Highly parallel direct RNA sequencing on an array of nanopores

Daniel R Garalde[1], Elizabeth A Snell[1], Daniel Jachimowicz[1], Botond Sipos[1], Joseph H Lloyd[1], Mark Bruce[1], Nadia Pantic[1], Tigist Admassu[1], Phillip James[1], Anthony Warland[1], Michael Jordan[1], Jonah Ciccone[1], Sabrina Serra[1], Jemma Keenan[1], Samuel Martin[1], Luke McNeill[1], E Jayne Wallace[1], Lakmal Jayasinghe[1], Chris Wright[1], Javier Blasco[1], Stephen Young[1], Denise Brocklebank[1], Sissel Juul[2], James Clarke[1], Andrew J Heron[1] & Daniel J Turner[1]

Sequencing the RNA in a biological sample can unlock a wealth of information, including the identity of bacteria and viruses, the nuances of alternative splicing or the transcriptional state of organisms. However, current methods have limitations due to short read lengths and reverse transcription or amplification biases. Here we demonstrate nanopore direct RNA-seq, a highly parallel, real-time, single-molecule method that circumvents reverse transcription or amplification steps. This method yields full-length, strand-specific RNA sequences and enables the direct detection of nucleotide analogs in RNA.

A cell's transcriptome contains rich information, including the structure of genes (such as splice variants and fusion genes), different expression levels of transcripts, and antisense transcription[1].

A method to best capture this information is one that is accurate, strand specific, quantitative across a wide dynamic range, does not need prior knowledge of sequence, is capable of revealing the presence and identity of modified bases, and can detect antisense transcripts without a concern that these are artifacts of library preparation[2]. Ideally, the method would also generate continuous sequence reads that span any splice junctions. Current sequencing-based transcriptomic analyses (RNA-seq), based on the high-throughput sequencing of complementary DNA (cDNA), have enabled us to build a more accurate picture of the active transcriptional patterns within organisms[1]. The most commonly used RNA-seq strategy involves either polydeoxythymidine (poly(dT)) priming or RNA fragmentation and random hexamer priming, followed by cDNA synthesis. These cDNA strands are amplified by PCR, which can introduce bias[3] such as reduced complexity of the resulting cDNA library, distortion of relative cDNA abundances and dropout of some RNA species. An amplification-free library prep would sidestep these issues[4]. Additionally, during PCR amplification, any modifications on the RNA are lost.

Two current methods do not require PCR amplification for RNA-seq library preparation, FRT-seq[5] and the DRS technique on the Helicos platform[6], but both of these approaches generate short sequence reads, which can make it difficult to correctly identify alternative splicing in eukaryotes[7,8]. This problem should be addressable by combining a long-read sequencing technology with a library preparation method that maintains the integrity of the RNA being analyzed[9].

All previous RNA-seq methods detect the products of a synthesis reaction rather than directly detecting the RNA molecule. Thus, sequences generated with these methods are subject to the processivity and error-rate limitations of reverse transcription and either cannot detect base modifications or cannot distinguish homopolymers[10].

Oxford Nanopore Technologies' nanopore-based platform detects single molecules of DNA, proteins and small molecules as they traverse through a nanopore, without the need for an enzymatic synthesis reaction. The platform consists of single nanopores embedded in an array of thousands of individual synthetic polymer membranes on a single flowcell. An electric potential drives DNA toward and into the nanopores[11]. When a single DNA molecule is captured in a pore and ratcheted through the pore at a consistent rate by an engineered motor protein, it creates perturbations of the nanopore current[12] which a recurrent neural network (RNN) converts into base sequences.

Here we assess an amplification-free method for sequencing RNA and detecting RNA modifications using the nanopore platform. To our knowledge, this is the first parallel, truly direct, RNA-seq method.

RESULTS

Direct RNA sequencing of yeast transcripts

To assess the performance of the direct RNA-seq method, we sequenced a direct RNA library (**Fig. 1a**) from yeast poly(A)$^+$ RNA on a MinION MkIb with R9.4 flowcells. Using the MinKNOW instrument software, we recorded the nanopore current as each strand of RNA translocated through a nanopore

[1]Oxford Nanopore Technologies Ltd., Oxford, UK. [2]Oxford Nanopore Technologies Inc., New York, New York, USA. Correspondence should be addressed to D.J.T. (dan.turner@nanoporetech.com).

