Oxford NANOPORE Technologies™
Research & Development

Comprehensive transcriptome analysis with latest cDNA sequencing chemistry from Oxford Nanopore

Single-molecule nanopore sequencing of cDNA enables easy, transcriptome-wide analysis of gene and transcript expression, untranslated regions (UTRs), splicing, and poly(A) site usage and tail length dynamics

Contact details: **sample.tech@nanoporetech.com**
More information at **www.nanoporetech.com**

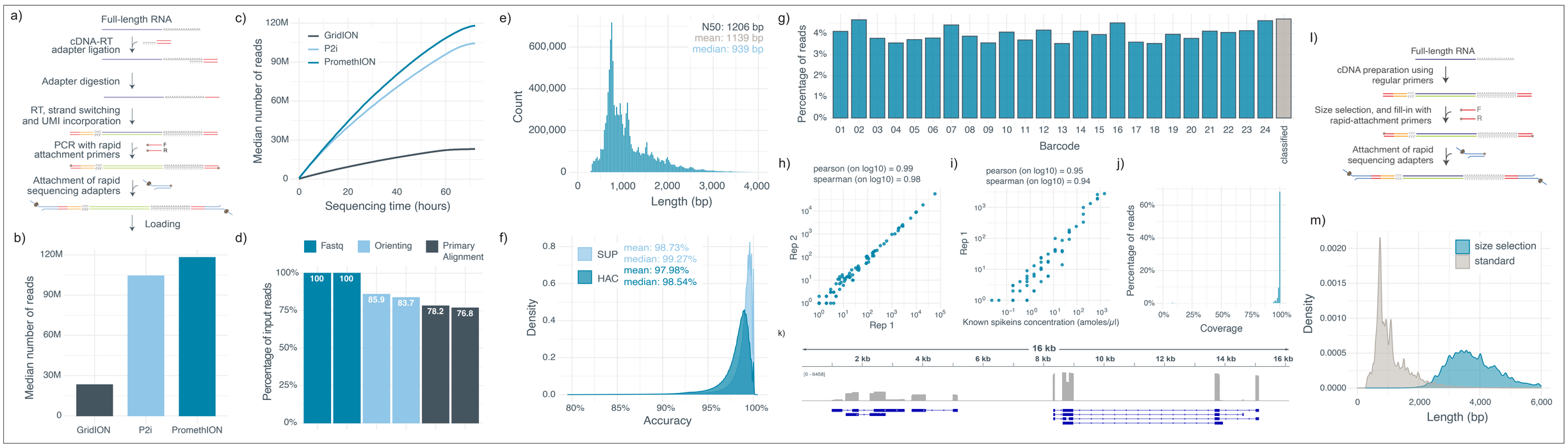

Fig. 1 Methods and metrics a) standard workflow b-c) raw-read output d) mapping rate e) raw-read length of mapped reads f) mapping accuracy g) barcode balance h) spike-in count correlation between replicates i) expected vs observed spike-in counts j) spike-in transcript coverage k) SIRV4 spike-in transcript l) size selection workflow m) size selected raw-read length of mapped reads

cDNA-PCR Sequencing Kit updated to the latest Kit 14 chemistry

The cDNA-PCR Sequencing Kit (SQK-PCS114) and its barcoded counterpart (SQK-PCB114.24) have been updated to Kit 14 chemistry, compatible with the latest 'R10' flow cells. The library preparation method incorporates a unique molecular identifier (UMI) and reverse transcription from a ligated adapter to avoid priming from poly(A) stretches within transcripts (Fig. 1a). We spiked Universal Human Reference RNA (UHRR) (Agilent) with a set of control transcripts of known lengths and concentrations (SIRV-Mix 4) (Lexogen), prepared libraries using SQK-PCS114 and SQK-PCB114.24, and sequenced these libraries in triplicate on GridION™, PromethION™ 48, and P2™ Integrated devices. The kit requires low input amounts of RNA (500 ng total, or 10 ng enriched RNA) and delivers 20M+ reads from MinION™ Flow Cells (GridION device) and 100M+ reads from PromethION Flow Cells (P2 and PromethION devices) in 72 hours of sequencing (Fig. 1b-c). To analyse the data, reads are basecalled during real-time sequencing using the Dorado basecaller from Oxford Nanopore via MinKNOW, or post-run using standalone Dorado. Reads are then strand-oriented with *pychopper* and aligned against a reference genome (or transcriptome) using *minimap2* (Li, H. 2021) (Fig. 1d). Median read length is just under 1 kb (Fig. 1e). High accuracy (HAC) basecalling is recommended for typical transcriptome analyses but super accuracy (SUP) basecalling is available to give an additional 1% increase in mapping accuracy (Fig. 1f). Up to 24 barcodes are available for sample multiplexing and achieve a good barcode balance (Fig. 1g). We used the spike-in transcripts to assess the method's ability to quantify expression and recover reads spanning full-length transcripts. Pairwise count correlation between samples (Fig. 1h) and relative to the know concentrations (Fig. 1i) is high, with majority of reads covering transcripts fully from 5' end to 3' end (Fig. 1j) and with clear delineation of exon/intron boundaries (Fig. 1k). Size selection using SPRI beads (Fig. 1l) can be used to enrich for very long transcript isoforms (Fig. 1m).

