Vol 461|8 October 2009|doi:10.1038/nature08390

LETTERS

Direct RNA sequencing

Fatih Ozsolak[1], Adam R. Platt[1], Dan R. Jones[1], Jeffrey G. Reifenberger[1], Lauryn E. Sass[1], Peter McInerney[1], John F. Thompson[1], Jayson Bowers[1], Mirna Jarosz[1] & Patrice M. Milos[1]

Our understanding of human biology and disease is ultimately dependent on a complete understanding of the genome and its functions. The recent application of microarray and sequencing technologies to transcriptomics has changed the simplistic view of transcriptomes to a more complicated view of genome-wide transcription where a large fraction of transcripts emanates from unannotated parts of genomes[1–7], and underlined our limited knowledge of the dynamic state of transcription. Most of this broad body of knowledge was obtained indirectly because current transcriptome analysis methods typically require RNA to be converted to complementary DNA (cDNA) before measurements, even though the cDNA synthesis step introduces multiple biases and artefacts that interfere with both the proper characterization and quantification of transcripts[8–18]. Furthermore, cDNA synthesis is not particularly suitable for the analysis of short, degraded and/or small quantity RNA samples. Here we report direct single molecule RNA sequencing without prior conversion of RNA to cDNA. We applied this technology to sequence femtomole quantities of poly(A)$^+$ *Saccharomyces cerevisiae* RNA using a surface coated with poly(dT) oligonucleotides to capture the RNAs at their natural poly(A) tails and initiate sequencing by synthesis. We observed transcript 3′ end heterogeneity and polyadenylated small nucleolar RNAs. This study provides a path to high-throughput and low-cost direct RNA sequencing and achieving the ultimate goal of a comprehensive and bias-free understanding of transcriptomes.

cDNA-based transcriptome analysis approaches being used today exhibit several shortcomings that prevent us from understanding the real nature of transcriptomes and ultimately genome biology. Some of these limitations are: (1) the tendency of various reverse transcriptases (RT) to generate spurious second-strand cDNA due to their DNA-dependent DNA polymerase activities[9,10,18]; (2) the generation of artefactual cDNAs due to template switching[8,13,16,17] or contaminating DNA and primer-independent cDNA synthesis[11,12]; and (3) the error-prone[15,19] and inefficient nature of RTs yielding low quantities of cDNA. Furthermore, most RNA analysis technologies require the synthesis of not just the first strand cDNA but also a second strand cDNA that are both subjected to further ligation/amplification steps, introducing yet more biases. These limitations pose problems for the determination of RNA strandedness[14,20], the identification of chimaeric transcripts, quantification of RNA species, and the analysis of low quantity (<1 nanogram) or short RNA species, such as those obtained from formalin-fixed, paraffin-embedded tissue samples. Because almost all transcript analysis technologies in use today suffer from the limitations briefly summarized above, there is an ever-growing

