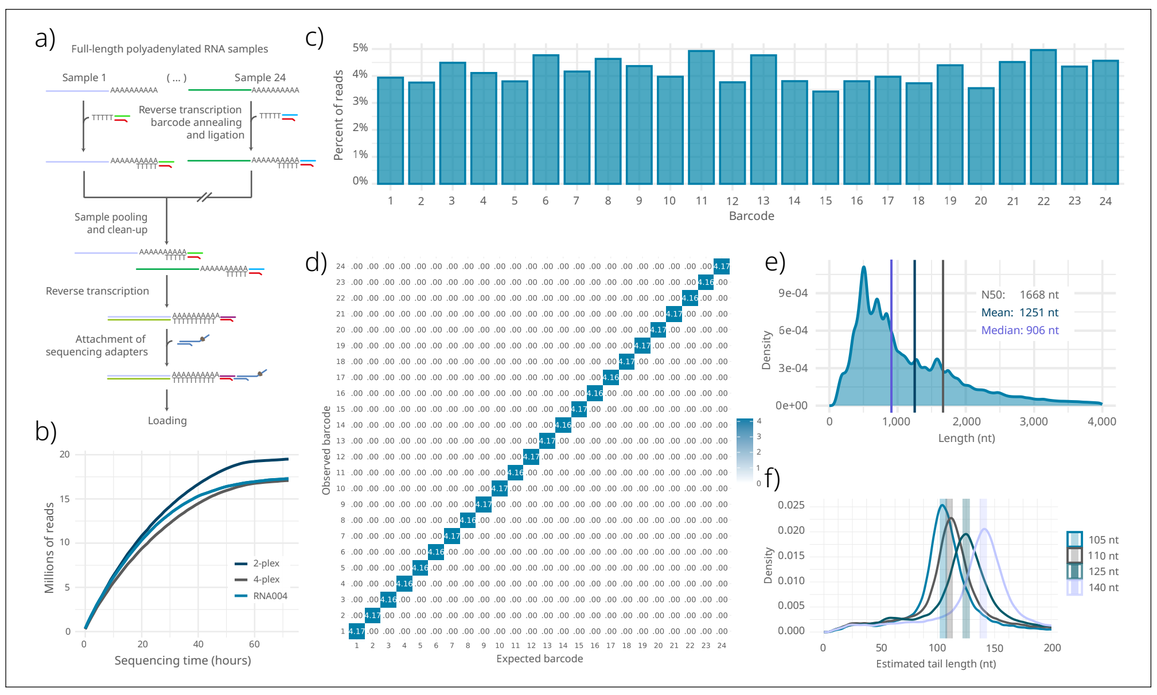
Oxford Nanopore Technologies

Transcriptome-wide expression and RNA modifications with full-length native RNA and cDNA sequencing

High-throughput native RNA and cDNA sequencing enables quantitative and transcriptome-wide capture of full-length transcripts, native detection of common RNA modifications, and poly(A) tail length measurement at single-molecule resolution



