https://doi.org/10.1038/s41592-022-01633-w

Advances in nanopore direct RNA sequencing

Miten Jain, Robin Abu-Shumays, Hugh E. Olsen and Mark Akeson

Nanopore direct RNA sequencing (DRS) reads continuous native RNA strands. Early adopters have used this technology to document nucleotide modifications and 3′ polyadenosine tails on RNA strands without added chemistry steps. Individual strands ranging in length from 70 to 26,000 nucleotides have been sequenced. In our opinion, broader acceptance of nanopore DRS by molecular biologists and cell biologists will be accelerated by higher basecall accuracy and lower RNA input requirements.

Nanopore direct RNA sequencing (DRS)[1] shares general features with nanopore DNA sequencing (Fig. 1). Briefly, a helicase motor regulates native RNA movement through a bespoke protein nanopore. As the RNA is driven through the pore by an applied voltage, monovalent ionic current varies depending on the identity of nucleotides in the pore. This ionic current signature is then converted into an RNA nucleotide sequence for individual strands by a neural network trained on a variety of RNA samples. In principle, nanopore DRS can provide a comprehensive picture of individual RNA strands as they exist in cells. Each individual read would include all exons, untranslated regions (UTRs), nucleotide modifications and end modifications (for example, 5′ capping and 3′ polyadenylation). Nanopore DRS has been used to analyze cellular mRNA and noncoding RNA, and numerous RNA viruses[2,3] including SARS-CoV-2[4,5].

State-of-the-art nanopore DRS

In 2019, a collaboration between six laboratories, including our group, used nanopore DRS to acquire ten million aligned poly(A) RNA reads from the model human cell line GM12878 (ref. [6]). These data were based on 30 MinION flow cell runs using 500 ng poly(A) RNA per flow cell. Throughput at the time was fairly low at 50,000–831,000 reads per flow cell. Aligned read lengths ranged from 85 to >21,000 nucleotides with a median basecall accuracy of 86%.

Subsequent experiments revealed improved metrics. In our hands, a MinION flow cell now typically generates 1–2 million aligned reads, and the documented read length range has been extended from a lower limit of 74 nucleotides for three *Escherichia coli* transfer RNAs (tRNAs)[7] to an upper limit of 26 kb for a coronavirus RNA genome[2]. Staff at Oxford Nanopore Technologies (ONT) recently described an updated nanopore DRS protocol in which 50 ng input poly(A) RNA can deliver robust throughput. This has been corroborated by Nadine Holmes (University of Nottingham), who observed 565,000 RNA reads compared to 823,000 reads for 50 ng and 500 ng brain poly(A) RNA, respectively (personal communication).

Basecall accuracy has also improved. Since 2019, two publications have reported median accuracy of ~91% for *Brassica napus*[8] and ~88–90% for *E. coli*[9]. These studies used updated versions of the Guppy software and the direct RNA sequencing kit. To verify these results, we reanalyzed human GM12878 poly(A) RNA data using Guppy v. 6.3.2 (default quality score cutoff = 7). For our group's data in the Workman study (~2.6 million reads)[6] and for the Mulroney et al. study (~3.8 million reads)[10], we found median accuracies of 90.6% and 89.8%, respectively, in agreement with the published results.

However, in our view, two technical issues need to be resolved to promote broader acceptance of nanopore DRS:

(1) Basecall accuracy should be >99%. Reliable documentation of short exons and exon boundaries for individual RNA strands will require accuracies well above 90%. It is reasonable to expect that further improvements in nanopore DRS accuracy are attainable because ONT DNA basecall accuracy is presently at 99%[11].

(2) Long RNA transcripts are underrepresented in nanopore DRS data. For example, the human *Xist* gene encodes a polyadenylated long noncoding RNA with isoforms up to 17 kb in length. We documented about 300 *Xist* mRNA reads in the GM12878 study[6]. As expected, most of these aligned to the paternal *Xist* allele; however, none of the reads corresponded to the full-length 17 kb isoform (in this case defined as extending from a 3′ poly(A) tail to within 25 nucleotides of the end of the 5′ exon).

