression level of each gene as its length-normalized read coverage, and normalized all values for the total number of reads.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

We used several standard quality measures20 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 error (RMSE) of the expression measurements in each library using the reference measurement as the expected level (Fig. 6b and Supplementary Table 2); and scatter, Q-Q, and MA21 plots (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 RMSE 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 *vs*. 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 RMSE measures, but with noticeably broader scatter across the expression range (Supplementary Fig. 4). The lowest 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 due to well-known technical limitations, incomplete method development, or high similarity to tested methods. These excluded protocols comprise single-stranded cDNA library synthesis22 (due to chimeric cDNAs created); deep sequencing of ribosome-protected mRNA fragments14 (because cDNA lengths are too short and the original method involves a complex procedure for RNA preparation; we have included published data from the non-protected library designated as the 3' split adaptor method; Supplementary Fig. 1); Helicos single-molecule digital gene expression23 and Direct RNA Sequencing24 (coverage heavily biased to the 5' or 3' ends of transcripts, respectively; the latter is currently still under development); ligation of adaptor to 5' end and C-tailing at 3' end of RNA25 and the double-random priming method26 (similar to NNSR). We did not include FRT-seq27 and SOLiD™ Whole Transcriptome Analysis Kit28 because they are similar to the two RNA ligation methods we tested and it would be difficult to distinguish differences due to library construction methods from those due to the different sequencing methods.

In addition to the formal criteria evaluated by our pipeline, 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-sized fragments (of greater than 152 bp) needed for paired-end sequencing with 76 base reads, since 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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The libraries also vary in the facility of computational analysis, in particular at early processing steps. The bisulfite method is the most computationally challenging, since reads must be aligned to two reference 'genomes' that have all the C bases converted to T 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 CV, 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), since these may be complex. We note, however, that the amount of starting material does not correlate with library complexity (Supplementary Tables 2 and 4).

The dUTP protocol provides 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 reconstruction10,11. The 3' split adaptor method14 excelled in measures critical for genome annotation, but is less well-suited for expression profiling. Finally, our compendium and analysis pipeline, which will be available online (www.broadinstitute.org/regev/rnaseqmethods) and as a GenePattern module (http://www.broadinstitute.org/cancer/software/genepattern/), provide important resources including a general benchmarking dataset and tools for testing the quality of future libraries.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank the Broad Genome Sequencing Platform for sequencing work, Jim Meldrim for advice on monotemplate sequencing issues, Tim Fennell for help with read processing, Shujun Luo and Gary Schroth (Illumina) for sharing their "Illumina RNA ligation protocol," Leslie Gaffney for assistance with figure graphics, and Jonathan Weissman for discussions. Work was supported by an 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 and Howard Hughes Medical Institute (AR), by the US-Israel Binational Science Foundation (NF

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and AR), by the Canadian friends of the Hebrew University (MY), and by National Human Genome Research Institute grant HG03067-05 (CN). AR is a researcher of the Merkin Foundation for Stem Cell Research at the Broad Institute.

REFERENCES

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 2009; 10:57–63. [PubMed: 19015660]
2. Wilhelm BT, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature. 2008; 453:1239–1243. [PubMed: 18488015]
3. Denoeud F, et al. Annotating genomes with massive-scale RNA sequencing. Genome Biol. 2008; 9:R175. [PubMed: 19087247]
4. Yassour M, et al. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. Proc. Natl. Acad. Sci. USA. 2009; 106:3264–3269. [PubMed: 19208812]
5. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008; 18:1509–1517. [PubMed: 18550803]
6. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods. 2008; 5:621–628. [PubMed: 18516045]
7. 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; 40:1413–1415. [PubMed: 18978789]
8. Wang ET, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; 456:470–476. [PubMed: 18978772]
9. Sultan M, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science. 2008; 321:956–960. [PubMed: 18599741]
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. 2010; 28:503–510. [PubMed: 20436462]
11. Trapnell C, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 2010; 28:511–515. [PubMed: 20436464]
12. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008; 322:1845–1848. [PubMed: 19056941]
13. Parkhomchuk D, et al. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. 2009; 37:e123. [PubMed: 19620212]
14. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009; 324:218–223. [PubMed: 19213877]
15. He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW. The antisense transcriptomes of human cells. Science. 2008; 322:1855–1857. [PubMed: 19056939]
16. Schaefer M, Pollex T, Hanna K, Lyko F. RNA cytosine methylation analysis by bisulfite sequencing. Nucleic Acids Res. 2009; 37:e12. [PubMed: 19059995]
17. Jaffe DB, et al. Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res. 2003; 13:91–96. [PubMed: 12529310]
18. Xu Z, et al. Bidirectional promoters generate pervasive transcription in yeast. Nature. 2009; 457:1033–1037. [PubMed: 19169243]
19. Guo J, Wu T, Bess J, Henderson LE, Levin JG. Actinomycin D inhibits human immunodeficiency virus type 1 minus-strand transfer in in vitro and endogenous reverse transcriptase assays. J. Virol. 1998; 72:6716–6724. [PubMed: 9658119]
20. Gentleman, R.; Carey, V.; Huber, W.; Irizarry, R.; Dudoit, S., editors. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Vol. 473. Secaucus, NJ: Springer; 2005.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