RECEIVED 23 JULY 2017; ACCEPTED 21 NOVEMBER 2017; PUBLISHED ONLINE 15 JANUARY 2018; DOI:10.1038/NMETH.4577

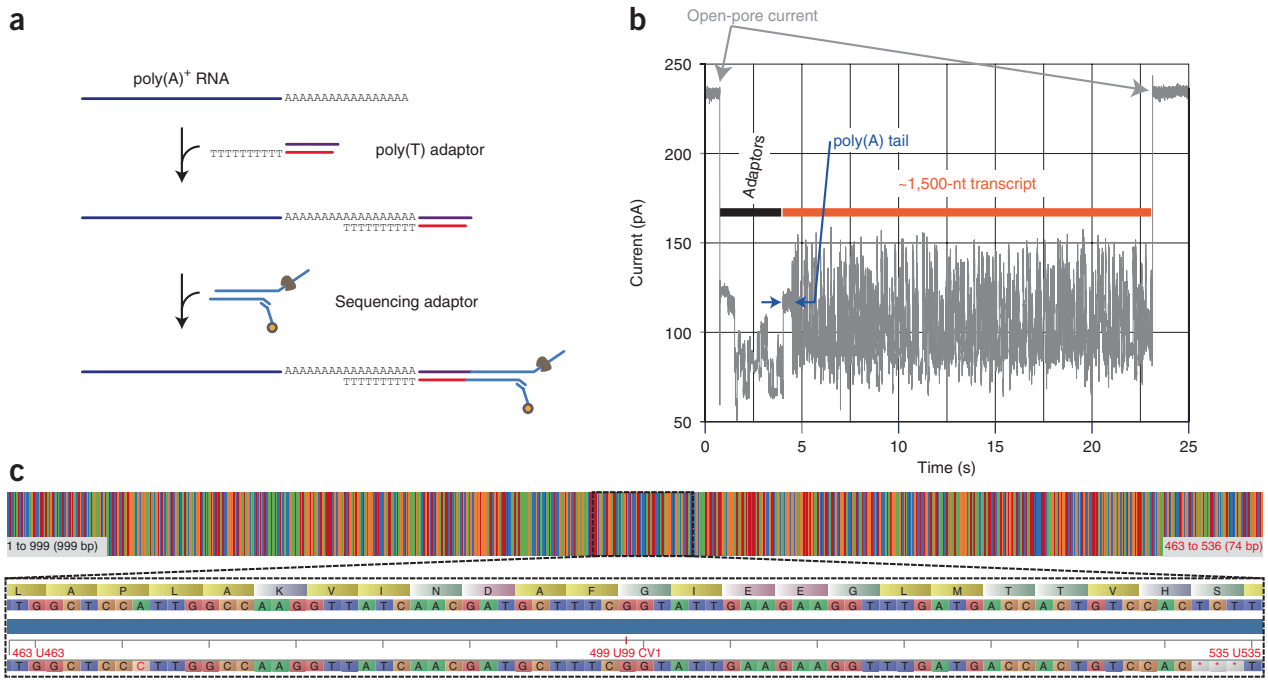

Figure 1 | Direct RNA-seq. (**a**) Library preparation method for direct RNA-seq. (**b**) Representative raw data 'squiggle' resulting from translocation of a single transcript through a pore in the MinION array. (**c**) Alignment of a typical *Saccharomyces cerevisiae* S228C read to the reference transcriptome.

(**Fig. 1b**). As an adaptor-ligated yeast RNA enters the pore, the adaptor oligo is detected first, followed by the poly(A) tail, then the body of the transcript. The nanopore current returns to a high open-pore level as the transcript exits the pore on the opposite side of the membrane.

We used Albacore 1.2.1 (Oxford Nanopore Technologies Ltd.) to call the bases, and we aligned the resulting reads to the *Saccharomyces cerevisiae* transcriptome. We also sequenced the same yeast mRNA sample on a MinION cDNA run and an Illumina 100-nucleotide paired-end run. The number of reads that mapped to the yeast transcriptome was 2,777,523 (79.29% of reads) for the direct RNA data set; 5,735,508 (90.36%) for the cDNA data set; and 572,206,890 (79.32%) for the Illumina data set. The read-length distributions for the direct RNA and nanopore cDNA data sets were similar (**Supplementary Fig. 1**). The modal accuracy of basecalled direct RNA reads is currently >90% (**Supplementary Fig. 2**).

We calculated read-count correlations between the three data sets as described, and we obtained good agreement. The direct RNA and cDNA nanopore data sets gave the highest correlation (Spearman's rho = 0.89), and both nanopore data sets gave similar correlation values to those of the Illumina data set (Spearman's rho = 0.81 for direct RNA and 0.79 for cDNA; **Fig. 2a**). Five technical replicates of different direct RNA yeast libraries correlated very well (Spearman's rho = 0.94–0.96; n = 6,713 transcripts), showing that the library prep and sequencing is reproducible (**Supplementary Fig. 3**).

We calculated the log fold change in coverage between the direct RNA and Illumina data sets. The number of direct RNA reads mapped to the yeast genome using GMAP[13] was 2,045,748 (63.43%)l; while the number of Illumina RNA-seq reads mapped by GSNAP was 708,592,030 (98.22%, **Fig. 2b,c**). Direct RNA gene coverage corresponds well with the Illumina results (Spearman's rho = 0.73).

Two of the yeast transcripts identified by direct RNA-seq mapped to isozymes of GAPDH, forms of the same enzyme that are encoded by similar genes at different loci. The sequences of the two genes are 95.8% identical, differing at 42 positions dispersed throughout the genes (data not shown). Even though the single-read accuracy of the direct RNA data is currently below the ~96% identity of the two genes, analysis of the reads mapping to each isozyme imply correct placement—first, multimapping (and hence randomly placed) reads are in the minority; and second, because each nanopore read covers the majority of the 42 divergent bases, a few incorrectly called bases cannot shift the mapping to the other gene (**Fig. 2d**).

Direct measurement of RNA with low bias

To assess the bias introduced by increasing numbers of PCR cycles, we prepared Nanopore cDNA libraries using the ERCC RNA spike-in mix with 5–40 PCR cycles (**Supplementary Fig. 4**). The results indicate that a major component of the bias comes from a decrease in the proportion of full-length reads, rather than a decrease in correlation between expected and observed read counts.

Next, we generated ~60,000 direct RNA reads from the same ERCC panel and calculated abundance as described in the Online Methods (**Fig. 3a**). The correlation between read counts and expected values (Spearman's rho = 0.93, $P = 1.9^{-40}$; the correlation coefficient and the corresponding two-sided P value were calculated using the stats.spearmanr function from the scipy Python package.) indicates low bias, regardless of the fragment length. The protocol gives good coverage of entire transcripts; a histogram of alignment coverage values shows the majority of values being close to 1, which indicates that alignments tend to cover full transcripts (**Fig. 3b,c**). Additionally, because the RNA strand is sequenced directly, the reads necessarily map back to the reference in a strand-specific manner (**Fig. 3c**).