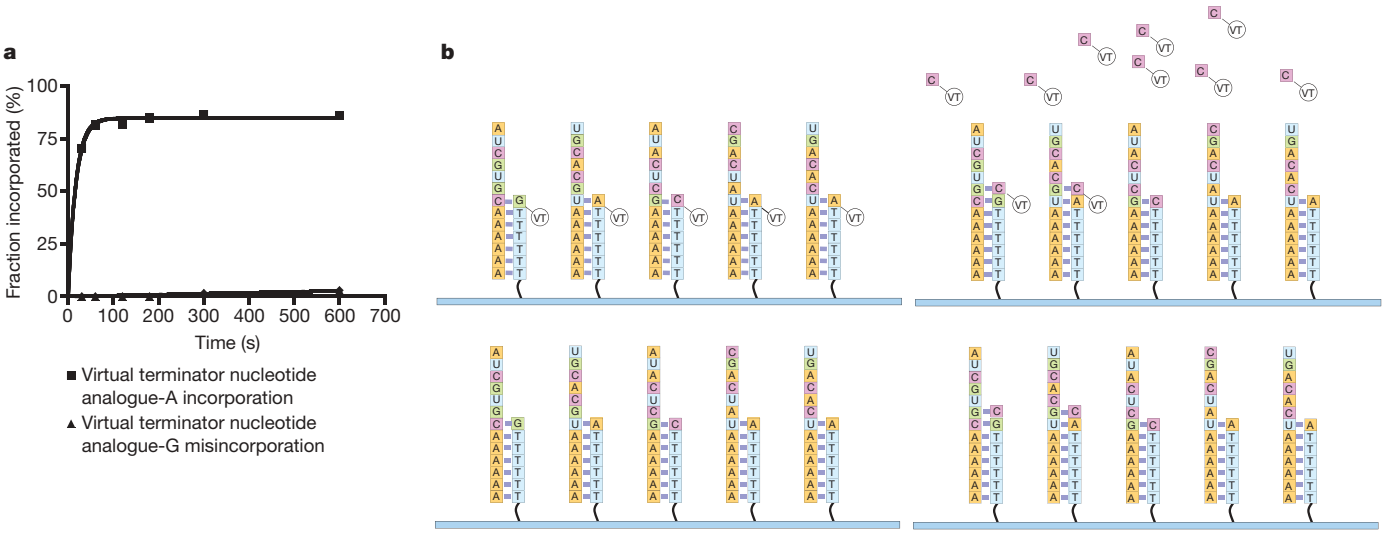

Figure 1 | DRS chemistry and sequencing steps. **a**, Under optimized conditions, polymerase exhibits fast correct nucleotide incorporation (VT-A) and slow misincorporation (VT-G) kinetics. **b**, DRS procedure. Top left: polyadenylated and 3′-blocked RNA is captured on surfaces coated with dT(50) oligonucleotide. A 'fill' step is performed with natural dTTP, and a 'lock' step with fluorescently labelled VT-A, -C and -G nucleotides. These steps correct for any misalignments that may be present in poly(A/T) duplexes, and ensure that the sequencing starts in the template rather than the poly(A) tail. Imaging is performed to locate the template positions. Bottom left: chemical cleavage of the dye-nucleotide linker is performed to prepare the templates for nucleotide incorporation. Top right: incubation with one VT nucleotide and polymerase is performed, followed by imaging to locate the templates that incorporated the nucleotide. Bottom right: chemical cleavage of the dye allows the surface and RNA templates to be ready for the next nucleotide addition cycle.

[1]Helicos BioSciences Corporation, One Kendall Square, Cambridge, Massachusetts 02139, USA.

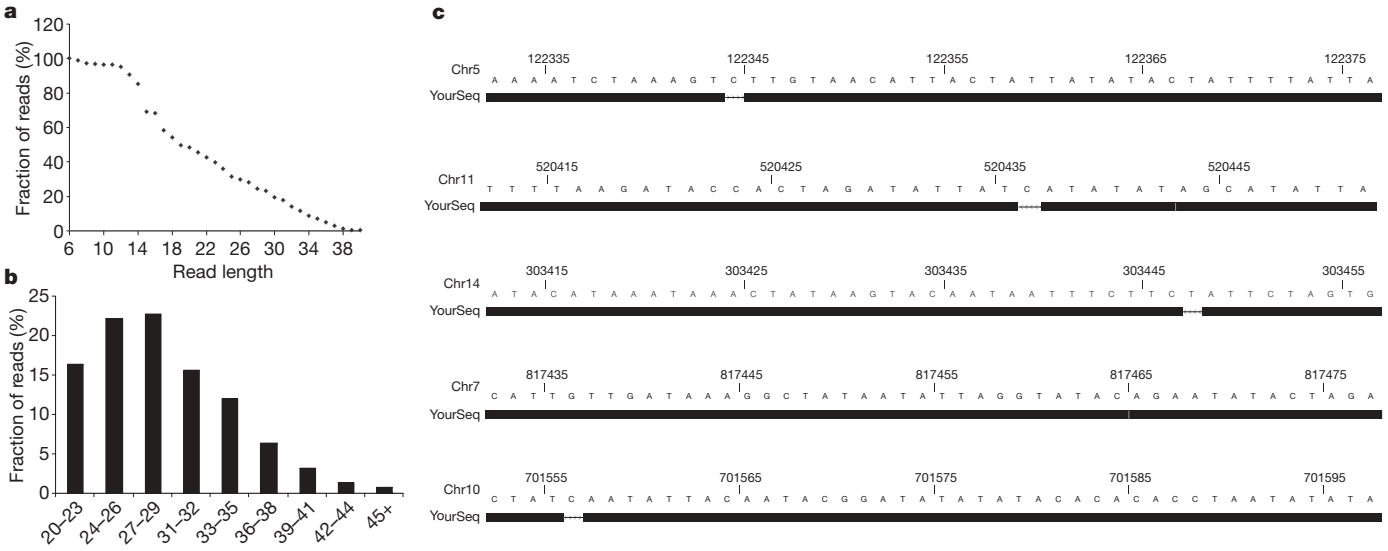

Figure 2 | DRS sequencing read-length statistics. **a**, Cumulative length distribution of reads obtained from oligoribonucleotides. *y* axis shows the fraction of reads at and above particular *x*-read lengths. **b**, Distribution of read lengths greater than 20 nucleotides aligned to *S. cerevisiae* genome. **c**, Several DRS reads are aligned with BLAT and visualized using the UCSC genome browser.

need for a method that would not be subject to the difficulties associated with RT behaviour, amplification, ligation and other cDNA synthesis/sample manipulation steps. A method allowing a comprehensive and bias-free view of transcriptomes using minute quantities of total RNA obtained from as few as one cell with no pre-treatment would stimulate great advances in the delineation of complex biological processes and be applicable across all biomedical research areas.