21. Yang YH, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 2002; 30:e15. [PubMed: 11842121]
22. Croucher NJ, et al. A simple method for directional transcriptome sequencing using Illumina technology. Nucleic Acids Res. 2009; 37:e148. [PubMed: 19815668]
23. Lipson D, et al. Quantification of the yeast transcriptome by single-molecule sequencing. Nat. Biotechnol. 2009; 27:652–658. [PubMed: 19581875]
24. Ozsolak F, et al. Direct RNA sequencing. Nature. 2009; 461:814–818. [PubMed: 19776739]
25. Project ACSHLET. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. Nature. 2009; 457:1028–1032. [PubMed: 19169241]
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. 2008; 105:20179–20184. [PubMed: 19088194]
27. Mamanova L, et al. FRT-seq: amplification-free, strand-specific transcriptome sequencing. Nat. Methods. 2010; 7:130–132. [PubMed: 20081834]
28. Linsen SE, et al. Limitations and possibilities of small RNA digital gene expression profiling. Nat. Methods. 2009; 6:474–476. [PubMed: 19564845]
29. Lister R, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 2008; 133:523–536. [PubMed: 18423832]
30. Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. Biotechniques. 2001; 30:892–897. [PubMed: 11314272]
31. Cloonan N, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat. Methods. 2008; 5:613–619. [PubMed: 18516046]
32. Armour CD, et al. Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. Nat. Methods. 2009; 6:647–649. [PubMed: 19668204]
33. Wapinski I, et al. Gene duplication and the evolution of ribosomal protein gene regulation in yeast. Proc. Natl. Acad. Sci. USA. 2010; 107:5505–5510. [PubMed: 20212107]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