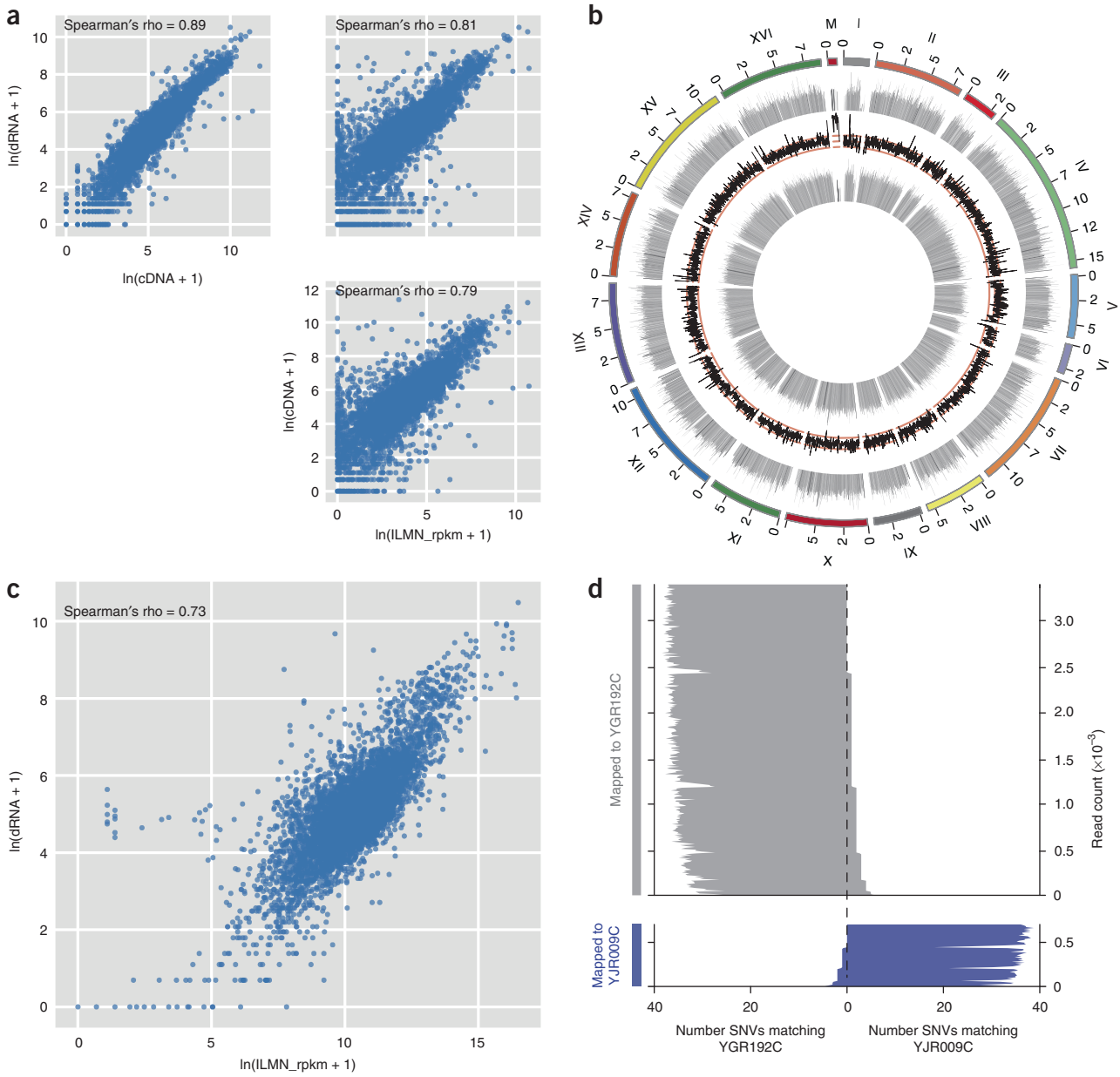

Figure 2 | Analysis of the *Saccharomyces cerevisiae* S228C transcriptome by direct RNA-seq. (**a**) Correlation between read counts after transcriptome mapping for direct RNA, cDNA and Illumina data sets calculated from n = 6,531 transcripts (each transcript appeared in at least one of the data sets). (**b**) Circos plot of reads aligned to the reference genome. The outer track shows the reference genome. Immediately inside this track are log gene coverage of direct RNA reads. The innermost track shows log gene coverage of cDNA reads. Between the two tracks is the $\log_2$ fold change of relative gene coverage between the data sets. The red lines indicate −2 and +2. (**c**) Correlation between gene coverage after genome mapping for direct RNA and Illumina data sets calculated from n = 6,692 genes. (**d**) Mapping results of two yeast GAPDH isozymes. Reads mapping to YGR192C are shown in gray, and reads mapping to YJR009C are shown in blue. For each mapped read, the plot shows the number of SNVs matching YGR192C and the number of SNVs matching YJR009C.

We investigated biases in read count correlated to transcript length or GC content for both direct RNA and Illumina yeast data sets (**Supplementary Fig. 5**). Transcript length has less of an effect on read count in direct RNA than Illumina (Pearson's $r = 0.13$, $P = 5.4 \times 10^{-29}$ and Pearson's $r = 0.3$, $P = 7 \times 10^{-141}$, respectively; **Supplementary Fig. 5a**,**b**). GC content appears to have a negligible influence (Pearson $r = 0.013$, $P = 0.29$) on read count in the direct RNA data set, substantially less than for the Illumina data set (Pearson $r = 0.19$, $P = 1.6 \times 10^{-58}$; **Supplementary Fig. 5c**,**d**), and the mean quality of aligned portions of the direct RNA reads does not appear to be strongly influenced by GC content (Pearson's $r = -0.082$; **Supplementary Fig. 5e**; $P = 0$, n = 2,777,523 alignments. The correlation coefficient and the corresponding two-sided P value were calculated using the stats.spearmanr function from the scipy Python package.).