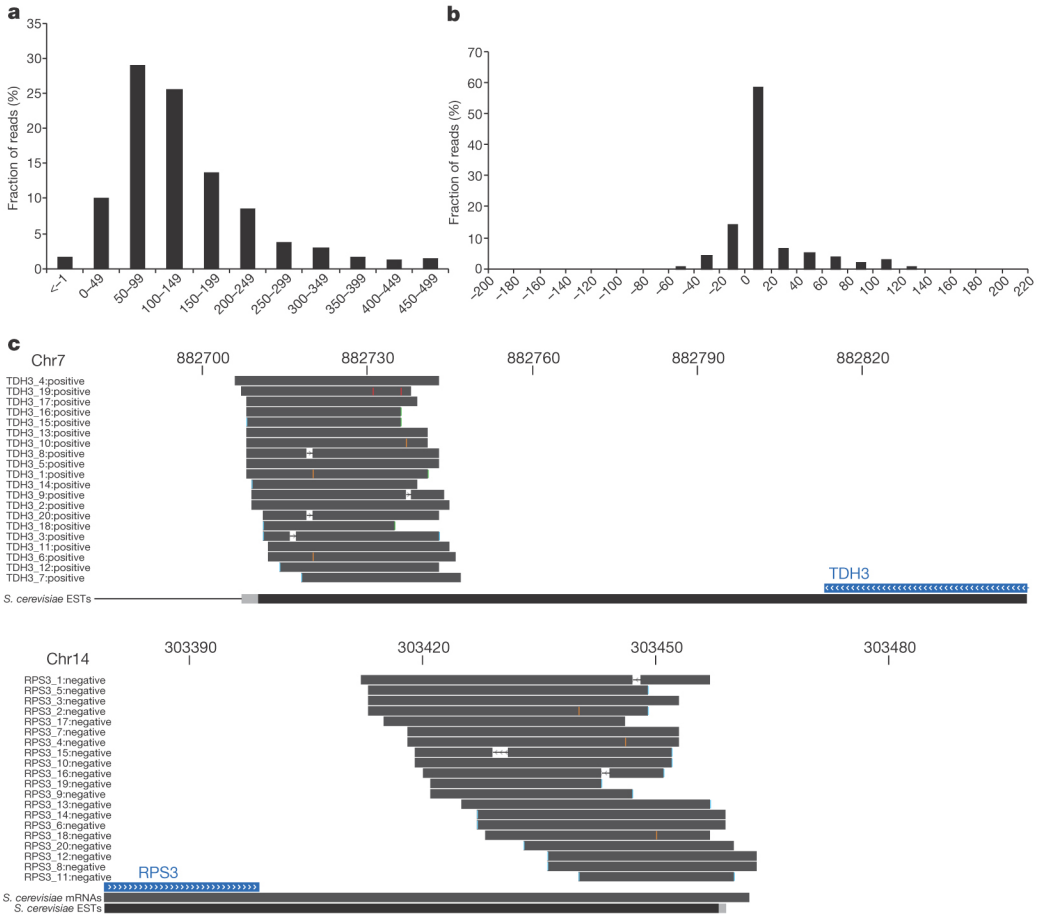

Figure 3 | DRS read distribution. **a**, **b**, Distance of aligned reads to *S. cerevisiae* coding sequence (**a**) and EST 3′ ends (**b**). *y* axis shows the fraction of reads at particular distance intervals indicated on the *x* axis. Most reads (91%) are 300 nucleotides immediately downstream of annotated gene 3′ ends. Reads aligning within the coding regions of transcripts are shown with negative distance in **a**. **b** represents the distances between the DRS reads and the closest EST clones. **c**, DRS reads aligning to *TDH3* and *RPS3* are exemplified. The alignment direction of each read (black bars) is indicated as 'positive' or 'negative'.