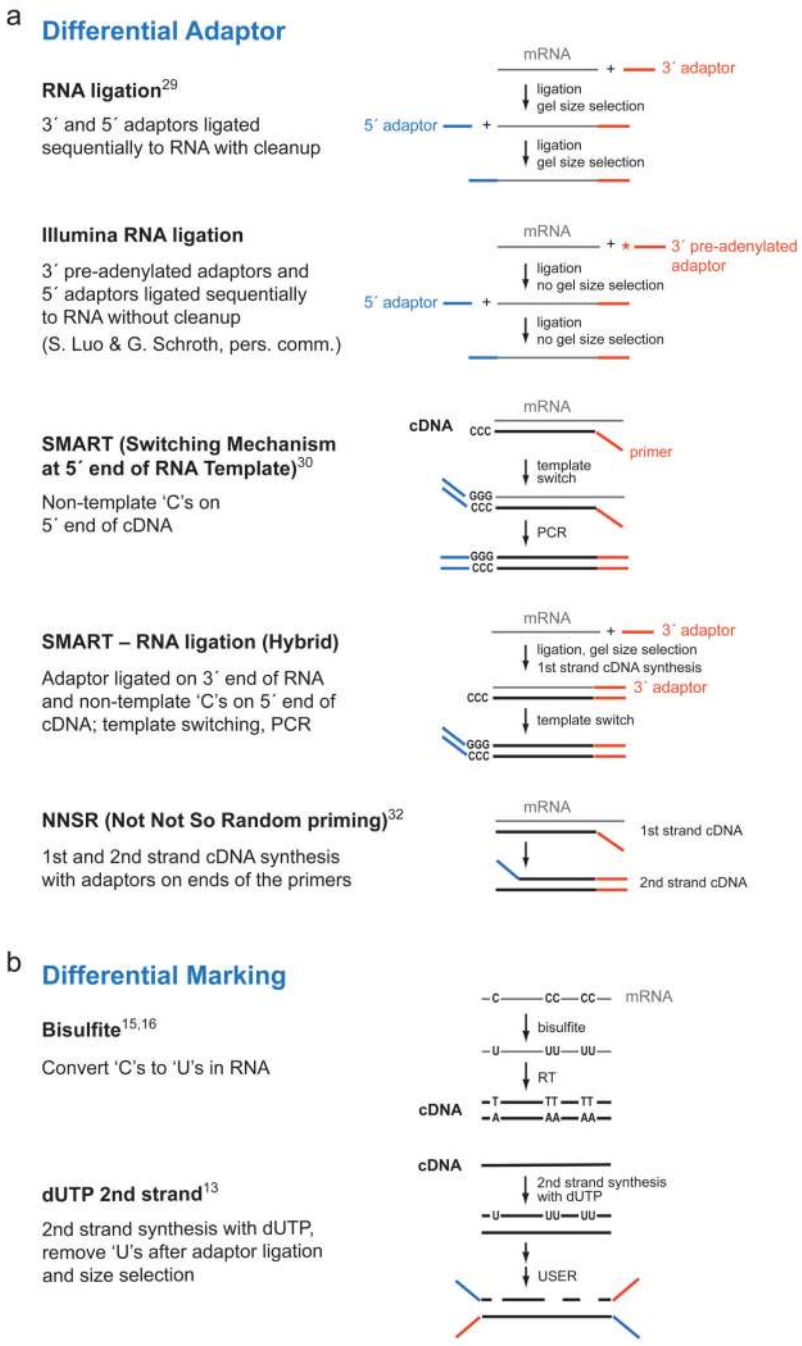

Figure 1. Methods for strand-specific RNA-Seq

Salient details for seven protocols for strand-specific RNA-Seq, differential adaptor methods (a) and differential marking methods (b). mRNA is shown in grey, and cDNA in black. For differential adaptor methods, 5' adaptors are shown in blue, and 3' adaptors in red.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