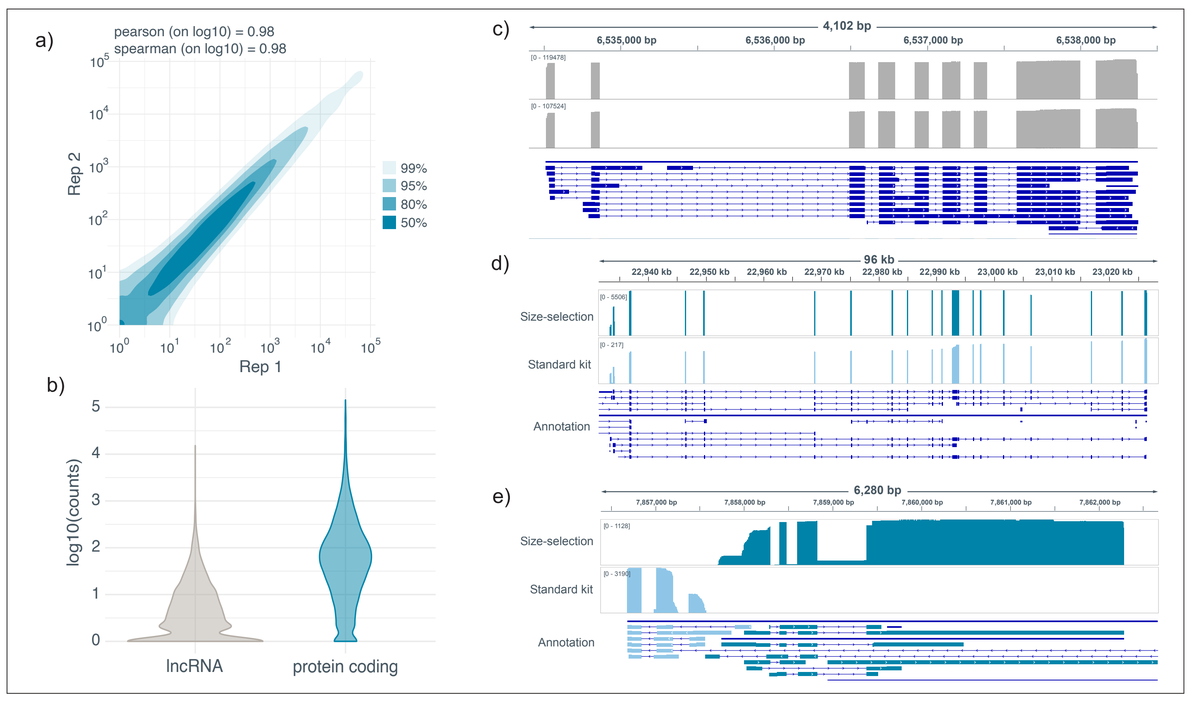

Fig 2. Human transcriptome quantification and isoform analysis a) count correlation between replicates b) gene counts by biotype c) IGV snapshot of *GAPDH* d) *RBBP8* e) *TMEM88/NAA38*

Quantifying human transcriptome expression with cDNA-PCR sequencing

The cDNA sequencing kits SQK-PCS114 and SQK-PCB114.24 enable robust, transcriptome-wide gene expression quantification and capture different biotypes of polyadenylated transcripts. Gene counts are highly correlated between two replicate SQK-PCS114 runs on UHRR samples, and the counts have a high dynamic range (Fig. 2a). The method quantifies both protein coding and long non-coding RNA transcripts (Fig. 2b). Replicate libraries show full 5' to 3' coverage along the GAPDH gene (Fig. 2c). A higher number of reads covering full-length transcripts to the 5' end was observed for a size-selected library (dark blue) compared with a standard library (light blue) for long transcripts of RBBP8 (Fig. 2d). Size selection captures expression and full-length coverage of long transcripts while reducing short ones (Fig. 2e).