Here we report the successful development of direct RNA sequencing (DRS), allowing massively parallel sequencing of RNA molecules directly without prior synthesis of cDNA or the need for ligation/amplification steps. DRS represents an extension of single-molecule DNA sequencing technology (tSMS)[21,22] that relies on the stepwise synthesis and direct imaging of billions of single DNA strands on a planar surface. The sequencing-by-synthesis reaction is performed using a modified polymerase and proprietary fluorescent nucleotide analogues, called Virtual Terminator nucleotides (VT), that contain a fluorescent dye and chemically cleavable groups that allow step-wise sequencing. The first step for the development of DRS was the identification of an optimal polymerase, VT nucleotide analogues and buffer combination. Several DNA-dependent DNA polymerases have previously been shown to have reverse transcriptase activity[23–25]; we therefore tested DNA polymerases in addition to known RTs. After screening studies performed in solution, we identified conditions with satisfactory reaction kinetics (Fig. 1a) that could be attempted in a single-molecule sequencing system. The DRS procedure is summarized in Fig. 1b. Briefly, *Escherichia coli* poly(A) polymerase I (PAPI) is used to generate an A tail on 3′ ends of RNA molecules. The control of the A-tail length and the 3′ end blocking is performed by introducing 3′deoxyATP to the polyadenylation reaction shortly after the start of the tailing reaction, generating an A-tail of ~150 nucleotides. The blocking step is performed to prevent 'downward' nucleotide additions to the 3′ end of the template during the sequencing process. For RNA species containing poly(A) tails, such as mRNAs, poly(A) tailing is not required; only 3′ blocking is needed. Polyadenylated and 3′-blocked RNAs are hybridized to poly(dT)-coated surfaces. To begin sequencing at the unique region adjacent to the poly(A) tail, each RNA molecule is 'filled' in with dTTP and polymerase, and then 'locked' in position with VT-A, -C and -G addition, stopping subsequent nucleotide additions. After washing away the unincorporated dye-labelled nucleotides, images are taken, and then the fluorescent dye and inhibitor are cleaved off the incorporated nucleotide, rendering it suitable for additional rounds of incorporation. Each molecule is then provided the opportunity to extend (alternating C, T, A or G) followed by rinsing, imaging and cleavage. Repeating this cycle many times provides a set of images that are aligned and then used to generate sequence information for each individual RNA molecule with real-time image processing.

We first used chemically synthesized 40-mer RNA oligoribonucleotides as a model system to develop and optimize DRS chemistry. After sequencing on a prototype sequencer with 120 cycles of alternating VT-C, -T, -A or -G additions, we aligned the resultant sequence reads to the input oligonucleotide reference sequences and observed 48.5%