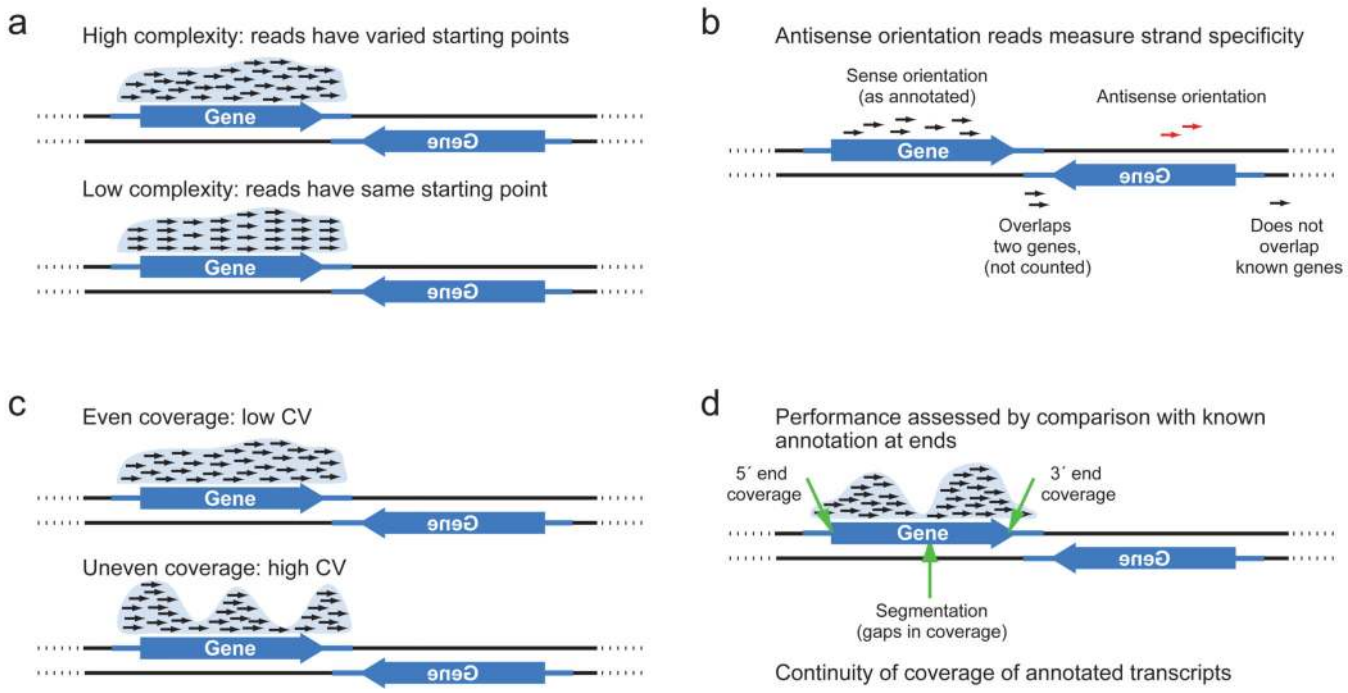

Figure 2. Key criteria for evaluation of strand-specific RNAseq libraries

Four categories of quality assessment. Double stranded genome (black parallel lines), with Gene ORF orientation (thick blue arrow) and UTRs (thin blue line), along with mapped reads (short black arrows – reads mapped to sense strand; red – reads mapped to antisense strand). (**a**) Complexity. (**b**) Strand Specificity. (**c**) Evenness of coverage. (**d**) Comparison to known transcript structure..

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

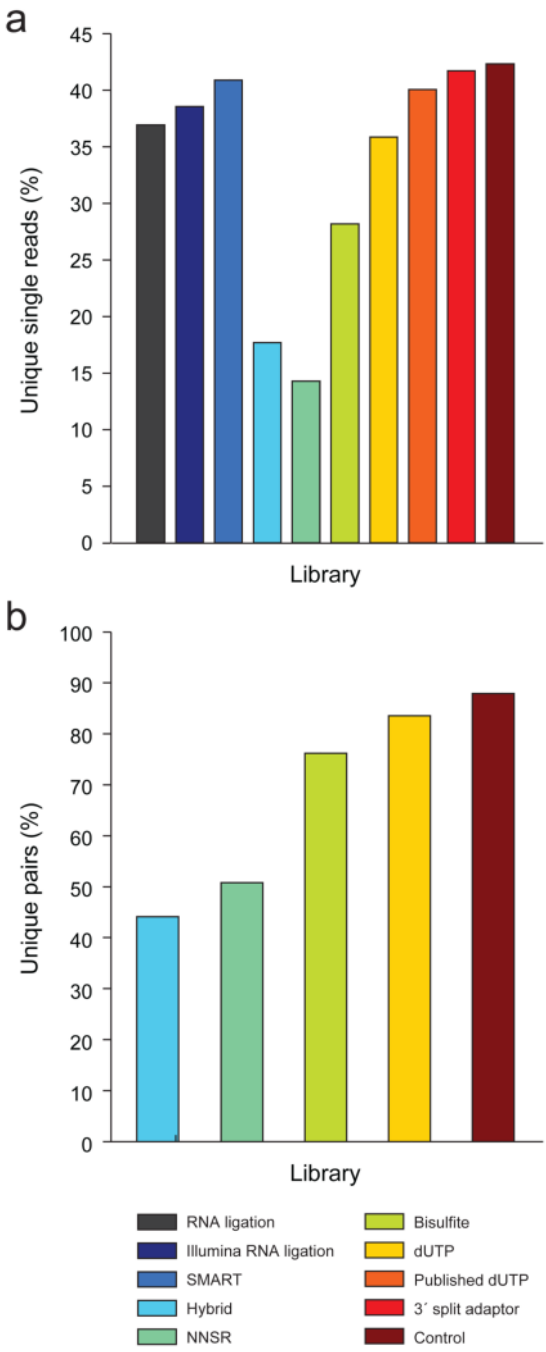

Figure 3. Complexity of single- and paired-end libraries

Bar graphs comparing library complexity by the fraction 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). Libraries are ordered as in Figure 1. Full data for all library variations are presented in Supplementary Table 2.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