Splice variation

We evaluated the unambiguous detection of splice variants using Lexogen's Spike-in RNA Variant Control Mixes (SIRVs). When quantifying isoform levels in the E2 SIRV data set using the

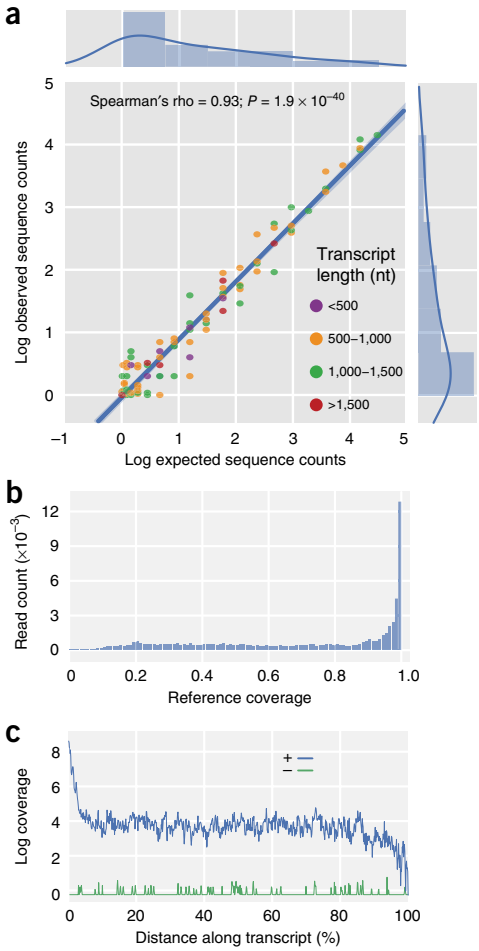

Figure 3 | Analysis of quantitative and length biases using the ERCC spike-in panel. (**a**) Correlation between read counts and expected abundances for ERCC RNAs (Spearman's rho = 0.93, $P = 1.9 \times 10^{-40}$, $n = 92$ ERCC transcripts). The correlation coefficient and the corresponding two-sided P value were calculated using the stats.spearmanr function from the scipy Python package. (**b**) The global reference coverage histogram of covered reference portions in the ERCC RNAs. (**c**) The global fragment-coverage plot for ERCC RNAs. Fragment coverage is the number of alignments starting at the respective position.

transcriptome-alignment-based strategy, we found strong correlation with the known mix concentrations (Spearman's rho = 0.8, $P = 3 \times 10^{-16}$; **Fig. 4a**). On the gene level the rank correlation was perfect (Spearman's rho = 1.0, $P = 0$; **Fig. 4b**), which suggests that discriminating between similar isoforms by transcriptome mapping is the factor limiting the correlation on transcript level.

With the spliced-alignment-based quantification strategy, we found a quantitatively lower, but still strong, correlation with the known E2 isoform concentrations (Spearman's rho = 0.62; **Supplementary Fig. 6**). The lower correlation suggests that, while this strategy is viable, the additional complexity of spliced mapping and the downstream quantification approach might introduce additional biases. These can hopefully be alleviated by optimizing the analysis tools further for use with long reads.

When we evaluated the coverage of individual exons in the E0 SIRV data set, we did not find any trends, such as missing first or last exons (**Supplementary Fig. 7**). Indeed, all exons had considerable coverage, with the exceptions of most exons in SIRV502, the two first exons of SIRV505 and the last exon of SIRV704. Hence, we conclude that there are reads supporting the existence of all but a few transcripts in the correct annotation. Guided assembly of the E0 SIRV data set achieved a transcript-level sensitivity of 100% (all 69 transcripts recovered) and specificity of 95.8% (three false-positive transcripts reported). No exons were missed or novel exons reported. These results reinforce our conclusion that our data set contains reads supporting all correct transcripts. The unguided assembly, however, recovered only 14 correct transcripts (sensitivity, 20.3%; precision, 43.8%). It is possible that further optimization of StringTie[14] will improve its performance with direct RNA reads.

Detection of modified bases

Modifications on RNA cause a characteristic current blockade within the nanopore that can be measured by direct RNA-seq. To determine the effect of two common RNA modifications, N^6-methyladenosine (m^6A) and 5-methylcytosine (5-mC), on the current blockade, we sequenced the FLuc transcript in which the relevant nucleoside was either modified in every position, or was unmodified (Trilink Biotechnologies Inc.). We constructed an HMM to independently describe each position in the sequence. After training, the resulting models can be considered consensus squiggles—they encode the average current observed for each position along the sequence. We compared the mean current levels for unmodified (red line) versus m^6A modification (blue line, **Fig. 5a**) or 5-mC modification (blue line, **Fig. 5b**). The current level is perturbed locally near nucleoside modifications, but is otherwise similar.

DISCUSSION

The direct RNA-seq method here described has many potential advantages over other RNA-seq strategies; namely, (i) it is amplification free so does not suffer from PCR bias[3] or RT bias; (ii) it is compatible with very long reads, which are particularly useful, for example, in the study of splice variants; (iii) it measures the RNA directly so it can detect nucleotide analogs; and (iv) it is strand specific.

There are several areas where we are currently working to improve the direct RNA-seq method. First, the current basecalling model appears to be slightly overfitted to yeast (data not shown); and although the effect is not substantial, yeast data sets will be called with higher accuracy than others. Training the basecaller on a wider range of data sets will be required to remove this effect, and this will increase overall accuracy. Although slight read-length effects can be seen in the yeast transcriptome data, these are lower than for the Illumina data set. In the absence of an unambiguously correct answer for the expression levels of this particular yeast sample, we cannot rule out the possibility that expression level in this sample is genuinely correlated with transcript length, though this seems unlikely.

