Nanopore sequencing of RNA and cDNA molecules in *Escherichia coli*

Felix Grünberger[1,*], Sébastien Ferreira-Cerca[2,3] and Dina Grohmann[1,2,*]

[1] Institute of Biochemistry, Genetics and Microbiology, Institute of Microbiology and Archaea Centre, Single-Molecule Biochemistry Lab & Biochemistry Centre Regensburg, University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany

[2] Regensburg Center of Biochemistry (RCB), University of Regensburg, 93053 Regensburg, Germany

[3] Institute for Biochemistry, Genetics and Microbiology, Regensburg Center for Biochemistry, Biochemistry III, University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany

* To whom correspondence should be addressed.

Tel.: 0049 941 943 3147; Fax: 0049 941 943 2403; Email: dina.grohmann@ur.de

Present Address: Dina Grohmann, Department of Biochemistry, Genetics and Microbiology, Institute of Microbiology, University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany

Tel.: 0049 941 943 3267; Fax: 0049 941 943 2403; Email: felix.gruenberger@ur.de

Present Address: Felix Grünberger, Department of Biochemistry, Genetics and Microbiology, Institute of Microbiology, University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany

Running title: Nanopore RNA-seq in *E. coli*

Keywords: Nanopore, RNA-seq, transcriptome, bacteria

ABSTRACT

High-throughput sequencing dramatically changed our view of transcriptome architectures and allowed for ground-breaking discoveries in RNA biology. Recently, sequencing of full-length transcripts based on the single-molecule sequencing platform from Oxford Nanopore Technologies (ONT) was introduced and is widely employed to sequence eukaryotic and viral RNAs. However, experimental approaches implementing this technique for prokaryotic transcriptomes remain scarce. Here, we present an experimental and bioinformatic workflow for ONT RNA-seq in the bacterial model organism *Escherichia coli*, which can be applied to any microorganism. Our study highlights critical steps of library preparation and computational analysis and compares the results to gold standards in the field. Furthermore, we comprehensively evaluate the applicability and advantages of different ONT-based RNA sequencing protocols, including direct RNA, direct cDNA, and PCR-cDNA. We find that (PCR)-cDNA-seq offers improved yield and accuracy compared to direct RNA sequencing. Notably, (PCR)-cDNA-seq is suitable for quantitative measurements and can be readily used for simultaneous and accurate detection of transcript 5´ and 3´ boundaries, analysis of transcriptional units and transcriptional heterogeneity. In summary, based on our comprehensive study, we show that Nanopore RNA-seq to be a ready-to-use tool allowing rapid, cost-effective, and accurate annotation of multiple transcriptomic features. Thereby Nanopore RNA-seq holds the potential to become a valuable alternative method for RNA analysis in prokaryotes.

INTRODUCTION

In the last decade, next-generation sequencing (NGS) technologies (Levy and Myers, 2016) revolutionised the field of microbiology (Escobar-Zepeda et al., 2015), which is not only reflected in the exponential increase in the number of fully sequenced microbial genomes but also in the detection of microbial diversity in many hitherto inaccessible habitats based on metagenomics. Using transcriptomics, important advances were also possible in the field of RNA biology (Wang et al., 2009; Hör et al., 2018) that shaped our understanding of the transcriptional landscape (Croucher and Thomson, 2010; Nowrousian, 2010) and RNA-mediated regulatory processes in prokaryotes (Saliba et al., 2017). RNA sequencing (RNA-seq) technologies can be categorised according to their platform-dependent read lengths and the necessity of a reverse transcription and amplification step to generate cDNA (Stark et al., 2019). Illumina sequencing yields highly accurate yet short sequencing reads (commonly 100-300 bp). Hence, sequence information is only available in a fragmented form, making full-length transcript- or isoform-detection a challenging task (Tilgner et al., 2015; Byrne et al., 2019). Sequencing platforms developed by Pacific Bioscience (PacBio) and Oxford Nanopore Technologies (ONT) solved this issue. Both sequencing methods are *bona fide* single-molecule sequencing techniques that allow the sequencing of long DNAs or RNAs (Eid et al., 2009; Mikheyev and Tin, 2014). However, the base detection differs significantly between the two methods. PacBio-sequencers rely on fluorescence-based single-molecule detection that identifies bases based on the unique fluorescent signal of each nucleotide during DNA synthesis by a dedicated polymerase (Eid et al., 2009). In contrast, in an ONT sequencer, the DNA or RNA molecule is pushed through a membrane-bound biological pore with the aid of a motor protein attached to the pore protein called a nanopore. A change in current is caused by the translocation of the DNA or RNA strand through this nanopore, which serves as a readout signal for the sequencing process. Due to the length of the nanopore (version R9.4), a stretch of approximately five bases contributes to the current signal. Notably, only ONT-based sequencing offers the possibility to directly

sequence native RNAs without the need for prior cDNA synthesis and PCR amplification (Soneson et al., 2019). Direct RNA sequencing based on the PacBio platform has also been realised but requires a customised sequencing workflow using a reverse transcriptase in the sequencing hotspot instead of a standard DNA polymerase (Vilfan et al., 2013). Direct RNA-seq holds the capacity to sequence full-length transcripts and has been demonstrated as a promising method to discriminate and identify RNA base modifications (e.g. methylations (Liu et al., 2019; Smith et al., 2019; Parker et al., 2020; Begik et al., 2021; Jenjaroenpun et al., 2021)). ONT sequencing is a *bona fide* single-molecule technique and hence offers the possibility to detect molecular heterogeneity in a transcriptome (Workman et al., 2019). Recently, the technology was exploited to sequence viral RNA genomes (Keller et al., 2018; Boldogkői et al., 2019; Viehweger et al., 2019; Wang et al., 2021) to gain insights into viral and eukaryotic transcriptomes (Tombácz et al., 2019; Zhao et al., 2019; Sahlin et al., 2021) and to detect and quantify RNA isoforms in eukaryotes (Byrne et al., 2017; Workman et al., 2019; Parker et al., 2020; Dong et al., 2021; Seki et al., 2021). Essentially, the requirements, but also the possibilities in eukaryotes and prokaryotes, are the same (Choi, 2016), with a poly(A) tail being an essential prerequisite, which is required to capture the RNAs. Using enzymatic polyadenylation of prokaryotic RNAs that in general lack poly(A) tails, the applicability of Nanopore RNA-seq has already been demonstrated by metatranscriptomic sequencing of bacterial food pathogens (Yang et al., 2020) and by accurate estimation of gene expression levels in *Klebsiella pneumoniae* (Pitt et al., 2020). Despite these initial studies, a comprehensive analysis of the applicability of Nanopore RNA-seq for the analysis of prokaryotic transcriptomes is lacking.