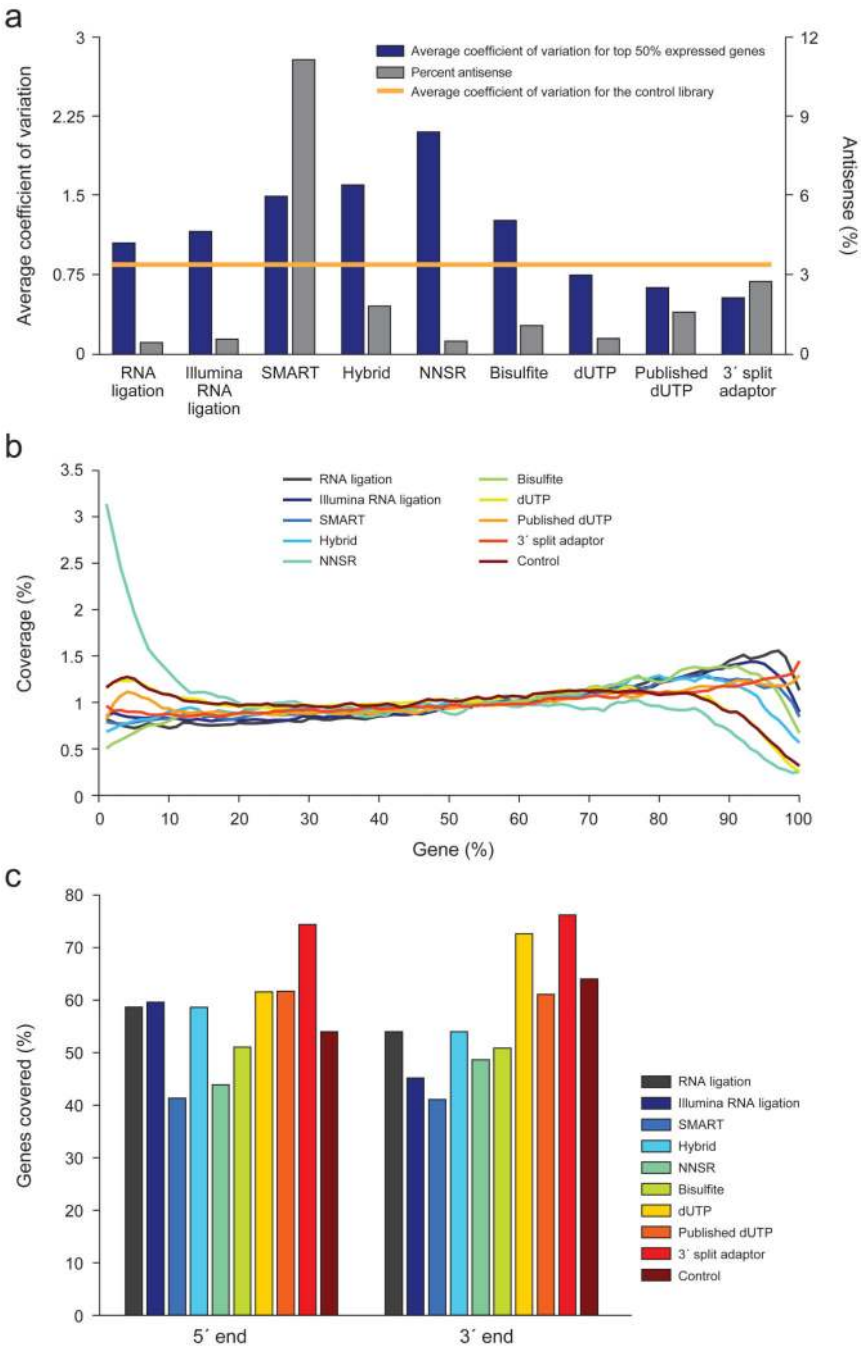

Figure 4. Strand specificity and evenness of transcript coverage

(**a**) Strand specificity (% antisense) and evenness of coverage (average coefficient of variation (CV)). The average CV of the control (non strand-specific library) is shown by an orange line. Libraries are sorted as in Figure 1. Full data for all library variations are presented in Supplementary Table 2. (**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. Full data for all library variations are presented in Supplementary Fig. 3. (**c**) 5' and 3' end coverage. Shown is the percentage of genes with 5' and 3' coverage (left and right bars, respectively;

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Online Methods) in each library. Full data for all library variations are presented in Supplementary Table 2.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