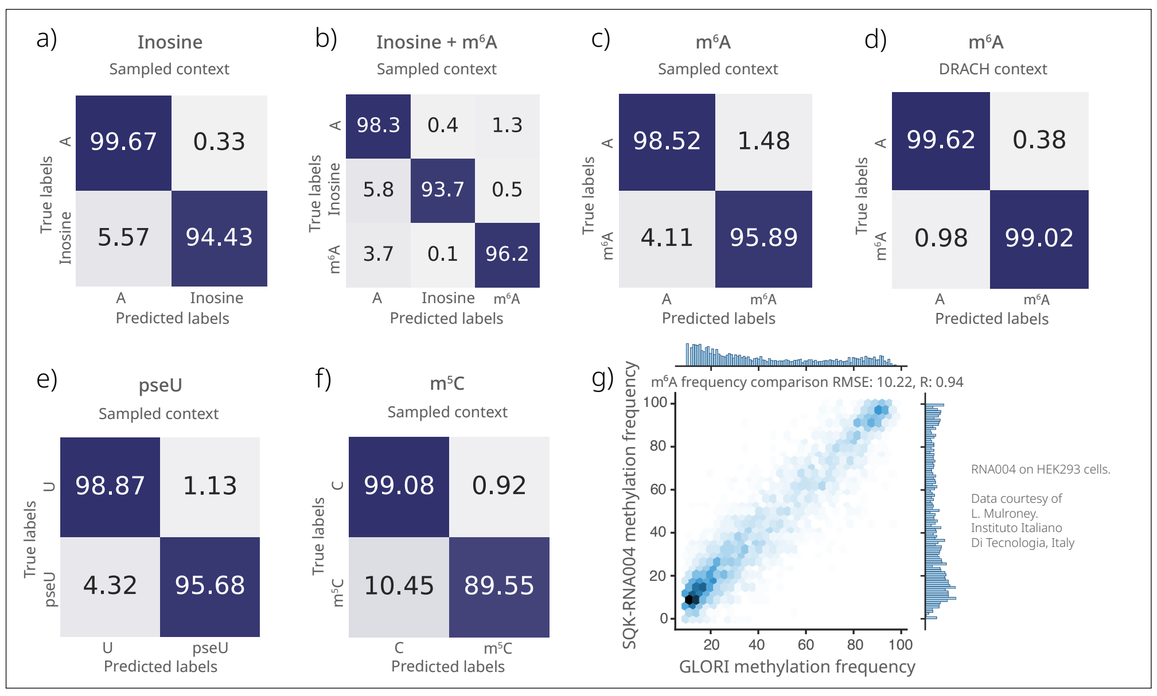


Fig. 1 RNA barcoding (a) protocol (b) mean read output (c) balance (d) count normalised classification (e) read length and (f) tail length estimation.

Fig. 2 Detection of (a) inosine (b) inosine with m⁶A (c) all-context m⁶A (d) DRACH-context m⁶A (e) pseU (f) m⁵C (g) m⁶A compared to GLORI.

RNA barcoding enables multiplexing up to 24 samples

Our new direct RNA barcoding kit enables combining up to 24 different samples on a single flow cell (**Fig. 1a**). We prepared poly(A)+ libraries from Universal Human Reference RNA (UHRR) and *in vitro* transcribed (IVT) transcripts and sequenced them on PromethION™ Flow Cells. The human libraries yielded approximately 18 million raw reads, comparable to SQK-RNA004 native RNA libraries (**Fig. 1b**). A 24-plex run of human poly(A)+ RNA showed good barcode balance (**Fig. 1c**). The 24 unique IVT transcripts, each assigned a unique barcode, were correctly classified by the Dorado basecaller (**Fig. 1d**). A barcoded run on human transcriptome yielded a median read length of 906 nt (**Fig. 1e**). A barcoded run with four IVT transcripts with distinct poly(A) tail lengths shows that direct RNA barcoding is compatible with poly(A) tail length estimation (**Fig. 1f**). Direct RNA barcoding is currently in beta testing.

Easy detection of common RNA modifications

The Direct RNA Sequencing Kit (SQK-RNA004), together with improved basecalling models, enables easy detection of RNA modification sites at single-molecule, single-nucleotide resolution directly from raw signal during basecalling without the need for additional experiments or treatments. Using a publicly available dataset derived from synthetic oligonucleotides (**epi2me.nanoporetech.com/rna-mod-validation-data**), we evaluated the performance of our latest new and improved RNA modified basecalling models for inosine (**Fig. 2a**), inosine and m⁶A (**Fig. 2b**), all context and DRACH-motif context m⁶A (**Fig. 2c–d**), pseU (**Fig. 2e**), and m⁵C (**Fig. 2f**). Each model showed high discrimination between modified and non-modified bases. The models are available to use with the new Dorado basecaller release. Direct detection of m⁶A with Oxford Nanopore sequencing has a high correlation with an orthogonal method GLORI (Liu *et al.*) (**Fig. 2g**).