Analysis of mitochondrial mRNA (mt-mRNA) transcripts helps explain this read length shortfall[6]. In human cells, mt-mRNAs are single exon and abundant (~10% of total mRNA reads), with lengths ranging from 349 to 2,379 nucleotides, and thus can serve as a useful internal control for nanopore DRS. When the ratio of full-length transcripts to total transcripts for ten mt-mRNA was plotted against the corresponding gene length, a linear anticorrelation was observed ranging from 0.92 for *MT-ND3* (349 nucleotides) to 0.55 for *MT-ND4/ND4L* (1,673 nucleotides) (a similar anticorrelation was recently observed for *Caenorhabditis elegans* mt-mRNA[12]). RNA degradation in MinION flow cells was a minor cause of shortened reads (~5% over 36 h); however, read truncations caused by enzyme stalls or spurious voltage spikes were common (>19% for 1.5-kb-long *MT-CO1*). Importantly, investigators in Nottingham, UK, showed that these truncated reads could be reconstructed in silico using continuous ionic current data[6]. As this phenomenon scales with length, it follows that 17 kb *Xist* transcripts would be significantly underrepresented by nanopore DRS. We expect that forthcoming ONT MinKNOW software updates will address this issue.

Software advances

In our opinion, Nanopolish[13] stands out among many academic computational tools for nanopore data analysis. Briefly, Nanopolish begins with a sequence of bases for individual RNA or DNA strands generated by ONT software. These sequences are aligned to reference sequences typically using minimap2 (ref. [14]). Nanopolish then works backwards, converting 5-nt-long sequences (pentamers) from the alignments into discrete ionic current segments ('events') from the original nanopore ionic current trace. The mean, standard deviation and dwell time for