Degraded RNA can hinder proper detection of splice variants, because the sample RNA is no longer full length. This issue can be addressed with a method for isolating intact transcripts; for instance, by targeting the eukaryotic 5′ cap. Including such a step in the library preparation should further improve the read-length distribution of direct RNA reads.

Many of the software tools we used for these analyses were not optimized for nanopore direct RNA data. Without this optimization, the fraction of reads mapping to the yeast

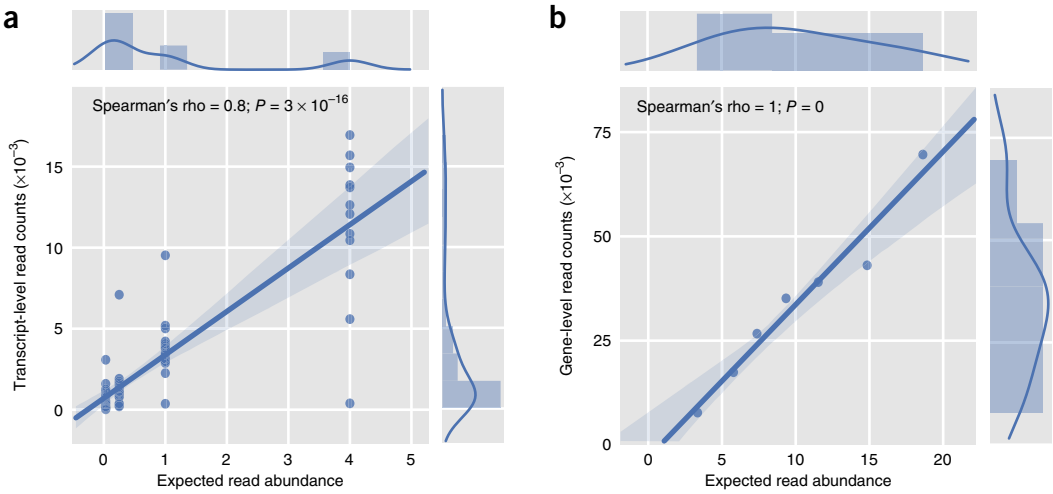

Figure 4 | Detection of splice variants using the SIRV E2 spike-in panel. Reads were aligned using the transcriptome-alignment strategy and correlations calculated at (**a**) transcript level (Spearman's rho = 0.8, $P = 3 \times 10^{-16}$, $n = 69$ transcripts) and (**b**) gene level (Spearman's rho = 1, $P = 0$, $n = 7$ genes). The correlation coefficients and the corresponding two-sided P values were calculated using the stats.spearmanr function from the scipy Python package.

transcriptome (~0.63) is lower than that of both the Illumina data set (~0.79) and the nanopore cDNA data set (~0.90). This lower mapping rate may be explained by the direct RNA reads having higher error than either of the cDNA data sets. However, in many applications the greater length of the direct RNA reads versus the Illumina reads may compensate for the lower accuracy. Results of our spliced mapping analysis suggest that the fraction of mapped direct RNA reads will increase with optimization of the alignment tools.

Although the data presented here on the detection of modified bases are preliminary, they reveal a clear and systematic difference between groups of molecules that differ only in the presence or absence of modified nucleotides in the sequence contexts analyzed. Others have also made this observation[15], and such analyses indicate that it may be possible to detect base modifications at a single-molecule level with single-nucleotide resolution on a transcriptome-wide scale. Although the computational cost will grow with the number of base modifications included in the basecaller, many cases of practical relevance target limited choices of base analogs at specific loci in a reference sequence—analyses which can be performed by less computationally expensive algorithms.

In some cases, synthesis of a cDNA strand opposite the RNA template before sequencing improves throughput, possibly by reducing intramolecular secondary structure of the RNA. Analogous to the 1D[2] method of DNA sequencing, it is possible to create an RNA–cDNA hybrid in such a way that the cDNA strand is sequenced immediately after its parent RNA strand. The data from both strands could then be combined into a single, higher accuracy read. Such an approach could also be used for *de novo* identification of modified bases, since the synthetic strand provides a built-in reference sequence.

In this paper, the RNA adaptor is ligated onto the 3′ poly(A) tail of RNA. The approach lends itself readily to the sequencing of eukaryotic mRNAs, but it is necessary to add a tail to other types of RNA. We have found enzymatic addition of a 3′ poly(A) tail to be efficient. In addition, the direct RNA adaptors are modular, so the poly(T) splint can be replaced with a user-provided, sequence-specific splint for targeting specific non-poly(A) 3′ sequences such as ribosomal RNAs.

For the work presented here, we operated the motor protein at a speed around 85 nucleotides per s. Refinements to our DNA-sequencing process have allowed us to increase the DNA motor speed approximately ten-fold, with scope to increase this further. Screening a wider range of motor protein mutants will allow us to find enzymes with the best combination of steady movement, good processivity and high processing speed, which will increase throughput and data quality.

We are currently using the same *Escherichia coli* CsgG-derived nanopore for both direct RNA and DNA sequencing[16], and this allows both DNA and RNA strands to be sequenced together on the same flowcell, but we continue to engineer this pore as well as

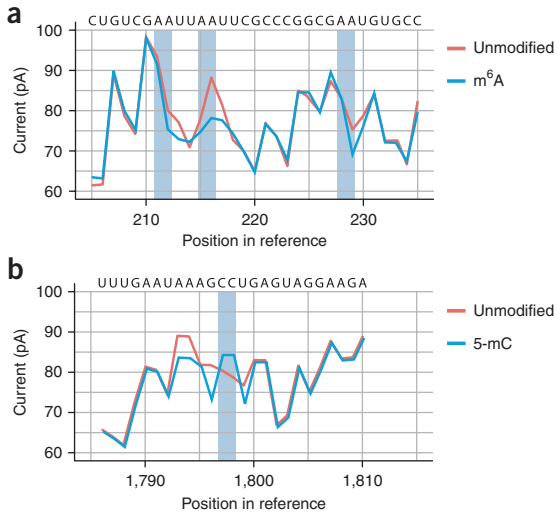

Figure 5 | Detection of modified bases in synthetic RNA strands. Using a Hidden Markov Model, we generated average current levels for each k-mer in the strands with or without the relevant modification, and we aligned these average levels to (**a**) a region of a strand containing m^{6}A, U, G and C compared to an unmodified strand and (**b**) a region of a strand containing A, U, G and 5- C compared to an unmodified strand.