In this study, we applied and compared all currently available ONT library preparation methods to analyse RNAs in the prokaryotic model organism *Escherichia coli* K-12. These include direct sequencing of native RNAs, direct sequencing of cDNAs, and sequencing of PCR-amplified cDNAs. The goal was to create a robust workflow for the simultaneous determination of multiple transcriptional features. To this end, we analysed the reproducibility

and comparability of transcript quantification, evaluated the accuracy of transcript boundary identification and the potential of long-read ONT RNA-seq to capture the complexity of bacterial transcriptional units. Noteworthy, due to the single-molecule resolution of ONT sequencing, in-depth analysis of transcription units becomes possible. In addition, we point out practical and technical considerations of the different methods such as the effects of rRNA depletion on the sequencing depth, the possibility to enrich for full-length transcripts in the cDNA protocols and the effects of read trimming.

RESULTS

Experimental design for comprehensively comparing Nanopore sequencing of RNA and cDNA molecules in *Escherichia coli*

Currently, three different protocols from ONT are available for the analysis of RNAs including i) direct sequencing of native RNAs (SQK-RNA002, referred to as DRS in this study), ii) direct sequencing of cDNAs (SQK-DCS109, referred to as cDNA in this study) and iii) sequencing of PCR-amplified cDNAs (SQK-PCB109, referred to as PCR-cDNA in this study) (Figure 1A, Supplementary Fig. 1). Since there is a crucial difference between sequencing RNA or DNA, we additionally use the combined term (PCR)-cDNA-seq, which refers to both cDNA and PCR-cDNA approaches. In short, all methods rely on polyadenylated RNAs as starting material since RNAs are either annealed to an oligo(dT) primer for (PCR)-cDNA approaches or ligated to a double-stranded oligo(dT) splint adapter in the DRS approach. Although reverse transcription is optional for DRS, it is highly recommended by ONT and the community to resolve secondary structures in the RNA and to decrease the probability of pore blockage, which ultimately results in an increase in total throughput (Workman et al., 2019). However, only the RNA strand carries the motor protein and is subsequently sequenced. The (PCR)-cDNA protocols take advantage of the template-switching ability of the reverse transcriptase, which adds a few non-templated Cytosines to the end of the cDNA (Matz et al., 1999). This allows the enrichment of full-length sequenced transcripts during the analysis (Supplementary Fig. 1). After RNA digestion, the second strand is synthesized, followed by barcode ligation, PCR amplification in the PCR-cDNA protocol, attachment or ligation of sequencing adapters and sequencing.

We performed all three protocols using unfragmented total RNA prepared from the prokaryotic model organism *E. coli* K-12 strain MG1655 grown at 37°C in rich medium. The aim was to discuss current limitations and best practices analysing prokaryotic transcriptomes using Nanopore sequencing of RNA and cDNA molecules and to compare the results to other full-length sequencing protocols and platforms. Two biological replicates

for each library preparation method were sequenced on a MinION using R9.4 flow cells controlled by MinKNOW. The key steps of library preparation and sequencing are depicted in Figure 1A,B and are briefly summarized in the following: after purification of high-quality RNAs using silica-membrane columns with a cut-off size of about 200 nucleotides, RNAs were immediately polyadenylated to make them amenable for library preparation and to preserve the 3´ends from further degradation during the next steps of the library preparation. Since full-length sequencing of RNAs and cDNAs is dependent on the quality of the source material, we only used RNAs with integrity values (RIN) greater than 9.5 (Schroeder et al., 2006). Also, Bioanalyzer analysis was used to confirm efficient polyadenylation based on a shift in ribosomal RNA peaks and to check fragment size of PCR-amplified cDNAs (Supplementary Fig. 2A,B). To increase the proportion of sequenced mRNAs, ribosomal RNAs, which usually make up the main part of the RNA pool, can be depleted. However, the input quantity requirements currently still make it challenging to use rRNA-depleted RNAs in a sensible and cost-efficient way, especially for DRS. The input amounts are currently listed to be 500 ng polyA+ (DRS), 100 ng polyA+ RNA (cDNA) and 1 ng (PCR-cDNA), respectively. Therefore, we used non-depleted RNA for DRS sequencing, a mix of depleted (40%) and non-depleted RNA (60%) for the cDNA protocol and fully depleted RNA for the PCR-cDNA approach. Additionally, we tested the compatibility with other RNA treatments using the commonly applied digestion of 5´-monophosphorylated non-primary RNAs with a 5´-Phosphate-dependent Terminator Exonuclease (TEX) as an example (Figure 1A). However, it should be noted that we deliberately chose reaction conditions not sufficient for complete digestion of all non-primary RNAs. The intention of this design was not to distinguish primary from processed transcripts but rather to minimize the rRNA content even further.

Overall run and raw read characteristics and analysis of mapped reads

Sequencing throughput on a single FLO-MIN106 flow cell is dependent on the kit chemistry and currently listed by ONT to typically range between 1 to 4 Gb for DRS, more than 8 Gb for cDNA and about 10 Gb for the PCR-cDNA kits. Considering that a higher yield could be

expected for the cDNA kits and the (partial) depletion of ribosomal RNAs, cDNA runs were multiplexed and aborted as soon as a sufficient number of reads (> 0.5Gb) was reached. All sequencing parameters and run statistics are listed in Supplementary Table 1 and shown in Supplementary Fig. 3,4. The sequencing yield of unfiltered reads ranged between 0.09 and 2.21 million reads, or 0.08 Gb to 1.57 Gb, respectively (Supplementary Fig. 3). Read qualities, which are specified as mean qscore values, were similarly distributed within the three library types, showing median values of 8.8 (DRS), 9.7 (cDNA) and 10.5 (PCR-cDNA) (Supplementary Fig. 4A). As expected, the read length distributions of the individual samples were highly dependent on the effect of rRNA depletion (Supplementary Fig. 4B). Although we could confirm the reports of previous studies that very short direct RNA reads are associated with bad quality (Soneson et al., 2019; Workman et al., 2019), we did not see a pronounced effect in other library types or for very long RNAs in our datasets (Supplementary Fig. 4C).

We next aligned the unfiltered reads to the *E. coli* K-12 genome using minimap2 (Figure 1C). 71.4% (DRS), 64.7% (cDNA) and 48.9% (PCR-cDNA) of the reads mapped to the genome, which corresponds to 78.0% (DRS), 64.7% and 47.2% of the bases, respectively (Supplementary Fig. 5). The moderate numbers arise from short reads with low quality, which dominate the class of unmapped reads and are particularly common for the direct RNA datasets but also occur in the (PCR)-cDNA approaches (Supplementary Fig. 6A-D). The lower total number of mapped reads in the PCR-cDNA samples is due to the preference for over-amplification of short fragments in the PCR, which is more pronounced at higher cycle numbers. This suggests that successful sequencing can be estimated reasonably well already from the Bioanalyzer results of PCR-amplified cDNA (Supplementary Fig. 2B). Based on the length distribution, which is similar to the unamplified cDNA and the proportion of mapped reads (62%), we concluded that 12 PCR cycles are sufficient to obtain high quality sequencing data (Supplementary Fig. 5,6).