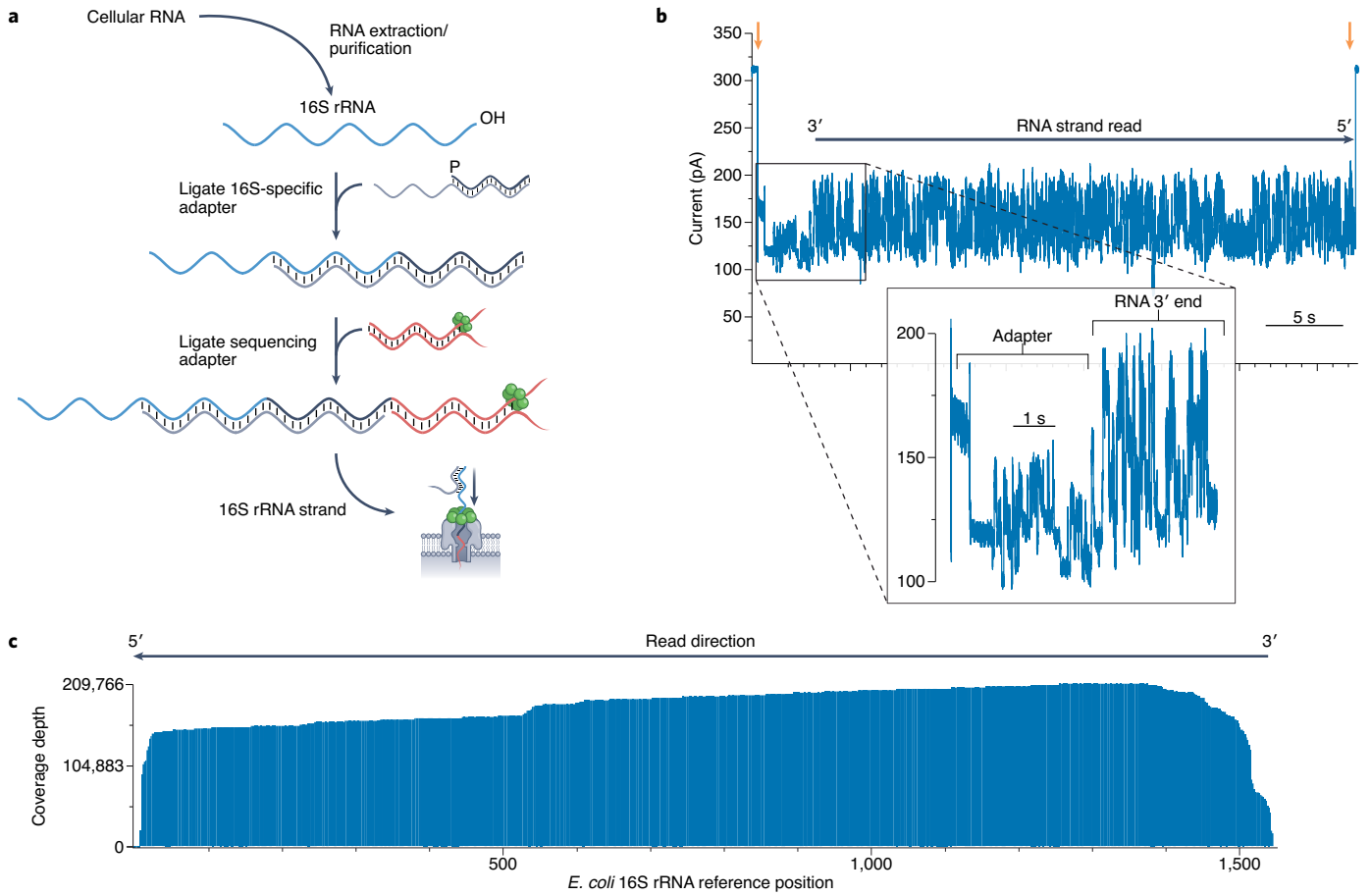

Fig. 1 | Nanopore sequencing of individual *E. coli* 16S ribosomal RNA strands. **a**, Library preparation for MinION sequencing. Following RNA extraction, a 16S rRNA-specific adapter is hybridized and ligated to the 16S rRNA 3′ end. Next, a sequencing adapter bearing a RNA motor protein is hybridized and ligated to the 3′ overhang of the 16S rRNA adapter. The sample is then loaded into the MinION flowcell for sequencing. **b**, Representative ionic current trace during translocation of a 16S rRNA strand from *E. coli* strain MRE600 through a nanopore. Following capture of the 3′ end of an adapted 16S rRNA, the ionic current transitions from open channel (310 pA; gold arrow) to a series of discrete segments characteristic of the adapters (inset). This is followed by ionic current segments corresponding to base-by-base translocation of the 16S rRNA. The trace is representative of thousands of reads collected for individual 16S rRNA strands from *E. coli*. **c**, Alignment of 200,000+ 16S rRNA reads to *E. coli* strain MRE600 16S rRNA *rrnD* gene reference sequence. Reads are aligned in 5′ to 3′ orientation, after being reversed by the basecalling software. Numbering has been done according to the canonical *E. coli* 16S sequence. Coverage across reference is plotted as a smoothed curve. In this experiment, 94.6% of reads that passed quality filters aligned to the reference sequence. Data presented here are from a single flow cell. P, monophosphate; OH, hydroxyl. Figure adapted from ref. [23] under a CC-BY license.

these events are then calculated and used to model Gaussian distributions for downstream analysis. In the following text, we highlight how Nanopolish-integrated programs are enabling the analysis of RNA 3′-tails and RNA modifications.

Nanopore DRS analysis of RNA 3′ ends

Nanopolish combined with nanopore DRS can directly quantify biological poly(A) RNA tail lengths without added chemistry steps[6]. For example, Tudek et al.[15] documented the relationship between tail length, RNA expression levels and yeast growth conditions. Further, nanopore DRS combined with in vitro polyadenylation facilitated discovery of antisense transcripts in *Pseudomonas*[16], and has been applied to archaeal transcriptomes[17]. Extension of 3′ termini with polyinosine augments this technology by allowing analysis of both non-polyadenylated and polyadenylated RNA molecules[18,19].

RNA modifications

In addition, nanopore DRS has been used to detect RNA modifications[20–22]. The general schemes are illustrated in Fig. 2. As an RNA strand translocates through the nanopore, the monovalent ionic current is altered depending on nucleotides that occupy the pore sensor (Fig. 2a). These ionic current data are basecalled and aligned to a reference sequence (Fig. 2b). The ionic current alignment can be visualized (Fig. 2c). To date, the ONT RNA basecaller is only trained to recognize canonical nucleotides; therefore, base-level errors can arise at modified positions (Fig. 2d). These basecall errors have served as useful coarse-grained indicators of RNA modifications, in some cases supported by orthogonal validation[23,24].