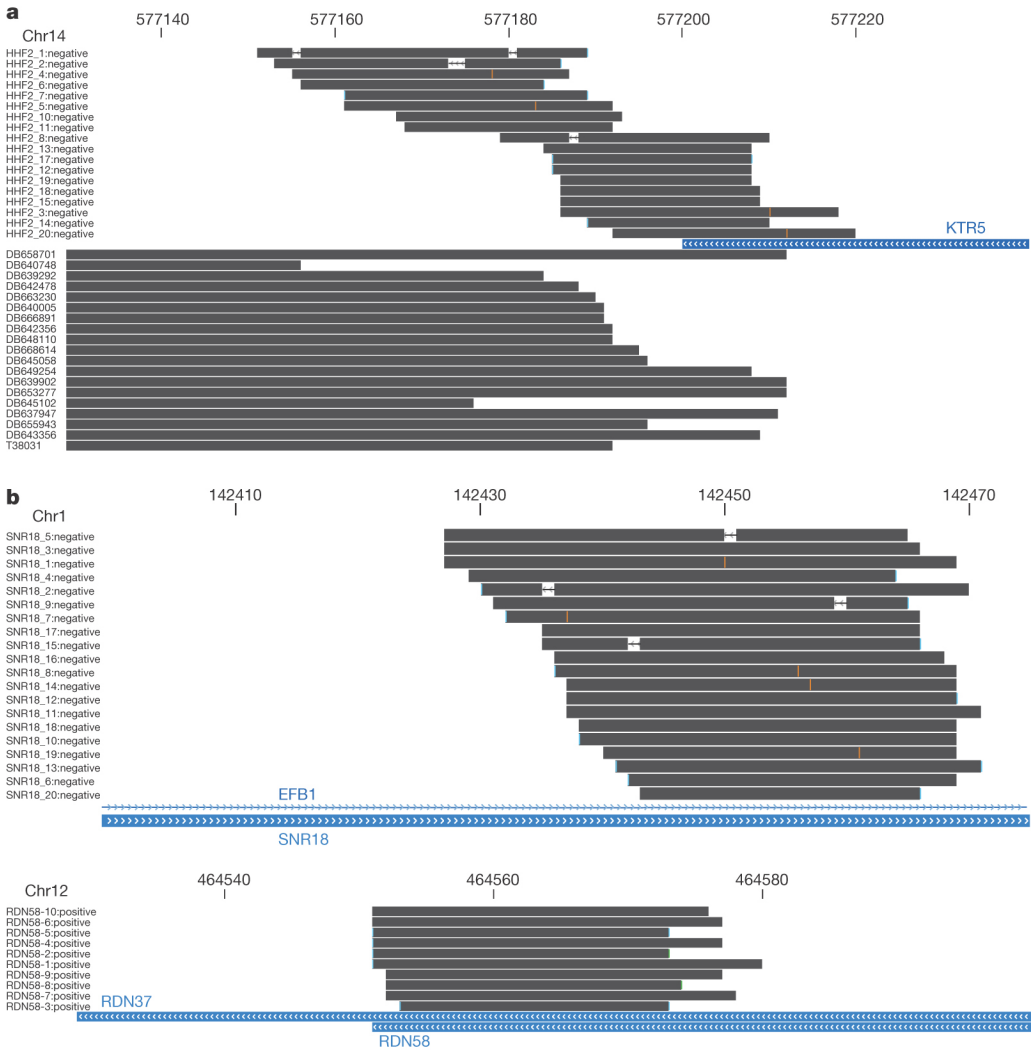

Figure 4 | *S. cerevisiae* poly(A)$^+$ RNA DRS suggests overlapping transcription units and polyadenylated snoRNA and rRNA species. a, DRS reads in this region are aligned in the reverse direction, suggesting their origination from the *HHF2* transcript (not shown) located ~200 nucleotide upstream relative to annotated *KTR5* 3′ end and transcribed in the forward direction. Three reads extend into the *KTR5* coding sequence, suggesting that the *HHF2* transcription can extend into *KTR5* coding sequence. **b**, Figure exemplifies randomly selected reads aligning to intronic *SNR18* (top panel) and 5.8S rRNA (bottom panel). Note that the reads are aligning within the 3′ ~40-nucleotide annotated regions of snoRNAs and rRNAs, suggesting their polyadenylation during or after their maturation.

of aligned reads to have a sequence length of at least 20 nucleotides, with the longest perfect match (no errors) being 38 nucleotides (Fig. 2a and Supplementary Fig. 5). In terms of total strand yield, DRS was efficient, providing on average 972 reads per 1,000 μm^2 flow cell surface area compared to ~1,100 for tSMS DNA sequencing in similar conditions. Because the sequencing process relies on incubating the templates with one base at a time, all errors are single base errors in the form of deletions (failure to detect incorporation), insertions (for example, failure to rinse VT analogues from the flow cell between each addition cycle) and substitutions. Total raw base error rate for DRS is currently approximately 4%, dominated by missing base errors (2–3%), whereas the insertion rate is 1–2% and the substitution error rate is 0.1–0.3%. Although further improvements in error rates are in progress, the read lengths and error rates achieved here are sufficient to allow the use of standard computational methods to align sequences to reference transcriptomes and genomes.

We then sequenced *Saccharomyces cerevisiae* poly(A)$^+$ RNA with DRS. Because this RNA sample already contains a natural, pre-existing poly(A) tail, no additional tailing was needed. Two femtomoles (~2 nanograms) of 3′ end-blocked yeast poly(A)$^+$ RNA was hybridized to dT(50) flow cells with no additional sample preparation procedures. One hundred and twenty sequencing cycles were performed on a prototype sequencing system over 3 days. RNA stability remained at high levels during the run, as demonstrated by the relatively constant number of nucleotides added per addition cycle to RNA templates (Supplementary Fig. 2). The sequence run generated 41,261 reads greater than 20 nucleotides, of which 19,501 reads (48.4%) aligned to the yeast genome using the BLAT algorithm[26]. The average aligned read length was 28.7 nucleotides, with the longest perfect match aligned read being 50 nucleotides (Fig. 2b, c). Of the aligned reads 91% were within 400 nucleotides downstream of annotated yeast gene 3′ open reading frame sequence ends (Fig. 3a). Such a wide distribution is expected, as yeast 3′ gene annotations mark mostly the coding sequence end point rather than the polyadenylation site. As expected, the alignment orientation of these sequences was in the opposite direction to the known gene transcription direction. This is a result of using unmodified, intact yeast poly(A)$^+$ RNA without any additional sample preparation steps, and therefore, the sequence read matches the direction opposite to that of the transcript. Because the 3′ ends of yeast protein-coding genes are not well annotated, we compared our findings to the yeast expressed sequence tags (EST) database as well. As exemplified in Fig. 3b and c, our data are supported by the EST data, with most of our reads being in close proximity to EST 3′ ends, aligned in the direction opposite to transcription. The reads that did not align to proximal 3′ ends of yeast genes or EST clones were caused by their localization beyond the 500 nucleotide downstream regions examined, at the 3′ ends of potentially transcriptionally active retrotransposons or at the 3′ ends of transcripts classified as dubious (Supplementary Fig. 3). Comparison of the DRS read localizations to the transcript 3′ ends identified previously[5] using high-throughput cDNA sequencing revealed a high concordance, with 81% of the DRS reads being within ±20 nucleotides of their 3′-end annotations (Supplementary Fig. 6). We observed transcripts extending into the coding sequence of neighbouring genes (Fig. 4a), some of which were supported by the available EST data and as described[5]. Interestingly, ~2% of the total reads aligning within the coding regions of transcripts (Fig. 3a) were from the ribosomal RNAs (rRNAs) and a portion of small nucleolar RNAs (snoRNAs). Mature forms of these RNAs are produced from longer precursor RNAs through cleavage steps. Our observation that DRS reads map to the 3′ end of the mature snoRNAs and rRNAs indicates that at least a fraction of snoRNAs[27,28] and rRNAs[29] can be polyadenylated post-transcriptionally, possibly during their 3′-end processing and/or RNA quality control steps[28] (Fig. 4b and Supplementary Fig. 7). We independently validated the 3′ polyadenylation site heterogeneity and the existence of polyadenylated snoRNAs by amplifying 3′ polyadenylation sites with polymerase chain reaction (PCR) in a manner preserving the variability in the 3′ ends, followed by sequencing of the PCR products with tSMS DNA sequencing to identify the 3′ polyadenylation sites (Supplementary Fig. 11). Our data add further support to the suggestion that many yeast genes possess a heterogeneous set of 3′ ends for genes[5].