To allow a detailed analysis of the mapped reads, they were first classified into transcript features and classes using the annotation found in GenBank entry U00096.3 (Riley et al., 2006). Most of the reads in the non-coding RNA class originate from the *ssrA* gene in our sample conditions, producing the transfer-messenger RNA (tmRNA), which explains the uniform length distribution. The tmRNA has tRNA-specific base modifications (Himeno et al., 2014) that lead to an altered current profile, which presumably explains the lower read quality in the DRS approach (Supplementary Fig. 6B). After the RNA is transcribed into cDNA, the modifications are lost, and the quality of the sequenced reads increases significantly. As expected, the raw read length of rRNA-mapping reads was largely dependent on the pre-designed depletion efficiency and subsequent TEX treatment (Supplementary Fig. 6A). Indeed, the number of reads mapping to ribosomal RNAs is significantly reduced in TEX-treated samples compared to the non-treated counterparts (Supplementary Fig. 7).

In the following, we will focus on mRNA-originating reads performing an in-depth analysis of transcriptomic features in *E. coli*. Reads mapping to mRNAs in fully rRNA-depleted libraries make up about 33% of all mapped reads (PCR-cDNA samples with 12 PCR cycles), which corresponds to 42% of all mapped bases (Supplementary Fig. 7). The aligned read length distribution of mRNA-mapping reads was similar between all library types with median values of 406 (DRS), 372 (cDNA) and 395 (PCR-cDNA) bases (Supplementary Fig. 8A). Despite these relatively short median values, there is also a proportion of reads in all library types that are very long and cover large operon structures in one read, which is exemplarily shown for the rpsP-rimM-trmD-rplS operon Figure 1D. This is particularly clear when looking at the mean aligned length of the 100 longest reads in each protocol, which are 4738 (DRS), 6567 (cDNA) and 6132 (PCR-cDNA) bases. At this point, it should be mentioned that the 100 shortest reads have mean lengths of 89 (DRS), 80 (cDNA) and 80 (PCR-cDNA) bases, which is caused by the mapping tool minimap2 used with standard parameters. As previously reported, the mapped read identity of direct RNA reads (88.1%) is substantially

lower as compared to cDNA reads (96% cDNA, 94% PCR-cDNA) (Supplementary Fig. 8B) (Soneson et al., 2019). However, we noticed that the read identity improved when using the RNA002 chemistry instead of the meanwhile outdated RNA001 kit. Although template-switching and second-strand synthesis enriches explicitly for full-length transcripts in all cDNA protocols, no clear difference was detected in the aligned length distribution. The difference in the number of PCR cycles leads to significant differences in mean read lengths (15 cycles: 310 bases, 12 cycles: 526 bases), although there is no effect on read quality and identity.

Comparing raw read length with aligned read length, we noticed that many reads in the cDNA protocol are twice as large as their mapped counterpart, which is caused by reverse transcription artefacts (Perocchi et al., 2007; Tuiskunen et al., 2010) that generate 2D-like reads, containing both strands of a transcript (Supplementary Fig. 9A). Interestingly, the reverse complement part of the read has much lower quality scores than the reverse transcribed RNA. This is confirmed by a correlation analysis between the raw read qualities and the mapped read identity (Supplementary Fig. 9B). Direct RNA and PCR-cDNA reads with low quality led to a lower identity score. This is not observed in the direct cDNA dataset since the distribution is dominated by the low-quality peak of the reverse complement. In most of the cases, only the good-quality first part of the 2D-like read maps to the genome and the aligned read identity is high. The second part of the read, however, is discarded (Supplementary Fig. 9C,D). However, it should be noted that these 2D-like reads make up the majority of reads that have mapped to mRNAs in the cDNA libraries (Supplementary Fig. 9E). Although much less common, the artefact also occurs in PCR-cDNA reads and as expected, is not found in direct RNA reads (Supplementary Fig. 9E).

Reproducibility and comparability of gene quantification

Since the first strand is always sequenced, 2D-like reads are not expected to distort the quantification of reads. To test this and to determine the overall comparability and robustness of the data in absolute quantitative terms, we compared the count data based on

untrimmed, uncorrected Nanopore reads with published short-read (Illumina) and full-length long-read (SMRT-Cappable-Seq protocol, PacBio) cDNA sequencing data from *E. coli* sampled under similar conditions (Yan et al., 2018). Since we only consider reads that map to mRNAs for this purpose, we first looked at the sequencing depth of each dataset to assess whether representative statements can be made. Sequencing depth was dependent on rRNA depletion, TEX treatment and the total number of reads sequenced. Therefore, sequencing depths between 0.2-fold (DRS, RNA002, replicate 2) and 52-fold (PCR-cDNA, 12 cycles, replicate 1) reflect the design of the particular experiment and are mostly comparable to the selected SMRT-Cappable (replicate 1: 51-fold, replicate 2: 7-fold) and short-read Illumina (70-fold) datasets (Supplementary Fig. 10A). Considering the sequencing strategies, (PCR)-cDNA Nanopore sequencing offers a more straightforward way to produce comprehensive data sets to analyse mRNA features. Almost 90% of known genes were covered by at least one read in all (PCR)-cDNA libraries. In contrast, direct RNA libraries only covered 70% (RNA001, replicate 3), 44% (RNA002, replicate 1) and 13% (RNA002, replicate 2) of the genes (Supplementary Fig. 10B). In order to evaluate how many reads are needed to cover at least 75% of all genes, we subsampled the reads of the representative rRNA-depleted PCR-cDNA sample (12 PCR cycles). We found that a sequencing depth of about 10-fold is sufficient for this purpose, corresponding to 70,000 mRNA-mapping reads (Supplementary Fig. 10C).

In order to evaluate if Nanopore RNA-seq data can be used for quantitative measurements, we looked into the correlation between replicates and when using different library types (Figure 2A,C, Supplementary Table 3). Despite different sequencing platforms, protocols for sample preparation and sequencing depths, we observed a decent correlation between expression data from published short-read Illumina RNA-seq data and ONT datasets (Figure 2B,C). Nevertheless, we found that a higher number of PCR cycles resulted in particularly GC-rich genes being underrepresented, leading to an overall more insufficient correlation in the PCR-cDNA datasets (Supplementary Fig. 11A,B). However, since we observed a similar

effect with the non-amplified direct cDNA sample, which overall showed the best correlation to the Illumina data, other biases cannot be ruled out completely. For example, the SMRT-Cap protocol includes stringent size-selection filtering for fragments bigger than 1 kb. Consequently, from a purely quantitative perspective, the SMRT-Cap data are not fully comparable to the Nanopore data, but this may also be partly due to the sequencing depth. Considering these interfering factors, we have obtained very good correlation between the cDNA replicates (0.97) and to the other library methods (DRS-Rep2 to cDNA-Rep2: 0.91; PCR-cDNA-Rep4 to cDNA-Rep2: 0.94).