In our view, more principled software tools use the underlying ionic current signal to detect RNA modifications (Fig. 2e). These tools are summarized in two recent reviews[20,21]. Most use Nanopolish to

assign discrete ionic current events to pentamers and then distinguish between modified and unmodified *k*-mers using Gaussian mixture models or neural networks[25]. For this scheme, independent validation is warranted for modification predictions, as is true for the base miscall scheme (Fig. 2f). Here, we summarize nanopore-DRS-based detection of three common RNA modifications.

m6A. In 2020, Parker et al. used nanopore DRS to document transcript abundance for *Arabidopsis thaliana* mRNA. A key component of this study was the characterization of m6A modifications and their role in circadian rhythm[26]. A similar study employed nanopore DRS to profile m6A in *Populus trichocarpa*[27]. Importantly, both studies performed orthogonal tests of their predictions.

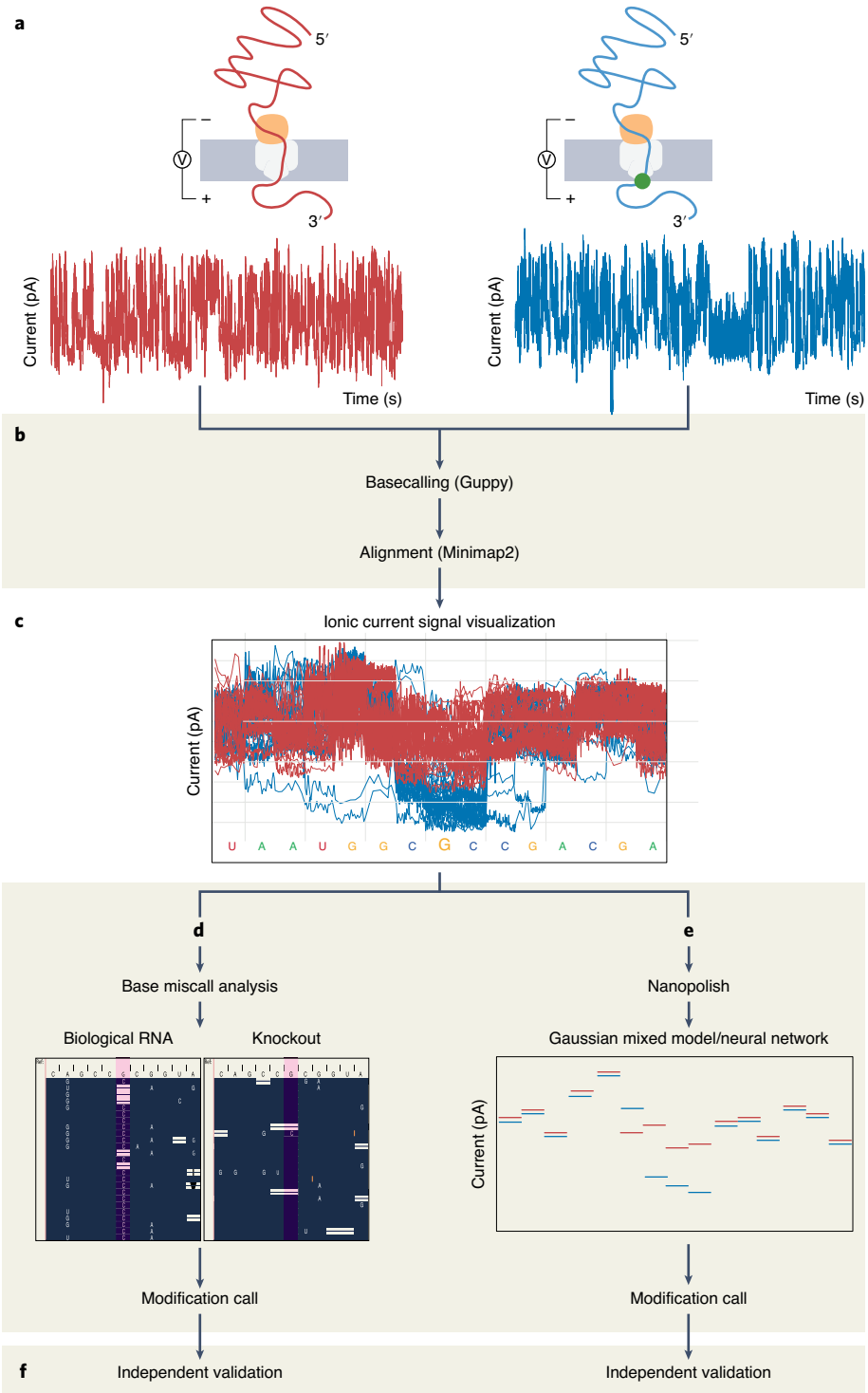

Fig. 2 | General schemes for modification detection using nanopore DRS. a, RNA strands with (green ball) and without RNA modifications are captured and translocated through the nanopore sensor producing ionic current signatures. **b**, ONT proprietary software (Guppy) converts the ionic current signatures into basecalls, which can be aligned to a reference sequence using Minimap2. **c**, Ionic current signatures corresponding to basecalls can be aligned and visually compared. This is facilitated by 'time warping' events to a fixed time interval (in reality, the dwell times vary substantially between events). **d**, The path to identifying RNA modifications using base miscalls. This strategy uses base-level sequence alignments to visualize and count miscalls in biological RNA strands and in control RNA standards where modifications are absent. In the cartoon example, C miscalls (pink) are common for the biological sample at one position, suggesting a modification; for the corresponding knockout strain, most of the basecalls fall under G in agreement with the reference sequence suggesting no modification at that position. **e**, Path for identification of miscalls using Gaussian mixture models or neural networks. First, Nanopolish assigns ionic current segments (events) to pentamers and yields their mean (in pA), standard deviation (in pA) and dwell time (in ms). These events are then used by machine-learning approaches (Gaussian mixture models or neural networks) to learn ionic current