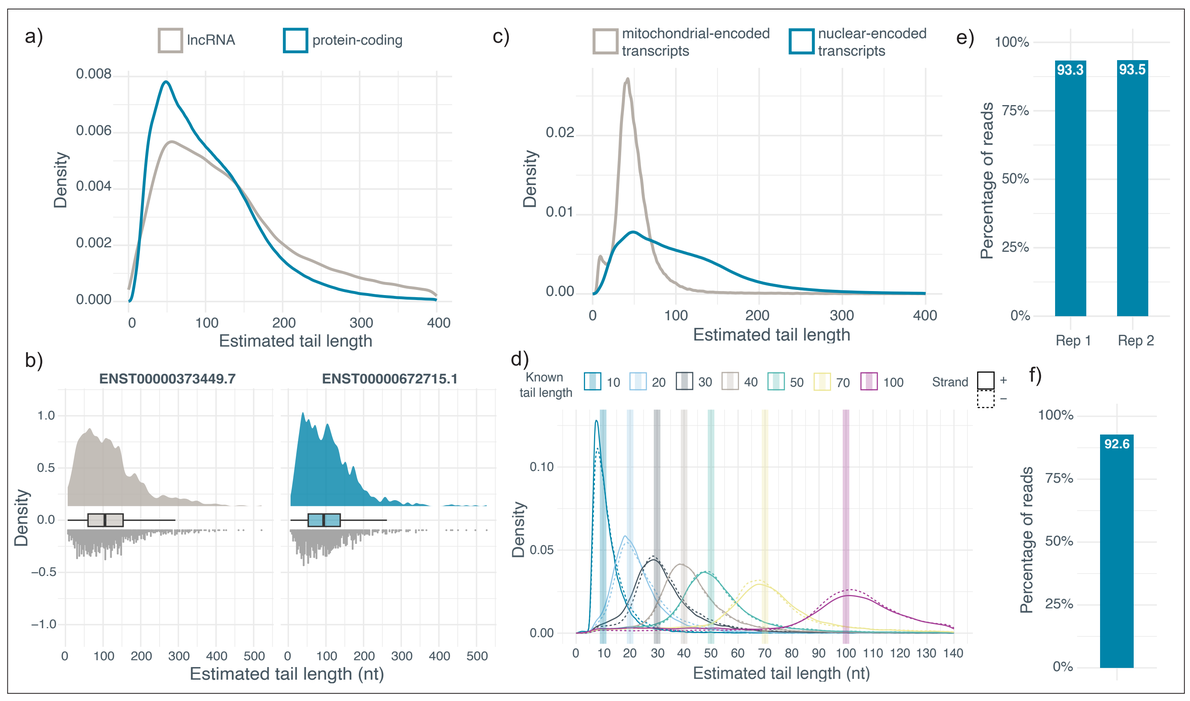

Fig. 3 Poly(A) tail lengths a) by biotype b) by isoform c) by nuclear vs mitochondrial transcripts d) tail standards e) human reads with a tail length estimate f) tails standard reads with a tail length estimate

Estimating per-transcript poly(A) tail lengths using cDNA-PCR sequencing and Dorado

Library preparation with SQK-PCS114 or SQK-PCB114.24 enriches for full-length poly(A) tails. Poly(A) tail length information can be easily estimated for each read during basecalling with the Dorado basecaller. A range of poly(A) tail lengths is captured for reads assigning to human protein coding and long non-coding transcripts (Fig. 3a). Poly(A) tail lengths of transcripts of the same gene can be compared (Fig. 3b) as can transcripts encoded by nuclear or mitochondrial genes (Fig. 3c). Tail length standards were correctly estimated for in-house standards ranging from 10-100 nucleotides (Fig. 3d). A poly(A) tail was detected by Dorado in >92% of all human (Fig. 3e) and tail standard (Fig. 3f) reads.