Taken together, the ONT data are very consistent and allow a quantitative analysis of various transcriptomic features, which we will discuss in more detail below. However, the PCR bias is a critical point and researchers should carefully determine the number of PCR cycles required for their sample of choice.

Identification and trimming of full-length sequenced transcripts

To accurately quantify and identify the number of full-length sequenced reads, we used Pychopper (https://github.com/nanoporetech/pychopper), a tool developed by ONT. This tool allows the detection and trimming of full-length sequenced cDNA reads based on the presence of strand-switching primer (SSP) and anchored oligo(dT) VN primer (VNP). In addition, it orients the sequenced reads (Supplementary Fig. 12A). As already evident from the length distribution, 2D-like reads make up a significant portion of the direct cDNA samples, which is confirmed by the low percentage of the Pychopper-detected full-length sequenced transcripts of about 34% in contrast to over 80% of full-length sequenced reads in all PCR-cDNA samples using standard settings (Supplementary Fig. 12B). However, a direct-cDNA specific setting in Pychopper, which can handle 2D-like reads better and allows to rescue many reads, almost doubled the number of full-length sequenced reads detected (59%). When comparing the aligned read lengths, we detected only a minimal difference between untrimmed and full-length-filtered reads, which is probably caused by random mapping of adapter sequences (Supplementary Fig. 12C). Despite Pychopper trimming, we

observed that Adenine, caused by polyadenylation, and Guanine, caused by non-templated addition of nucleotides by the template-switching RT, are overrepresented at the 3´ and 5´ ends of cDNA reads, respectively (Supplementary Fig. 13A). To enable precise determination of transcript boundaries, we successfully trimmed off long poly(A) tails that are not expected to be found at the 3´ends of bacterial transcripts, and remaining SSP adapters from the 5´ends (Supplementary Fig. 13A,B).

Identification of transcript boundaries

Long-read ONT sequencing of RNA and cDNA molecules allows the simultaneous readout of 5´ and 3´ transcript boundaries (Figure 3A). Since full-length sequenced read starts and ends are expected to be enriched at functional relevant terminal positions and not randomly distributed, we first applied a peak calling algorithm on bedGraph files from terminal read positions that determines the positions of all local maxima. In the next step, comparable to the evaluation of SMRT-Cap or short-read datasets, we defined the highest accumulation of 5´ ends in a peak 300 bp upstream of an annotated gene as the primary 5´-transcript boundary (Figure 3A). Each additional peak in this region was designated as secondary and enriched intergenic peaks as internal. Because our samples do not only contain primary transcripts, we deliberately did not designate these ends as transcription start sites, although a considerable overlap is expected. We were able to define between 549 and 5,019 5´ transcript ends in representative data sets, which varied depending on sequencing depth and trimming (Supplementary Fig. 14A, Supplementary Table 4). As described in other studies, the majority of enriched 5´ ends are localized in internal regions. However, we could also identify up to 1,248 primary sites. Unexpectedly, untrimmed reads had a higher agreement in 5´ ends at the single-nucleotide level to other comparable methods such as short-read differential RNA-seq and SMRT-Cap than trimmed reads (Figure 3B, Supplementary Fig. 14B) (Thomason et al., 2015; Yan et al., 2018). Ends determined by direct RNA sequencing are about 12 nt shorter, which is in line with previous observations (Soneson et al., 2019; Workman et al., 2019; Parker et al., 2020) and can be rationalised by

a lack of control of the RNA translocation speed after the motor protein falls off the 5´ ends of the RNA (Figure 3B). PCR-cDNA and cDNA 5´ ends are very clearly defined and predominantly end at the same base. In contrast, DRS leads to fuzzy 5' ends, presumably caused by a lower mapping accuracy. TEX treatment had neither a positive nor negative effect on 5´ end detection or the number of reads starting at the enriched 5´ ends. This may be due to the short treatment time and the digestion of the remaining ribosomal RNA leaving mRNA-mapping primary transcripts unaffected. Primary 5´ transcript ends highly correlated between all different library types provided that enough reads support the enriched position (Supplementary Fig. 15). The moderate correlation (0.67) to SMRT-Cap 5´ ends can mainly be attributed to the different library preparation approaches (Supplementary Fig. 16A): In contrast to our data, only primary transcripts are specifically captured and subsequently small transcripts naturally occurring in *E. coli* are intentionally lost due to size selection. The correlation drastically improved, when considering the positions of secondary 5´ ends determined during ONT read analysis (Supplementary Fig. 16B). We found that for some genes, the SMRT-Cap primary site coincides with the ONT secondary, but not primary site. Although no specific enrichment for primary transcripts was performed for most of the samples, the 5´ UTR distributions and the bacterial-typical nucleotide contents of upstream regions lead to the assumption that ONT sequencing is capable of accurately determining transcription start sites (Supplementary Fig. 17).

Peak enrichment analysis and 3´ end annotations were performed as described for the 5´ ends (Figure 3A). Overall, the number of enriched 3´ ends found in the respective categories was slightly lower as compared to the 5´ ends (Supplementary Fig. 18A, Supplementary Table 5). In contrast to the rather detrimental effect of trimming on the accuracy of 5´ end detection, trimming increased the number of 3´ ends that are identical to Term-seq (Supplementary Fig. 18B) (Dar and Sorek, 2018). However, it should be noted that *in vitro* polyadenylation and trimming can affect detection accuracy of 3´ ends that have a naturally occurring terminal polyA sequence. This is for example the case with the RNAse E

processing site of the *ssrA* gene, where raw reads are artificially too long and contain additional adenines, which are later trimmed off during read processing (Supplementary Fig. 19) (Lin-Chao et al., 1999). Although 3' end detection is highly reproducible and 3´ ends overall highly correlate with SMRT-Cap detected ends, ONT 3´ ends are fuzzier and tend to be up to 3 nucleotides shorter (Figure 3C, Supplementary Fig. 20). Since we cannot exclude that 3´ to 5´ exoribonucleases degrade RNAs after transcription, enriched sites may either represent genuine termination sites or enriched processed 3´ends. Nevertheless, the 3´ UTR lengths of primary 3´ ends and the poly(T) termination motif, which is typical for intrinsic terminators, suggest that most detected primary 3´ ends are genuine transcription termination sites (Supplementary Fig. 21A,B).