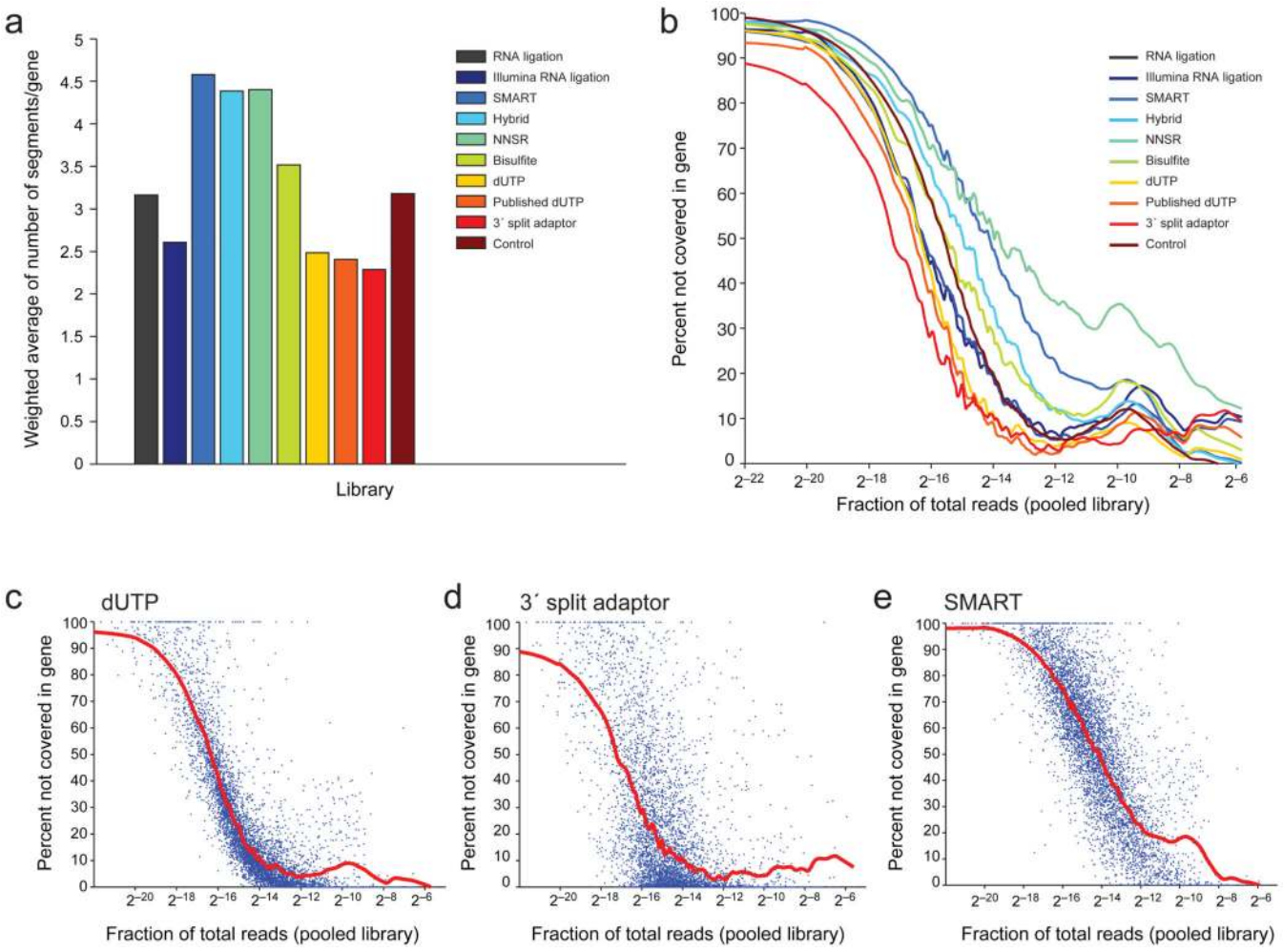

Figure 5. Continuity of transcript coverage

(**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. Full data for all library variations are presented in Supplementary Table 2. (**b–e**) Fraction of bases not covered by reads for each gene (blue dot) in the genome, plotted against the fraction of total reads for that gene in the pooled library, for the dUTP method (c), the 3' split adaptor method (d) and the SMART method (e). In each case, a Lowess fit is shown as a red curve, with fits from all libraries shown in (b). Full data for all library variations are presented in Supplementary Fig. 2.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