signatures that are associated with specific modifications in known sequence contexts. This process yields trained models that are used to predict modifications. **f**, Independent validation of modification calls using orthogonal techniques. These include mass spectrometry, knockouts and knockdowns, reverse transcription stops and chemical adducts. Note that the subpanels are intended to represent concepts; together, they do not represent an actual experiment. V, applied voltage.

BOX 1

Advice for new users

In the main text, published nanopore DRS users are described as 'early adopters' because they fit that category in the classic text *Diffusion of Innovations*[35]. A larger group is the 'early majority', which, in our field, we associate with molecular biologists and cell biologists. Here we offer advice to investigators in this group who are considering nanopore DRS for their laboratories.

- Acquire bioinformatics skills or collaborate with someone who has those skills. For example, you should be able to use Nanopolish and Minimap2.
- Combine nanopore cDNA sequencing (R10.4 platform) with nanopore DRS to provide a more comprehensive picture of the RNA expression landscape. The former is more accurate with higher throughput, whereas the latter contains more information.
- Stay up-to-date with ONT improvements on sample preparation and sequencing protocols.
- Consider foregoing poly(A) RNA selection prior to ONT adapter ligation. Viscardi and Arribere[36] found that the poly(dT)-bearing ONT adapter alone gave satisfactory poly(A) enrichment without prior poly(A) selection. RNA strand read counts and read lengths were comparable using this abbreviated protocol.
- For RNA modification analysis, Nanocompore is a good place to start. This software is well documented, easy to install, and threshold statistics can be tuned. Nanocompore requires a reference standard wherein the modification of interest has been deleted (for example, by a modifying enzyme knockout or by in vitro transcription). For m6A modifications in mRNA, xPore[37] and Yanocomp[38] are accessible and sound alternatives. Users should also be cautious of nanopore-focused software tools from the academic community that have not been implemented beyond the host laboratory.
- We recommend MinION for new users because it is inexpensive and robust. However, in our hands the PromethION can deliver substantially more reads (4–8 million per flow cell) than MinION (1–2 million per flow cell) using comparable amounts of input RNA.