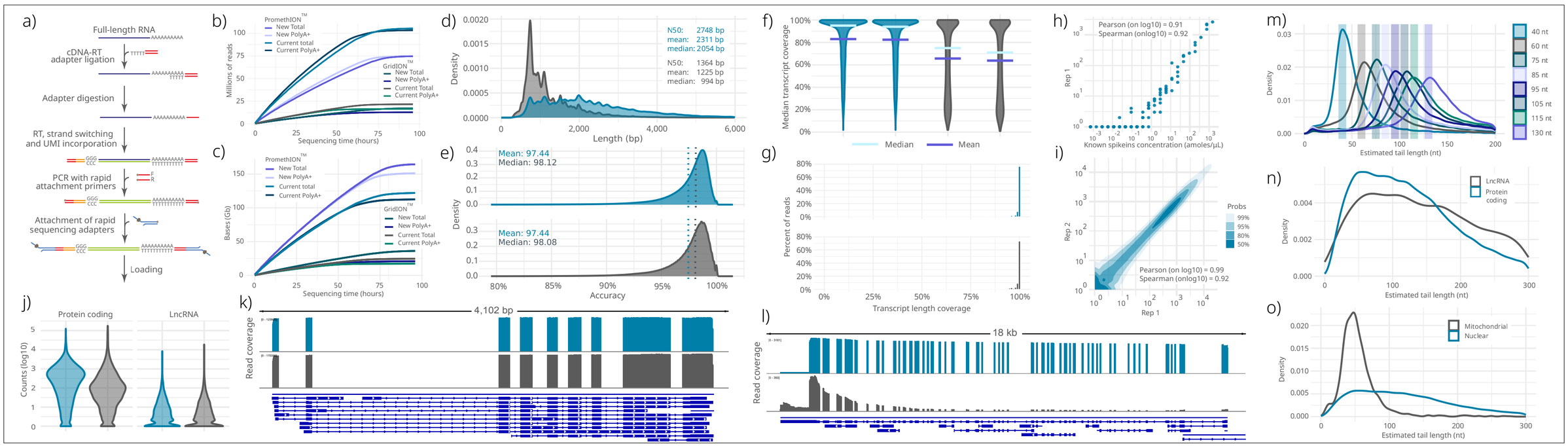


Fig. 3 cDNA sequencing (a) protocol (b) mean read output (c) mean base output (d) read length (e) accuracy (f) median human transcript coverage (g) distribution of spike-in transcript coverage (h) spike-in expression (i) human expression (j) gene counts by biotype (k) GAPDH on IGV (l) COL1A1 on IGV (m) tail lengths for IVT controls (n) tail lengths for nuclear-encoded protein-coding and lncRNAs (o) tail lengths for mitochondrial and nuclear-encoded transcripts. The blue plot colour indicates the new method and the grey indicates the current method, unless otherwise stated.

Human transcriptome analysis with an updated cDNA sequencing protocol that captures long transcripts

The cDNA-PCR Sequencing Kit (SQK-PCS114) enables global gene and transcript expression analysis from PCR-amplified material (**Fig. 3a**). A new protocol, currently in development, enhances full-length capture of long transcripts. We prepared replicate libraries from poly(A)+ and total RNA transcriptome samples from UHRR with 1% of Lexogen Spike-in RNA Variant (SIRV) Control Set 4, added using the current and augmented protocol, before sequencing the libraries on MinION™ and PromethION™ Flow Cells. We also pooled, prepared, and sequenced eight IVT transcripts with known tail lengths. The new protocol yielded approximately 75% read output (**Fig. 3b**) and improved gigabase output (**Fig. 3c**) compared with the current protocol, with an increased read length profile (**Fig. 3d**) and similar mapping accuracy (**Fig. 3e**). Full-length coverage along human and spike-in transcripts was improved (**Fig. 3f–g**) due to better recovery of long transcripts. The obtained read counts were highly correlated with the known concentration of SIRVs (**Fig. 3h**), as were the human gene counts between replicates (**Fig. 3i**). Most read counts came from protein coding genes, but lower abundance lncRNAs were also captured (**Fig. 3j**). Full transcript length coverage on IGV remained good for GAPDH (**Fig. 3k**) while improving for others such as COL1A1 (**Fig. 3l**). Tail length estimation on tail length controls recovered expected tail lengths (**Fig. 3m**). Poly(A) tail lengths were estimated to be shorter for nuclear-encoded protein coding genes than lncRNAs (**Fig. 3n**), while mitochondrial protein coding genes were estimated to have shorter tails than nuclear encoded genes (**Fig. 3o**).