search for a better pore. Pore mutations can improve the signal-to-noise ratio and RNA-capture efficiency, which would allow higher accuracy basecalls from less input RNA.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

AUTHOR CONTRIBUTIONS

D.R.G., A.J.H., J. Clarke and D.J.T. conceived the experiments. D.R.G. led the project. D.R.G., E.A.S., D.J., A.J.H., J.H.L., P.J., A.W., M.J., J.K., S.M. and L.M. designed and performed the experiments. J.H.L. tested, engineered and developed the motor protein. J.H.L., S.M., L.M., D.R.G., E.A.S., A.J.H., M.B., D.J., A.W. and E.J.W. designed or assessed motor protein mutations and the sequencing adaptor. D.J.T., D.R.G. and E.A.S. developed the library preparation. E.A.S. and J.K. created custom RNA templates. B.S. wrote custom analysis tools and performed analysis of all sequence data sets. N.P., T.A. and M.B. expressed and purified proteins. M.J., J. Ciccone and S.S. designed and prepared plasmids. M.J., E.J.W., L.J., S.Y., D.R.G., E.A.S., D.J., A.J.H., M.B., J.H.L. and D.B. assessed sequencing performance of buffers, voltages and pores. C.W. wrote squiggle-consensus algorithms. J.B., C.W., D.B., J.H.L., M.B. and S.Y. trained RNA basecallers or analyzed modified base data. D.J.T., B.S., D.R.G., S.J. and C.W. wrote the manuscript. A.J.H., S.Y. and P.J. contributed to the figures or to editing of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
2. 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).
3. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods* **6**, 291–295 (2009).
4. Lipson, D. *et al.* Quantification of the yeast transcriptome by single-molecule sequencing. *Nat. Biotechnol.* **27**, 652–658 (2009).
5. Mamanova, L. *et al.* FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat. Methods* **7**, 130–132 (2010).
6. Ozsolak, F. *et al.* Direct RNA sequencing. *Nature* **461**, 814–818 (2009).
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. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10**, 1177–1184 (2013).
9. Thomas, S., Underwood, J.G., Tseng, E. & Holloway, A.K. Long-read sequencing of chicken transcripts and identification of new transcript isoforms. *PLoS One* **9**, e94650 (2014).
10. Vilfan, I.D. *et al.* Analysis of RNA base modification and structural rearrangement by single-molecule real-time detection of reverse transcription. *J. Nanobiotechnology* **11**, 8 (2013).
11. Clamer, M., Höfler, L., Mikhailova, E., Viero, G. & Bayley, H. Detection of 3′-end RNA uridylation with a protein nanopore. *ACS Nano* **8**, 1364–1374 (2014).
12. Smith, A.M., Abu-Shumays, R., Akeson, M. & Bernick, D.L. Capture, unfolding, and detection of individual tRNA molecules using a nanopore device. *Front. Bioeng. Biotechnol.* **3**, 91 (2015).
13. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
14. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
15. Byrne, A. *et al.* Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* **8**, 16027 (2017).
16. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nat. Biotechnol.* **34**, 518–524 (2016).

ONLINE METHODS

Screening motor proteins. We expressed and tested a large number of helicases and polymerases from an internal library of candidate enzymes, and from published literature, for suitable strand movement properties—in particular, processive, single-base, unidirectional movement on RNA with a speed of 80 to 1,000 bases per second. Consistent with published literature, we found that many DNA translocases lack efficient binding activity or processivity on RNA. However, our enzyme screen identified a candidate translocase that exhibited movement on RNA in the 3′ to 5′ direction.

In addition to using fluorescence assays to measure displacement proficiency of candidate RNA translocases, we analyzed translocase movement properties on the *in vitro* RNA transcripts eGFP, FLuc, and β-galactosidase (Trilink Biotech) on a MinION using R9.4 flowcells. Briefly, we used in-house scripts to create a consensus of current levels generated by different reads of the same sequence. The consensus data provided a rough measure of the consistency of the signal generated each time an RNA strand was ratcheted through the pore by the translocase.

We used these limited data sets to train Hidden Markov Models (HMMs) to predict the sequence that generated a given nanopore signal. Once we could use an HMM to map a nanopore signal to an RNA sequence, we used more diverse training sets, such as the transcriptome from yeast strain S228C, to improve the model. We used an HMM to align the nanopore signal to the reference nucleotide sequence resulting in the input data for training a Recurrent Neural Network (RNN) basecaller. The RNN training process can overcome deficiencies in the HMM labeling of the training data to yield an improved basecaller. In summary, we found that a properly trained basecalling model allowed us to generate high-quality basecalls using the single-molecule RNA data generated using this translocase. We engineered this translocase further to increase stability, binding, movement and speed. We also engineered a closed-complex version to topologically lock the enzyme around the polynucleotide substrate, enabling both essentially unlimited processivity and the ability to prepare stable enzyme-preloaded adapters.