Inosine. Complementary DNA (cDNA) sequencing protocols read A-to-I substitutions as guanosine, but are limited by short read lengths and incomplete reference annotations. Nanopore cDNA sequencing coupled with nanopore DRS can largely overcome these limitations[28].

Dinopore, an inosine-specific software tool, used a convolutional neural network to learn nanopore ionic current signatures for 81 RNA pentamers in which inosine was the centermost nucleotide[28]. This approach distinguished inosine from adenosine and guanosine in human, mouse and *Xenopus* transcriptomes[28]. These authors were also able to estimate the modification rate at each A-to-I editing site[28].

Pseudouridine. Nanopore DRS has also been applied to pseudouridine detection. Early work used U-to-C miscalls, which are coincidental errors in nanopore RNA basecalling software.

Recently, a more principled software tool based on the ionic current signal (NanoPsu) was developed for Ψ detection in human mRNA[29]. This tool identified interferon inducible pseudouridines in interferon-stimulated human transcripts[29]. Similar approaches were used to identify modifications in yeast ribosomal RNA (rRNA) (supported by small-nucleolar-RNA-based knockouts)[30], and to identify Ψ sites in human RNA[31].

An important, but often-overlooked phenomenon, is the impact of RNA enzyme motor function on strand translocation rate. For example, Fleming et al.[32] showed that pseudouridine in SARS-CoV-2 RNA slowed strand translocation when the modification resided in the motor.

Further technical improvements to broaden nanopore DRS use

The previous text summarized recent advances in nanopore DRS. In Box 1, we offer our thoughts on how new users can best implement those advances. Below we propose improvements in the technology that will take nanopore DRS to the next level.

Higher basecall accuracy. Nanopore single-strand ('simplex') DNA basecall accuracy is 99%[11]. We believe that raising nanopore DRS accuracy to that level would revolutionize the field.

Decreased RNA input. For differentiated cells — for example, cells along the path from embryonic stem cells to engineered beta cells[33] — harvesting 500 ng of poly(A) RNA is impractical. Early unpublished evidence (discussed in the previous text) indicated that useful RNA read counts can be achieved using only 50 ng of mRNA. There is still headroom for improvement because the vast majority of RNA strands applied to nanopore flow cells are not sequenced.

Routine validation of RNA modification calls. Until nanopore DRS modification calls become more precise and quantitative, we believe that it is essential to routinely validate those calls, most notably using mass spectrometry.

Full-length reads. The term 'full length' is widely used in nanopore DRS papers, but apart from tRNA[7] this is inaccurate because the nanopore enzyme motor typically releases the captured strand 10–12 nucleotides from the 5′ terminal base. This obscures the true identity of that strand end. In our hands, even when we demonstrably sequenced poly(A) RNA from the 3′-poly(A) tail to the 5′-m7G cap, we could not resolve the final six nucleotides[10]. Consequently, it is important that investigators either achieve truly full-length reads or that they define what is meant by the term.

Validation of newly discovered mRNA isoforms. Nanopore DRS facilitates discovery of previously unknown mRNA isoforms. These preliminary discoveries require validation using orthogonal data such as reverse transcription (RT)–PCR, and transcription start site markers such as DNase-seq, polymerase II chromatin immunoprecipitation (ChIP)-seq, SPI1 ChIP-seq and CAGE (cap analysis of gene expression)[10]. For mRNA, unambiguous validation requires documentation of an associated protein[34].

Straightforward implementation of software developed by the academic community. To our knowledge, most academic software tools designed for nanopore DRS modification analysis have not been

used or validated outside the host laboratory. This should be rectified to facilitate software adoption and replication of experiments.

Miten Jain[1] ✉, Robin Abu-Shumays[2], Hugh E. Olsen[2] and Mark Akeson[2] ✉

[1]Northeastern University, Boston, MA, USA. [2]University of California, Santa Cruz, CA, USA.