Gene body coverage of long-read Nanopore reads

In contrast to DRS, the cDNA protocols provide access to full-length sequenced transcripts due to the template-switching behaviour of the RT (Supplementary Fig. 12). Accordingly, it is expected that the 5´ and 3´ ends are covered to the same extent and that the coverage distribution over a gen is overall flat, which should improve an accurate transcriptional unit analysis. However, previous studies have shown that both DRS and direct cDNA reads are often truncated at the 5´-end (Sessegolo et al., 2019; Soneson et al., 2019; Workman et al., 2019). The reasons for this observation are still not completely clear but could be related to the fact that RNAs are directly sequenced starting from the 3´ ends, to problems during template-switching, or sequencing-related issues like current spikes. To estimate the effect of the 3´-coverage bias, we looked at the gene body coverage profile between all samples and used previously introduced metrics, like the quartile coefficient of variation (QCoV) (Parker et al., 2020), to quantify coverage drops along the transcripts (Figure 4A,B). For the DRS and cDNA samples, we can confirm that 5´ ends are less covered compared to the 3´ ends (Figure 4A-E). This has a particularly dramatic effect on the 5´coverage of long transcriptional units, exemplarily shown for units ending at the *hslU* gene (Supplementary Figure 22). Overamplification during PCR results in both ends being more enriched

compared to the transcript centre. In contrast, at 12 cycles the reads are equally distributed across the gene body deviating on average less than 5% from the median coverage (Figure 4C). As expected, quality filtering and selection of full-length sequenced cDNA reads with both recognition adaptors results in an enrichment of the 5´ ends for all cDNA samples (Figure 4D). However, we see that transcripts longer than 2 kb are less well uniformly covered (Figure 4E). It should be noted that these do not occur very often in our selected data that rely on the previous annotation of 5´ and 3´ ends, which could influence the distribution.

Nanopore sequencing captures the complexity of bacterial transcriptional units

The distribution of reads over the gene body confirmed that ONT sequencing can cover both ends of a transcript. Since read lengths are theoretically only limited by the transcript size, ONT sequencing has the potential to accurately define complex transcriptional unit structures by finding overlaps between the mapping coordinates of individual reads and the transcript positions (Figure 5A). Following the annotation approach from the SMRT-Cap protocol, the unique combination of genes within a transcriptional unit was defined as the transcriptional context of a gene. Transcriptional unit prediction was performed exemplarily for one each of the DRS (RNA001 replicate 1), cDNA (DCS109 replicate 2) and PCR-cDNA (PCB109 replicate 4) libraries (Supplementary Table 6). Thereby, 788 (DRS), 2264 (cDNA) and 2433 (PCR-cDNA) unique transcriptional units were defined, respectively (Figure 5B). Mainly limited by the sequencing depth, the vast majority of defined transcriptional units (PCR-cDNA: 90%, cDNA: 83%, RNA: 90%) overlapped between the different protocols (Figure 5B). Hence, rare transcriptional unit variants stretching over multiple genes are not detected at low sequencing depth and stringent detection filters, which is also reflected in the mean number of genes encoded in a transcriptional unit: 1.14 for the DRS, 1.18 for the cDNA and 1.26 for the PCR-cDNA approach, respectively. This is in agreement with the observation that particularly long transcriptional units are underrepresented in our dataset. Therefore, the overall agreement with SMRT-Cap (43%) and the RegulonDB database (50%)

is only moderate, which is presumably additionally heavily influenced by the respective detection algorithms, library preparation and sample conditions (Supplementary Fig. 23). Nevertheless, the distribution of transcriptional contexts in the PCR-cDNA dataset is in good agreement with the results from the SMRT-Cap analysis, showing that many genes are transcribed in more than one context (Figure 5C).

Note that without prior enrichment or treatment, quantification of the individual transcriptional contexts should consider that prokaryotic transcripts are subject to various degradation and processing events: Therefore, it was not surprising that we captured a mix of 3´ or 5´ intact transcripts, which are often processed from the other end, as indicated by the ONT single-read tracks (Figure 5A, Supplementary Fig. 24,25). Effects that arise from RNA processing could be analysed in more detail when sequencing transcriptomes of exonuclease knock-out strains or with protocols that specifically enrich for primary transcripts (compare Send-seq and SMRT-Cap protocol) (Yan et al., 2018; Ju et al., 2019). However, after the explicit enrichment of full-length sequenced transcripts and under the valid assumption that transcripts are not strongly degraded (compare RIN values) the extensive transcriptional heterogeneity is surprising. This can not only be seen in Figure 5A, but also in other examples, such as the RegulonDB-annotated operon *rpsP-rimM-trmD-rplS* (Supplementary Fig. 22) or a section of the genome containing many ribosomal proteins (Supplementary Fig. 23). The annotation of transcriptional units fits very well with the prediction of primary 5´ and 3´ end and illustrates that long-read ONT RNA-seq can more easily identify transcripts that arise from a shared promoter and have heterogeneous 3´ ends. As already shown in the SMRT-Cap data, the *tff-rpsB-tsf* unit, which is identical to the operon annotated in the RegulonDB, is terminated in a stepwise manner. However, an additional termination site can be detected directly after the putative small RNA *tff*, which is otherwise lost through size selection (Figure 5A).

Taking advantage of the 5´ to 3´ connectivity of the reads is one of the key advantages of single-molecule sequencing, which we used to perform transcriptional unit prediction.

However, this feature can also be used to explore transcription, processing and degradation patterns of individual transcripts. We exemplify this capacity using the well-described decay of the *rpsO* mRNA, encoding the ribosomal protein S15 (Supplementary Fig. 26a) (Régnier and Portier, 1986; Régnier and Hajnsdorf, 1991; Hajnsdorf and Régnier, 1999). Nanopore (PCR)-cDNA sequencing captures that the majority of the transcripts are derived from promoter P1 and end at the 3´hairpin (PCR-cDNA: 44%, cDNA: 53%) protecting the primary transcript from degradation. Consequently, this represents the most abundant transcript (PCR-cDNA data shown in Supplementary Fig. 26b,c,d). Additionally, frequent degradation events from the 3´end after processing at M2 and minor populations (e.g. transcript cleavage at M3, transcription from a second upstream promoter or termination readthrough) can be observed.

In summary, Nanopore sequencing is capable of not only accurately detecting complex transcriptional unit structures but can also aid in quantification or in deciphering the unprecedented transcriptional heterogeneity, which may be improved by using specialized strains or conditions depending on the scientific question.

DISCUSSION

In this study, we performed a comprehensive comparison of all currently available kits from Oxford Nanopore for the analysis of RNAs, including direct sequencing of native RNA (RNA001, RNA002), direct cDNA (DCS109) and PCR-cDNA sequencing (PCB109), in the bacterial model organism *Escherichia coli* K-12. As a result, we demonstrate that multiple properties of the transcriptome can be examined simultaneously with high accuracy. This study therefore provides the first extensive analysis of ONT RNA-seq methods in prokaryotes. Furthermore, after screening important quality control metrics of the sequenced libraries, we show that Nanopore RNA-seq is suitable for making quantitative measurements and correlates well with data of the most commonly used short-read Illumina RNA-seq data. Additionally, we provide a bioinformatics workflow that allows accurate determination of transcript boundaries and quantitative analysis of transcriptional units applicable to all prokaryotes.