Adaptor design. We designed a sequencing adaptor to attach to the 3′ end of the RNA template strands. This adaptor has a 3′ terminus that is low in secondary structure, allowing the 3′ end of the adaptor to be captured by a nanopore. The motor protein is prebound to the adaptor, which contains a short section of nonpolynucleotide linker designed to prevent the protein from processing through the RNA strand until captured on the pore. The adaptor also contains a hydrophobic portion that encourages the library strands to associate with the membrane in which the nanopores are embedded.

Library preparation and sequencing. We prepared direct RNA libraries using the protocol depicted in **Figure 1**. Briefly, 100–500 ng of poly(A)$^+$ RNA was ligated to a poly(T) adaptor using T4 DNA ligase. This ligase was found empirically to give a higher yield of ligated products than RNA ligases 1 and 2 (approximately 80% compared to 40% for RNA ligases) and to reach its maximum yield in a shorter amount of time. Following adaptor ligation, the products were purified by adding a 1.8-fold excess of Agencourt RNAClean XP beads and following the Agencourt purification protocol. Sequencing adaptors preloaded with motor protein (200 mM annealed, preloaded adaptor, 200 mM NaCl, 50 mM Tris–HCl pH 7.5, 5% (*w/v*) glycerol, 0.1% (*w/v*) β-OTG, 0.1 mM EDTA) were then ligated onto the overhang of the previous adaptor using T4 DNA ligase (**Fig. 1a**). Excess buffer was removed using Agencourt RNAClean XP beads with a modified purification protocol described in the Direct RNA Sequencing kit documentation[17]. The RNA library was eluted from the RNAClean beads in 21 µl of elution buffer (50 mM Tris–HCl pH 8.0, 20 mM NaCl, 200 mM oligonucleotide with hydrophobic portion). 1 µl of the RNA library was quantified using a Qubit fluorometer using the manufacturer's RNA assay. Immediately before sequencing, the remaining 20 µl of RNA library was mixed with 17.5 µl of nuclease-free water and 37.5 µl of undiluted 2× running buffer (940 mM KCl, 50 mM HEPES pH 7.0, 20 mM $MgCl_2$, 22 mM ATP), making 75 µl of the final RNA library. The final RNA libraries were added to FLO-MIN106 flowcells and run on an MkIb MinION.

Library preparation and 100 nucleotide, paired-end Illumina sequencing of the yeast RNA sample was performed by the sequencing service at the Wellcome Trust Centre for Human Genetics, Oxford, UK.

Data analysis. 1. Basecalling: sequencing runs were performed using MinKNOW version 1.5.5 or 1.5.15 software (Oxford Nanopore Technologies Ltd.) by executing the EXPERIMENT_RNA_Baseline_Sequencing.py script. MinKNOW is the instrument control software that runs on the host computer to which the MinION is connected. MinKNOW carries out several core tasks: data acquisition; real-time analysis and feedback of experimental progression; data streaming while providing device control, including selecting the run parameters and ensuring that the platform chemistry is performing correctly to run the samples. The data output from MinKNOW consists of a single file per sequence read, in an HDF5 format[18] called FAST5.

FAST5 files were basecalled using Albacore 1.2.1 (Oxford Nanopore Technologies Ltd.), a proprietary recurrent neural network basecaller, with the following parameters: read_fast5_basecaller.py -i <input_dir> -t 10 -c r94_70bps_rna_linear.cfg -s <output_dir> -o fastq,fast5

2. Yeast transcriptome analysis: for the yeast analyses, we have used Ensembl release 89 genome, annotation and cDNA collection. For mapping the nanopore and Illumina reads to the transcriptome we used bwa mem (0.7.15-r1140) with parameters -x ont2d and bwa aln/sampe, respectively. Reads mapping to the different transcripts were counted by the bam_count_reads.py script from the wub package (https://github.com/nanoporetech/wub), with a minimum required mapping quality parameter of 5. The RPKM values for the Illumina data set were calculated by performing length normalization using the length_normalise_counts.py script from the wub package. The correlation of counts from different runs and experiments was calculated using the correlate_counts.py script from the wub package.

To assess the biases in the ssRNA, cDNA and Illumina data, we first generated read count files from the reads aligned to the transcriptome using the bam_count_reads.py script from the wub package with the -z argument specified and with a minimum required mapping quality of 5. We then used the bias_explorer.py script from wub to calculate correlation of transcript counts with the length and GC content of originating transcripts.

Mapping of the nanopore reads to the yeast genome was performed by GMAP (version 2017-05-08)[13] with parameters –n 1–cross-species -f samse. The Illumina reads were mapped to the genome using gnsap (version 2017-05-08)[13] with parameters -N 1 -n 1. The gene coverage was generated from the genomic alignments using the annotation and bedtools coverage (v2.25.0)[19] and visualized using circos (v0.69-5)[20]. When calculating the $\log_2$-fold ratios for the middle track of the Circos plot, we added a pseudocount of 10^{-7} to avoid division by zero when a gene is not covered. The pairwise alignment used for the isozyme figure was generated using Clustal W (version 2.1)[21].

3. ERCC panel quantitative analysis: the ERCC panel (ThermoFisher Scientific) is a set of 92 polyadenylated RNAs, ranging from 250 to 2,000 nucleotides in length, which are present in the mixture at defined concentrations. Reads were mapped to the ERCC transcriptome reference using bwa mem (version 0.7.15-r1142-dirty)[22] with parameters -Y -M -L 300 -x ont2d. Reads mapping to the different transcripts were counted by the bam_count_reads.py script from the wub package. The correlation of read counts and known abundance was plotted using in-house scripts (https://github.com/nanoporetech/dRNA-paper-scripts). The fragment coverage plots and reference coverage histograms were produced from the ERCC transcriptome alignments using bam_frag_coverage.py from the wub package. More information on the study design and analysis software can be found in the **Life Sciences Reporting Summary**.

4. SIRV splice panel analyses: in order to assess the quantification of isoforms in the E2 SIRV ssRNA data set, we used two strategies: one based on transcriptome alignment and another based on spliced alignment to the genome.