✉e-mail: mi.jain@northeastern.edu; makeson@soe.ucsc.edu

Published online: 6 October 2022

References

1. Garalde, D. R. et al. *Nat. Methods* **15**, 201–206 (2018).
2. Viehweger, A. et al. *Genome Res.* **29**, 1545–1554 (2019).
3. Wongsurawat, T. et al. *Front. Microbiol.* **10**, 260 (2019).
4. Kim, D. et al. *Cell* **181**, 914–921 (2020).
5. Ugolini, C. et al. *Nucleic Acids Res.* **50**, 3475–3489 (2022).
6. Workman, R. E. et al. *Nat. Methods* **16**, 1297–1305 (2019).
7. Thomas, N. K. et al. *ACS Nano* **15**, 16642–16653 (2021).
8. Rousseau-Gueutin, M. et al. *Gigascience* **9**, giaa137 (2020).
9. Grünberger, F., Ferreira-Cerca, S. & Grohmann, D. *RNA* **28**, 400–417 (2022).
10. Mulroney, L. et al. *RNA* **28**, 162–176 (2022).
11. Sereika, M. et al. *Nat. Methods* **19**, 823–826 (2022).
12. Li, R. et al. *Genome Res.* **30**, 287–298 (2020).
13. Loman, N. J., Quick, J. & Simpson, J. T. *Nat. Methods* **12**, 733–735 (2015).
14. Li, H. *Bioinformatics* **34**, 3094–3100 (2018).
15. Tudek, A. et al. *Nat. Commun.* **12**, 4951 (2021).
16. Pust, M.-M., Davenport, C. F., Wiehlmann, L. & Tümmler, B. *J. Bacteriol.* **204**, e0041821 (2022).
17. Grünberger, F. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2019.12.18.880849 (2020).
18. Vo, J. M. et al. *RNA* **27**, 1497–1511 (2021).
19. Drexler, H. L. et al. *Nat. Protoc.* **16**, 1343–1375 (2021).
20. Furlan, M. et al. *RNA Biol.* **18**, 31–40 (2021).
21. Abebe, J. S., Verstraten, R. & Depledge, D. P. *mBio* **13**, e0370221 (2022).
22. White, L. K., Strugar, S. M., MacFadden, A. & Hesselberth, J. R. Preprint at *bioRxiv* https://doi.org/10.1101/2022.05.29.493267 (2022).
23. Smith, A. M., Jain, M., Mulroney, L., Garalde, D. R. & Akeson, M. *PLoS ONE* **14**, e0216709 (2019).
24. Begik, O. et al. *Nat. Biotechnol.* **39**, 1278–1291 (2021).
25. Leger, A. et al. *Nat. Commun.* **12**, 7198 (2021).
26. Parker, M. T. et al. *eLife* **9**, e49658 (2020).
27. Gao, Y. et al. *Genome Biol.* **22**, 22 (2021).
28. Nguyen, T. A. et al. *Nat. Methods* **19**, 833–844 (2022).
29. Huang, S. et al. *Genome Biol.* **22**, 330 (2021).
30. Bailey, A. D. et al. *eLife* **11**, e76562 (2022).
31. Tavakoli, S. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2021.11.03.467190 (2022).
32. Fleming, A. M., Mathewson, N. J. & Burrows, C. J. *ACS Cent. Sci.* **7**, 1707–1717 (2021).
33. Pagliuca, F. W. et al. *Cell* **159**, 428–439 (2014).
34. Miller, R. M. et al. *Genome Biol.* **23**, 69 (2022).
35. Rogers, E. M. *Diffusion of Innovations* 5th edn (Simon and Schuster, 2003).
36. Viscardi, M. J. & Arribere, J. A. *BMC Genomics* **23**, 530 (2022).
37. Pratanwanich, P. N. et al. *Nat. Biotechnol.* **39**, 1394–1402 (2021).
38. Parker, M. T., Barton, G. J. & Simpson, G. G. Preprint at *bioRxiv* https://doi.org/10.1101/2021.06.15.448494 (2021).

Acknowledgements

M.J., H.E.O. and R.A.S. were supported by NIH grant HG010053.

Competing interests

M.A. holds options in ONT and is a paid consultant to ONT. H.E.O. and M.J. received reimbursement for travel, accommodation and conference fees to speak at events organized by ONT.