The simplicity of the DRS sample preparation steps presented here, the requirement for only femtomole quantities of RNA and the potential of DRS to eliminate biases introduced by cDNA synthesis, end repair, ligation and amplification procedures will be useful for applications requiring minute quantities of RNA and/or short RNA species that are challenging for analysis with existing cDNA-based methodologies. This ability, combined with further improvements in DRS sample preparation, single molecule sequencing surface capture, throughput and computational tools, will ultimately allow us to understand and quantify the 'true' nature of transcriptomes in a high-throughput, low-cost and bias-free manner.

METHODS SUMMARY

RNA oligoribonucleotide templates were obtained from IDT. *S. cerevisiae* poly(A)$^+$ RNA was obtained from Clontech. Polyadenylation of oligoribonucleotides was performed by using a poly(A) tailing kit (Ambion). 3′deoxyATP (cordycepin triphosphate, Jena Biosciences) was introduced 10 min after the initiation of the polyadenylation reaction for 3′-end blocking and tail length limitation. DRS reads obtained are listed in Supplementary Table 4.

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

Received 18 May; accepted 5 August 2009.
Published online 23 September 2009.

1. Denoeud, F. *et al.* Annotating genomes with massive-scale RNA sequencing. *Genome Biol.* **9,** R175 (2008).
2. Kapranov, P., Willingham, A. T. & Gingeras, T. R. Genome-wide transcription and the implications for genomic organization. *Nature Rev. Genet.* **8,** 413–423 (2007).
3. 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).
4. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628 (2008).
5. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320,** 1344–1349 (2008).
6. 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).
7. Wilhelm, B. T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453,** 1239–1243 (2008).
8. Cocquet, J., Chong, A., Zhang, G. & Veitia, R. A. Reverse transcriptase template switching and false alternative transcripts. *Genomics* **88,** 127–131 (2006).
9. Gubler, U. Second-strand cDNA synthesis: classical method. *Methods Enzymol.* **152,** 325–329 (1987).
10. Gubler, U. Second-strand cDNA synthesis: mRNA fragments as primers. *Methods Enzymol.* **152,** 330–335 (1987).
11. Haddad, F. *et al.* Regulation of antisense RNA expression during cardiac MHC gene switching in response to pressure overload. *Am. J. Physiol. Heart Circ. Physiol.* **290,** H2351–H2361 (2006).
12. Haddad, F., Qin, A. X., Giger, J. M., Guo, H. & Baldwin, K. M. Potential pitfalls in the accuracy of analysis of natural sense-antisense RNA pairs by reverse transcription-PCR. *BMC Biotechnol.* **7,** 21 (2007).
13. Mader, R. M. *et al.* Reverse transcriptase template switching during reverse transcriptase-polymerase chain reaction: artificial generation of deletions in ribonucleotide reductase mRNA. *J. Lab. Clin. Med.* **137,** 422–428 (2001).
14. Perocchi, F., Xu, Z., Clauder-Munster, S. & Steinmetz, L. M. Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. *Nucleic Acids Res.* **35,** e128 (2007).
15. Roberts, J. D. *et al.* Fidelity of two retroviral reverse transcriptases during DNA-dependent DNA synthesis *in vitro*. *Mol. Cell. Biol.* **9,** 469–476 (1989).
16. Roy, S. W. & Irimia, M. When good transcripts go bad: artifactual RT-PCR 'splicing' and genome analysis. *Bioessays* **30,** 601–605 (2008).
17. Roy, S. W. & Irimia, M. Intron mis-splicing: no alternative? *Genome Biol.* **9,** 208 (2008).
18. Spiegelman, S. *et al.* DNA-directed DNA polymerase activity in oncogenic RNA viruses. *Nature* **227,** 1029–1031 (1970).
19. Varadaraj, K. & Skinner, D. M. Denaturants or cosolvents improve the specificity of PCR amplification of a G + C-rich DNA using genetically engineered DNA polymerases. *Gene* **140,** 1–5 (1994).
20. Wu, J. Q. *et al.* Systematic analysis of transcribed loci in ENCODE regions using RACE sequencing reveals extensive transcription in the human genome. *Genome Biol.* **9,** R3 (2008).

21. Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S. R. Sequence information can be obtained from single DNA molecules. *Proc. Natl Acad. Sci. USA* **100,** 3960–3964 (2003).
22. Harris, T. D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320,** 106–109 (2008).
23. Karkas, J. D., Stavrianopoulos, J. G. & Chargaff, E. Action of DNA polymerase I of *Escherichia coli* with DNA-RNA hybrids as templates. *Proc. Natl Acad. Sci. USA* **69,** 398–402 (1972).
24. Rüttimann, C., Cotoras, M., Zaldívar, J. & Vicuna, R. DNA polymerases from the extremely thermophilic bacterium *Thermus thermophilus* HB-8. *Eur. J. Biochem.* **149,** 41–46 (1985).
25. Stenesh, J., Roe, B. A. & Snyder, T. L. Studies of the deoxyribonucleic acid from mesophilic and thermophilic bacteria. *Biochim. Biophys. Acta* **161,** 442–454 (1968).
26. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
27. Kim, M. *et al.* Distinct pathways for snoRNA and mRNA termination. *Mol. Cell* **24,** 723–734 (2006).
28. Grzechnik, P. & Kufel, J. Polyadenylation linked to transcription termination directs the processing of snoRNA precursors in yeast. *Mol. Cell* **32,** 247–258 (2008).
29. Slomovic, S., Laufer, D., Geiger, D. & Schuster, G. Polyadenylation of ribosomal RNA in human cells. *Nucleic Acids Res.* **34,** 2966–2975 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. Kerouac, P. Kapranov, L. Kung, C. Hart and D. Lipson for technical assistance and discussions.

Author Contributions F.O. conceived the project, designed the experimental plan, coordinated the studies and analysed the data. F.O., J.B., L.E.S. and P.M. performed the enzyme kinetics assays. F.O., D.R.J., A.R.P. and J.G.R. did the sequencing experiments. J.F.T. and M.J. provided experimental reviews. F.O. and P.M.M. wrote the manuscript, which was reviewed by all authors.

Author Information Sequencing data sets described in this study have been deposited at the National Center for Biotechnology Information (NCBI) Short Read Archive (SRA), accession no SRA 009023. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to F.O. (fatihozsolak@gmail.com) or P.M.M. (pmilos@helicosbio.com).

METHODS

Polymerase kinetics assay. Incorporations of VT nucleotide analogues in an RNA-template-directed manner by various enzyme and buffer combinations were screened by designing four 50-mer oligoribonucleotides (5′-UUCUUUUGCCUCUUUCGNCAGGGCAGAGGAUGGAUGCAAGGAUAAGUGGA-3′); the 5′ 25-nucleotide sequence of the oligoribonucleotides being complementary to a 25-mer 5′-rhodamine-labelled DNA oligo (5′-TCCACTTATCCTTGCATCCATCCTCTGCCCTG-3′), and the 26th nucleotide (denoted as N above) on the oligoribonucleotides being each of the four nucleotides. After hybridizing the 25-mer DNA oligonucleotide to RNA templates at 65 °C for 5 min in nuclease-free water, followed by incubation on ice for 2 min, the selected enzyme/buffer/VT combinations were added to the RNA-DNA hybrid mix. The reaction was stopped at different time points by an EDTA-quench and kinetics were measured by observing the lengthening of the 5′-rhodamine-labelled DNA oligonucleotide using capillary electrophoresis (ABI 3730 DNA Analyzer, Applied Biosystems). This assay allowed us to observe the kinetics of VT nucleotide incorporation into the 3′ end of the DNA primer in an RNA-template-directed manner. All oligonucleotides and oligoribonucleotides were ordered from IDT.