For the transcriptome-based approach, we first mapped the E2 reads to the SIRV transcriptome using bwa mem –x ont2d (version 0.7.15-r1140)[22] then counted the reads mapping to each transcript using the bam_count_reads.py script from the wub package. We used the plot_sirv_correlations.py script to assess the correlation of counts with the known E2 mix concentrations.

For the spliced-alignment-based approach, we aligned the E2 ssRNA reads to the SIRV 'genome' using GMAP (version 2017-05-08)[13] with the parameters –cross-species -n 1 -z sense_force. We then used StringTie (version 1.3.3)[14] along with the correct annotation (specified through the -G parameter) provided by Lexogen in order to quantify the transcripts (with the -e parameter specified). We used the scripts gtf_to_counts.py and plot_sirv_correlations.py to assess the correlation of FPKMs reported by StringTie with the known E2 mix concentrations.

In order to quantify the detection of splicing events through reconstructing transcripts, we used a data set generated from the SIRV E0 sample, which contains all transcripts in equimolar amounts. The spliced mapping to the SIRV 'genome' was performed as in the case of the E2 data set.

To evaluate the coverage of individual exons by aligned reads, we used StringTie in the quantification mode (-e parameter) along with the correct annotation, then we used the gtf_plot_exon_cov.py script to plot the log of average base coverage of individual exons in all transcripts. In order to assess the recovery of transcripts, we performed a StringTie transcript assembly guided by the correct annotation and also a completely *de novo* assembly (with additional parameters -f 0.05 -c 1.0). The gffcompare utility (version v0.9.8, https://github.com/gpertea/gffcompare) was used to evaluate the quality of annotations obtained by the assemblies by comparing it to the correct annotation.

5. Detection of methylated adenosine and cytosine: to show that direct RNA-seq can detect the presence of methylation, we trained Hidden Markov Models[23] from distinct samples; as a reference, we used synthetic RNA strands that contained only canonical nucleotides, and we used this to compare two modified strands (Trilink Biotechnologies, Inc., California)—one containing, m^6A, U, G and C; and the other containing A, U, G and 5-mC. For simplicity of exposition we choose to model all positions of the reference sequence independently; that is, the HMM comprises a state space containing as many states as there are bases in the reference sequence. Such modeling allows us to relax the assumption that only five bases contribute to the observed ionic current. To bootstrap the training, we do however use the emission parameterization of a 5-mer basecalling model—we index the parameters μ_p and σ_p on the five bases surrounding the reference position p. Having trained the HMM models, the final emission parameter sets represent a consensus 'squiggle' across all reads in the two data sets. To compare the consensuses we performed least-squares regression of μ_p and $\mu\_p(meth)$ at reference positions, which are not expected to be affected by the methylation under the 5-mer assumption.

Code availability. With the exception of scripts used to detect modified nucleotides, all custom scripts used to perform bioinformatics analyses are available from https://github.com/nanoporetech/wub and https://github.com/nanoporetech/dRNA-paper-scripts.

Life Sciences Reporting Summary. Further information on experimental design is available in the **Life Sciences Reporting Summary**.

Data availability. All data sets presented in this paper have been deposited in the Sequence Read Archive under BioProject accession number is PRJNA408327 and BioSample accession numbers SAMN07688322, SAMN07684568, SAMN07684569 and SAMN07684570.

17. Oxford Nanopore Technologies Ltd. Direct RNA sequencing https://community.nanoporetech.com/protocols/direct-rna-sequencing/v/drs_9026_v1_revj_15dec201 (2016).
18. The HDF Group. Hierarchical data format, version 5, 1997–2017. http://www.hdfgroup.org/HDF5/.
19. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
20. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
21. Larkin, M.A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
22. Li, H. & Durbin, R. Burrows–Wheeler Alignment Tool http://bio-bwa.sourceforge.net/bwa.shtml (2012).
23. Fariselli, P., Martelli, P.L. & Casadio, R. A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins. *BMC Bioinformatics* **6**, S12 (2005).

doi:10.1038/nmeth.4577

nature research

Corresponding author(s): Daniel J Turner

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

Sample sizes were not considered here. The manuscript reports a new way to sequence RNA, and in a sense, the sample size could be considered to be the number of reads used in the various analyses. However, these were governed by the performance of the method rather than by statistical considerations.

2. Data exclusions

Describe any data exclusions.

No data were excluded.

3. Replication

Describe whether the experimental findings were reliably reproduced.

In our analysis of reproducibility, all results were reported and the findings were reliably reproduced.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Randomization was not relevant as samples in this study are commercially available.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Not applicable, the manuscript concerns a new method of RNA sequencing, and the data we generate by this method is very distinctive.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Nature Methods: doi:10.1038/nmeth.4577

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	The third-party software used to analyse the data in this study is bwa mem ver. 0.7.15-r1142-dirty, GMAP ver. 2017-05-08, bedtools ver. V2.25.0., gnsap ver. 2017-05-08, circos ver. V0.69-5, StringTie ver. 1.3.3, Clustal W ver. 2.1. and affcompare ver. v0.9.8. Custom tools can be found in the wub package (https://github.com/nanoporetech/wub version string: a80af13) and scripts for this paper (https://github.com/nanoporetech/dRNA-paper-scripts version string c276036).

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	All materials used were from standard commercial sources

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	None were used

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	The only eukaryotic cells used in this work were a commercial strain of yeast
b. Describe the method of cell line authentication used.	We authenticated the cells by genome sequencing
c. Report whether the cell lines were tested for mycoplasma contamination.	The cells were not tested for mycoplasma contamination
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No commonly misidentified cell lines were used

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals were used

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	No human research participants were used

Nature Methods: doi:10.1038/nmeth.4577