However, at present, some disadvantages of Nanopore RNA-seq should be considered that are summarised in Figure 6A: First, it must be ensured that the polyadenylation reaction in the organism of choice works equally effectively for all RNAs. Second, direct sequencing of RNAs requires a large amount of starting RNA material (> 10 µg) to yield enough mRNA (500 ng) left after effective rRNA depletion. Since the depletion kits are usually not designed for these quantities, the additional reactions are another cost factor. Higher costs for DRS also originate from the slower sequencing speed, which negatively impacts throughput and the current lack of a barcoding options provided by ONT. Although there are already excellent options to build a custom set of DRS barcodes, this is not as straightforward to use as for (PCR)-cDNA libraries (Smith et al., 2020). Regarding 5´ end detection, it has been shown multiple times that about 12 bases are missing from the DRS 5´ ends. This observation can be explained by the motor protein falling off the end of a transcript resulting in a loss of control to guide the RNA through the nanopore, which is not the case for the (PCR)-cDNA data (Soneson et al., 2019; Workman et al., 2019). Another point of criticism

that is repeatedly discussed is the comparatively low accuracy, especially for DRS, but also for (PCR)-cDNA datasets (Garalde et al., 2018; Soneson et al., 2019; Workman et al., 2019). Although this is not a significant problem for most questions, it affected the base-accurate trimming of adapter sequences and thus influenced the accuracy of the determination of the transcript ends. In particular, up to four more bases are trimmed off at the 3´ ends since the homo-poly(A) sequence is usually low in quality and can only be trimmed inaccurately. Determining the 3´ ends without trimming, which performs better at the 5´ ends, performed even worse since long-read Nanopore mappers like minimap2 allow a higher number of errors (Li, 2018). In general, the choice of the mapping tool should be well considered, as it greatly impacts the quality of the analysis. We applied the widely used and actively developed minimap2, which fails to align small RNAs (~80 bases cutoff) (Li, 2018). While other mapping tools, like Magic-BLAST (Boratyn et al., 2019) or GraphMap2 (Sović et al., 2016) can align short transcripts, it is usually at the expense of other aspects, and the method of choice dependent on the respective question. Despite or even because of these limitations, the Nanopore community is very active and interested in providing solutions for the problems discussed. Indeed, there are already promising applications that will also further improve ONT RNA-seq in prokaryotes in the future, like the error-correction of (PCR)-cDNA reads using isONcorrect (Sahlin et al., 2021) or the improvement of 5´ end detection in DRS after 5´ dependent adapter ligation (Parker et al., 2020).

Based on our results and considering the most cost-effective way to create and sequence libraries, we conclude that (PCR)-cDNA sequencing is the method of choice for most scientific questions, except for the analysis of RNA modifications (Begik et al., 2021). As only 1 ng of rRNA-depleted RNA is sufficient to generate PCR-cDNA libraries, PCR-cDNA-seq is highly preferable for organisms or conditions where the amount of RNA isolated is a crucial criterion. Our data clearly show that the number of cycles in the PCR should be controlled with special care. Otherwise, small AT-rich transcripts are preferentially amplified and sequenced, which distorts the quantification and further analyses. However, if this is handled

correctly and the number of cycles is as low as possible, in our case 12, the PCR-cDNA data are highly comparable to the direct cDNA results. In any case, reverse transcription is a critical point for all (PCR)-cDNA libraries. Nevertheless, the ONT-recommended Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific) performed quite well for our samples as documented in the gene body coverage data and the reproducibly good quantification. Another advantage of the enzyme used is that the reaction temperature can be increased to transcribe sequences with exceptionally high GC content or secondary structures.

In fact, there are already some sophisticated ways to profile full-length transcripts in *E. coli*, including the SMRT-Cappable-seq (Yan et al., 2018) and the SEnd-Seq (Ju et al., 2019) protocols. Comparison of SMRT-Cap and ONT data show that both datasets are highly congruent, although the repeatedly discussed size selection in the PacBio libraries plays a critical role and is a disadvantage. Unfortunately, despite the introduction of these methods, they have not yet been used in the prokaryotic community for further studies, although the reasons for this may well be diverse. However, we can imagine that the low initial costs of purchasing a MinION and the excellent performance could encourage some laboratories to use Nanopore RNA-seq in prokaryotes. The additional costs and IT infrastructure requirements are also limited, with basecalling of the data representing the highest computational effort for these analyses.

Taken together, a key advantage of ONT RNA-seq is that multiple features can be addressed simultaneously with high accuracy (Figure 6B). This versatility distinguishes the technique from the various RNA-seq technologies designed to tackle only one specific question or biochemical assays. Furthermore, since Nanopore sequencing is a *bona fide* single-molecule method, molecular heterogeneity at the transcriptome level can be analysed. Additionally, even minor RNA populations can be detected that are inevitably lost in ensemble sequencing approaches. However, we observed a complex transcription pattern with multiple possible RNA variants. Given that transcription and translation are coupled in *E.*

coli, new questions about the translation efficiency and transcript stability of the transcript variants emerge (Proshkin et al., 2010; Wang et al., 2020; Webster et al., 2020; Irastortza-Olaziregi and Amster-Choder, 2021). Furthermore, high-quality long-read RNA-seq data can be used to analyse degradation or processing patterns to gain new insights into mRNA decay in prokaryotes. With this study, we not only show the applicability of ONT RNA-seq in prokaryotes, but also provide representative long-read transcriptome data from *E. coli* and a robust bioinformatical workflow to the community that can be used to tackle various questions.

MATERIAL AND METHODS

Cell growth and RNA extraction

Escherichia coli K-12 MG1655 cells were grown in rich medium (10 g tryptone, 5 g yeast extract, 5 g NaCl per litre, pH 7.2) to an OD_{600nm} of 0.5-0.6. To stabilize RNAs, two volumes of RNAlater (Thermo Fisher Scientific) were immediately added to the cultures and stored at -20°C until cells were harvested by centrifugation at 4°C.

Total RNA of all samples except RNA001 was extracted using RNeasy Mini Kit (Qiagen) according to the manufacturer´s instructions. RNA001 RNA was purified using the Monarch® Total RNA Miniprep Kit (New England Biolabs). The integrity of total RNA from *E. coli* was assessed via a Bioanalyzer (Agilent) run using the RNA 6000 Pico Kit (Agilent), and only RNAs with RNA integrity numbers (RIN) above 9.5 were used for subsequent treatments and sequencing. In short, the RIN value, calculated on a scale from 0 to 10, has evolved as a standard to estimate integrity of RNA samples from the size distribution and is calculated by an algorithm that is based on the combination of different features, like 16S and 23S rRNA areas (Schroeder et al., 2006).