Sample preparation for DRS. RNA oligoribonucleotide templates were ordered from IDT. The sequences were Oligo 1, 5′-AGAGUCCCAUCCUCACCAUCAUCACACUGGAAGACUGCAG-3′; Oligo 2, 5′-CUGGUGCAGCACUCUCGACGGCACCUAUCUGCCAUCGUAG-3′; Oligo 3, 5′-CGAUCGUCACUAUCUGCAUCAGUAGCUCUAGCAUACUGAG-3′; Oligo 4, 5′-UCUUUCGUCAGGGCAGAGGAUGGAUGCAAGGAUAAGUGGA-3′. Polyadenylation was performed by using a poly(A) tailing kit (Ambion). 3′deoxyATP (cordycepin triphosphate, Jena Biosciences) was introduced 10 min after the initiation of the polyadenylation reaction for 3′-end blocking and tail length limitation. Reaction products were cleaned by phenol-chloroform extraction and ethanol precipitation. Samples were analysed with microcapillary electrophoresis (Agilent Technologies) (Supplementary Fig. 1). Poly(A)$^+$ *S. cerevisiae* RNA strain DBY746 (*his3D1 leu2-3 leu2-112 ura3-52 trp1-289*), grown under standard conditions (yeast peptone dextrose, 30 °C) was obtained from Clontech (product number 636312). *S. cerevisiae* poly(A)$^+$ RNA (2 ng) was used for 3′-end blocking reaction with poly(A) polymerase and 3′deoxyATP.

Surfaces and template capture. Fifty-nucleotide poly(dT) primers (obtained from IDT) were covalently coupled to sequencing surfaces prepared on glass coverslips in one-channel or five-channel formats. Slides are available from Helicos BioSciences. Poly(A) tail containing RNA molecules were hybridized to the surface at 10–30 pM, requiring 0.5–1.5 fmol polyadenylated RNA per sequencing reaction. The surface was rinsed and the locations of the hybridized templates were determined by imaging after the 'fill and lock' step.

cDNA preparation, PCR amplification and DNA sequencing with the Helicos Genetic Analysis System. First-strand cDNA was prepared using a SuperScript III first-strand cDNA synthesis kit (Invitrogen) from 500 ng poly(A)$^+$ *S. cerevisiae* RNA according to manufacturer's instructions using 50 pmol dT/U-25-V primer (Supplementary Table 1). After cDNA synthesis, RNA was removed by RNase H (Invitrogen) and RNA If (New England Biolabs) digestion for 30 min at 37 °C followed by cleaning with a Nucleotide Removal Kit (Qiagen, 28304). The cDNA is then PCR-amplified with the dT/U-25-V primer and gene/snoRNA specific primers (Supplementary Table 1) using Taq polymerase (New England Biolabs, M0273) under the following thermal cycling conditions: 94 °C for 3 min, 30 cycles of 94 °C for 30 s, 48 °C for 30 s, 72 °C for 30 s, followed by a final 72 °C 10 min incubation step. The excess primers were removed by running the PCR products on 1% agarose gel. Because 3′ ends of genes are amplified, and multiple and variable size PCR products are expected, we extracted regions from the gels representing 50–500 base pairs (bp) size distribution (visible PCR products had 100–300 bp sizes) and isolated the DNA with a QIAEX II gel extraction kit (Qiagen, 20021). We chose this approach over commercial column/bead-based cleaning methods because: (1) PCR primers need to be removed as much as possible, otherwise they will be A-tailed by the terminal transferase (described below) and sequenced; and (2) as we expected multiple and variably sized PCR products, we did not want to use commercial systems that may have varying efficiencies of removing small fragments (generally $<$100 bp) and preserving larger DNA fragments. The PCR products were then treated with the USER enzyme (New England Biolabs, M5505) to eliminate/reduce the 5′-T/U tails left on the PCR products after PCR (Supplementary Fig. 11A). This step was performed to prevent potential competition of the 5′-T/U tail with the Poly(dT) primers on sequencing surfaces used for template capture and sequencing initiation. The USER reaction was cleaned with the Qiagen Nucleotide Removal Kit. Ten nanograms from each PCR product was combined and A-tailed with terminal transferase (New England Biolabs). Briefly, pooled PCR products were heat denatured at 95 °C for 5 min in the presence of the supplied 1× reaction buffer and 2.5 mM $CoCl_2$, followed by rapid snap-cooling on ice. Terminal transferase (40 U; New England Biolabs) and 900 pmol dATP were then added to the denatured DNA in 50 µl final reaction volume, incubated at 37 °C for 1 h, followed by the inactivation of the enzyme at 70 °C for 10 min. The blocking step was performed by adding 300 pmol ddTTP and 4 U of terminal transferase to the heat-denatured A-tailed reaction above, incubating at 37 °C for 1 h, followed by the inactivation of the enzyme at 70 °C for 20 min. After the tailing and blocking steps, the final DNA was then loaded directly into two channels of the 50 channel Helicos Genetic Analysis System without additional cleaning steps. cDNA sequencing data alignment to the yeast genome (October 2003 assembly) was performed using IndexDP[30].

30. Lipson, D. *et al.* Quantification of the yeast transcriptome by single-molecule sequencing. *Nature Biotechnol.* **27**, 652–658 (2009).