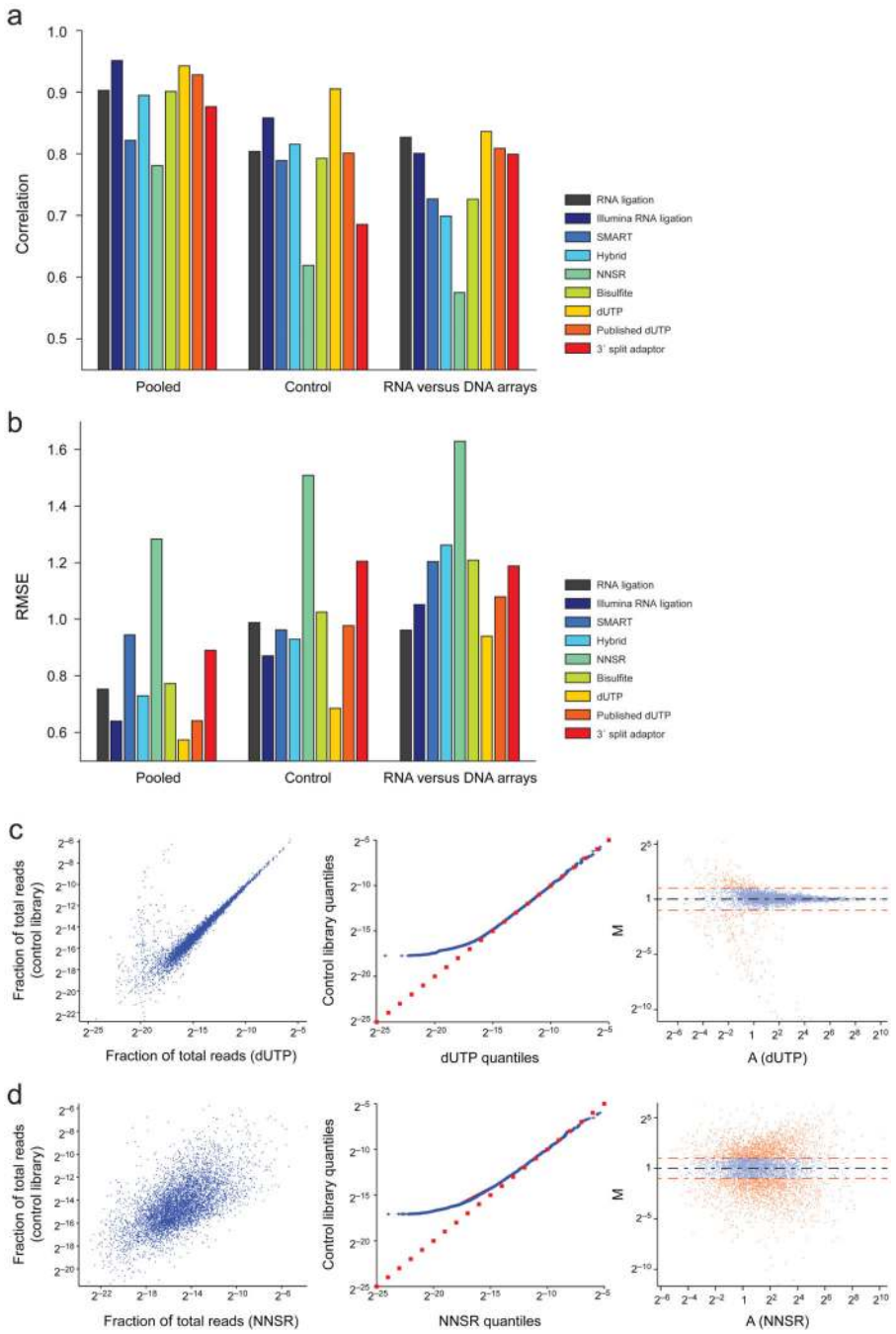

Figure 6. Digital expression profiling using strand-specific RNA-Seq

(**a, b**) Pearson correlation coefficient (a) and RMSE (b) for each library when compared to a pooled reference (left bars), the control library (middle bars) and Agilent microarrays (right bars). Full data for all library variations are presented in Supplementary Table 2. (**c, d**) Scatter (left panel), Q-Q (middle panel) and MA (right panel) 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)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of expression in the control library against the strand-specific library. A slope = 1 line is shown for reference (red crosses). The MA plot shows for each gene (dot) the difference in expression levels between the control and strand-specific libraries (Y axis) compared to their mean expression level (X axis). Red dashed lines – two fold difference in expression. Full data for all library variations are presented in Supplementary Fig. 4.