Poly(A) tailing, rRNA depletion and additional RNA treatment

Next, RNAs were heat incubated at 70°C for 2 min and snap cooled on a pre-chilled freezer block before polyadenylating RNAs using the *E. coli* poly(A) polymerase (New England

Biolabs). Briefly, 5 µg RNA, 20 units poly(A) polymerase, 5 µl reaction buffer and 1 mM ATP were incubated for 15 min at 37°C in a total reaction volume of 50 µl. Note that the identical reaction conditions were chosen here as described in the SMRT-Cap protocol that resulted in successfull and efficient poly(A)-tailing (Yan et al., 2018). To stop and clean up the reaction, poly(A)-tailed RNAs were purified following the RNeasy Micro clean-up protocol (Qiagen), which was used for all subsequent RNA clean-ups. The efficiency of poly(A)-tailing was evaluated via a Bioanalyzer run. Ribosomal RNA (rRNA) depletion was performed using the Pan-Prokaryote riboPOOL by siTOOLs, which effectively removes rRNAs from *E. coli*. For TEX-treated samples, partial digestion of RNAs that are not 5´-triphosphorylated (e.g. tRNAs, rRNAs) was achieved by incubation of the RNA with a Terminator 5´-Phosphate-Dependent Exonuclease (TEX, Lucigen). Therefore, 10 µg of RNA used in the RNA001 sample, were incubated with 1 unit TEX, 2 µl TEX reaction buffer and 0.5 µl RiboGuard RNase Inhibitor (Lucigen) in a total volume of 20 µl for 60 minutes at 30°C. Besides, 20 ng of rRNA-depleted samples subsequently used in the PCR-cDNA workflow (replicate 4 and 5), were only partially TEX-treated using the same enzyme and buffer concentrations but reducing the reaction time to 15 minutes. All reactions were terminated by adding EDTA and cleaned up following the RNeasy Micro clean-up protocol. Before library preparation, the extent of the remaining buffer and DNA contamination were tested by performing standard spectroscopic measurements (Nanodrop One) and using the Qubit 1X dsDNA HS assay kit (Thermo Fisher Scientific). Input RNAs were finally quantified using the Qubit RNA HS assay kit.

Library preparation and sequencing

Libraries for Nanopore sequencing were prepared from poly(A)-tailed RNAs according to protocols provided by Oxford Nanopore (Oxford Nanopore Technologies Ltd, Oxford, UK) for direct sequencing of native RNAs (SQK-RNA001, SQK-RNA002), direct cDNA native barcoding (SQK-DCS109 with EXP-NBD104) and PCR-cDNA barcoding (SQK-PCB109) with the following minor modifications: Agencourt AMPure XP magnetic beads (Beckman

Coulter) in combination with 1 µl of RiboGuard RNase Inhibitor (Lucigen) were used instead of the recommended Agencourt RNAclean XP beads to clean up samples. For reverse transcription, Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific) was used for all cDNA samples and for the RNA002 samples (SuperScript III Reverse Transcriptase from Thermo Fisher Scientific used for RNA001 sample). The amount of input RNA, barcoding strategy, number of PCR cycles and extension times can be found in Supplementary Table 1 and are also summarized in part in Figure 1A.

Nanopore libraries were sequenced using either a MinION Mk1B connected to a laptop with the recommended specifications for Nanopore sequencing or a Mk1C. All samples were sequenced on R9.4 flow cells and the recommended scripts in MinKNOW to generate fast5 files with live-basecalling enabled. In case of an observed drop in translocation speed and subsequent reduced read quality, the flow cells were refueled with flush buffer, as recommended by ONT. Flow cells were subsequently washed and re-used for further runs, provided a sufficient number of active pores left. To avoid cross-contamination of reads, a different set of barcodes was used for the next run. Also, the starting voltage of re-used flow cells was adjusted for the next run to account for the voltage drift during a sequencing run.

Data analysis

Basecalling, demultiplexing of raw reads and quality control of raw reads

All fast5 reads were re-basecalled using guppy (ont-guppy-for-mk1c v4.3.4) in high-accuracy mode (rna_r9.4.1_70bps_hac.cfg, dna_r9.4.1_450bps_hac.cfg) without quality filtering. While standard parameters were used for basecalling fast5s from cDNA sequencing, fast5 files from RNA sequencing were basecalled with RNA-specific parameters (--calib_detect, --reverse_sequence and --u_substitution). Next, basecalled fastq files from cDNA runs were demultiplexed in a separate step by the guppy suite command guppy_barcoder using default parameters and the respective barcoding kit. After that, relevant information from the guppy sequencing and barcode summary files were extracted to analyse the properties of raw

reads (Supplementary Table 1). Please note that in Supplementary Table 2, all figures created from numerical data are referenced and linked to the corresponding code in the Github repository https://github.com/felixgrunberger/microbepore.

Read alignment

Files were mapped to the reference genome from *Escherichia coli* K-12 MG1655 (GenBank: U00096.3) (Riley et al., 2006), using minimap2 (Release 2.18-r1015, https://github.com/lh3/minimap2) (Li, 2018). Output alignments in the SAM format were generated with -ax splice -k14 for Nanopore 2D cDNA-seq and -ax splice, -uf, -k14 for direct RNA-seq with i) -p set to 0.99, to return primary and secondary mappings and ii) with --MD turned on, to include the MD tag for calculating mapping identities. Alignment files were further converted to bam files, sorted and indexed using SAMtools (Li et al., 2009). To evaluate the alignments, we first calculated the aligned read length by adding the number of M(atch) and I(nsertion) characters in the CIGAR string (Soneson et al., 2019). Based on this, the mapping identity was defined as (1-NM/aligned_reads)*100, where NM is the edit distance reported taken from minimap2. Read basecalling and mapping metrics can be found in Supplementary Table 1. To analyse single reads in more detail with respect to the RNA type (mRNA, rRNA, other ncRNA, unspecified) they map to, bam files were first converted back to FASTQ using bedtools v2.29.2 (Quinlan and Hall, 2010). Next FASTQ files were remapped to a transcriptome file using minimap2 with the previously mentioned parameters to assign single read names with feature IDs. To handle multi-mapping reads, only the mapping location with i) the highest overall identity or if identical ii) the position with most aligned bases was kept for every read id.

Gene abundance estimation

A publicly available short-read Illumina dataset (SRR1927169) obtained from RNA-seq data of *E. coli* K-12 grown under rich conditions was downloaded from Gene Expression Omnibus (GEO) GSE67218. Reads were first quality trimmed using trimmomatic v0.39 (Bolger et al.,

2014) (leading:20, trailing:20, slidingwindow:4:20, minlen:12) and mapped to the reference genome using bowtie2 (-N 0, -L 26) (Langmead and Salzberg, 2012).

SMRT-Cap data obtained from sequencing data from rich-medium samples (SRR7533626, SRR7533627) were downloaded from GEO GSE117273 (Yan et al., 2018). PacBio reads were processed as described in the SMRT-Cap protocol using the pacbio_trim.py script downloaded from https://github.com/elitaone/SMRT-cappable-seq. In short, reads were filtered and trimmed using the respective filter and poly functions. Next, reads were mapped to the *E. coli* K-12 genome using minimap2 with PacBio-specific (-ax map-pb) options (Li, 2018). Bam files from Illumina and SMRT-Cap sequencing were converted to FASTQ format and remapped to the gene file as described before.

To estimate gene abundances from ONT, short-read Illumina and SMRT-Cap libraries, Salmon (v.1.4.0) was applied in alignment-based mode (Patro et al., 2017). Transcripts per million (TPM) were re-calculated using the salmon-computed effective transcript length, after dropping reads mapping to rRNAs, that are variable between non-depleted and depleted RNA sets.

Identification and trimming of full-length sequenced transcripts

Full-length cDNA reads containing strand-switching primer (SSP) and anchored oligo(dT) VN primer (VNP) in the correct orientation were identified using pychopper (v.2.5.0) with standard parameters using the default pHMM backend and autotuned cutoff parameters estimated from subsampled data (https://github.com/nanoporetech/pychopper). After a first round, a second round of pychopper was applied to the unclassified direct cDNA reads with DCS-specific read rescue enabled. Reads from rescued and full-length folders were merged and used for subsequent steps. To evaluate the influence of different trimming approaches on the accuracy of transcript boundary analysis, we applied additional 5´ and 3´ trimming steps using cutadapt v3.2 (Martin, 2011). To this end, polyA sequences were removed from the 3´ends (-a A{10}, -e 1, -j 0) and remaining SSP sequences were removed from the

5´ends (-g TTTCTGTTGGTGCTGATATTGCTGGG, -e 1, -j 0) of direct RNA and full-length sequenced cDNA reads. Finally, trimmed reads were mapped using minimap2 as described before. Reads with more than 10 clipped bases on either side were removed from the alignments using samclip (v.0.4.0, https://github.com/tseemann/samclip).

To assess the impact of trimmings on gene body coverage, a coverage meta-analysis was performed. First, a transcript file was created for all genes with an ONT-annotated primary 5´ and 3´ end (see next section). Based on this, strand-specific coverage files were created from the bam files and coverage analysis performed using a custom R script. The genomic coordinates and the counted reads per position were first scaled to values between 0 and 100 and the mean coverage distribution per normalised position was calculated. To evaluate the coverage profiles and the decay at the 5´ or 3´ ends, we calculated the quartile coefficient of variation (interquartile range/median) (Parker et al., 2020) and additionally compared the mean coverage in the first and last 10% of the positions to the median values.

Detection of transcript boundaries

The determination of enriched 5´and 3´ends was carried out in the same way, but independently of each other, and is briefly explained in the following: First, strand-specific read ends in bedgraph format were created from bam files using bedtools genomecov (-5 or -3 option, -bga) (Quinlan and Hall, 2010). Next, the previously published Termseq_peaks script (Adams et al., 2021) was used to call peaks for each sample individually without including replicates (https://github.com/NICHD-BSPC/termseq-peaks). This script is based on *scipy.signal.find_peaks*, which is running in the background of Termseq_peaks with lenient parameters (prominence=(None,None), width=(1,None), rel_height=0.75). However, we deliberately used Termseq_peaks since its ability to include replicates by applying an Irreproducible Discovery Rate method which can be applied to future studies. For end detection, only the leniently called peaks in the narrowPeak file were used after adding the number of counts for each position using bedtools intersect. Enriched positions were finally filtered and annotated based on the following criteria: i) For each peak the position with the

highest number of reads was selected. ii) Positions within 20 bases were merged and only the position with the highest number of reads retained. iii) Positions with less than three reads were filtered out. iv) Positions were assigned based on their relative orientation to a gene and their respective peak height as primary (depending on 5´ or 3´ detection: highest peak within 300 bases upstream or downstream of a gene, respectively), secondary (each additional peak 300 bases up/downstream of a gene) and internal (each peak in the coding range).

Reproducibility and comparability of primary 5´ and 3´ ends were evaluated based on Pearson coefficients calculated from pairwise complete observations. Additionally, 5´ and 3´ untranslated regions (UTR) were calculated based on the distance of the enriched primary site to the start or end of a coding region, respectively. The positions of primary sites called from direct RNA-seq data were corrected by 12 bases.

Detection and quantification of transcriptional units

Tables containing each read as a single row were created from the bam files using the R package Genomic alignments (Lawrence et al., 2013). Reads that mapped to the opposite strand of an annotated mRNA or ncRNA or that mapped to widely separated genomic positions were discarded. Next, all range overlaps sharing more than 100 bases were defined between the read table and the genomic feature table using the findOverlaps function from the GenomicRanges package. This way, multiple features can be assigned to each individual read. If their genomic positions are adjacent, the combination of features covered by a coverage-dependent number of reads (10 reads for PCR-cDNA replicate 4) are considered as a transcriptional unit. To enable a quantitative assessment of the transcriptional units and the respective context, the number of reads is first determined for each feature individually and then compared with the number of reads in each detected unit. We compared the transcriptional units with the operon tables from the RegulonDB database (Santos-Zavaleta et al., 2019) and the SMRT-Cappable-seq study (Yan et al., 2018).

Public data

In addition to the publicly available results from the SMRT-Cappable-seq study (Yan et al., 2018), the short-read Illumina data for gene expression comparison and the RegulonDB (Santos-Zavaleta et al., 2019) mentioned above, we also compared ONT RNA-seq 5´ends with the results of a differential RNA-seq study (Thomason et al., 2015) and 3´ ends with Term-seq results (Dar and Sorek, 2018).

AVAILABILITY

Data availability

To facilitate easier access basecalled and demultiplexed FASTQ, mapped BAM files from untrimmed reads and large read summary files are publicly available from https://zenodo.org/record/4879174#.YLSkjy221pQ.

Code availability

All scripts and code used in this work are available on GitHub (https://github.com/felixgrunberger/microbepore). Additionally, a more detailed documentation can be found at https://felixgrunberger.github.io/microbepore.

ACCESSION NUMBERS

Sequencing files in original FAST5 format are publicly available in the Sequence Read Archive SRA (RNA001: PRJNA632538, all other datasets: PRJNA731531).

ACKNOWLEDGEMENT

We thank all the members of the Ferreira-Cerca lab and of the Grohmann lab, especially Prof. Dr. Winfried Hausner and Martin Fenk for fruitful discussions.

FUNDING

This work was supported by the Deutsche Forschungsgemeinschaft [SFB960 TPA7 to D.G., and SFB960 TPB13 to S.F.-C..]; Funding for open access charge: Deutsche Forschungsgemeinschaft.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

Adams, P. P., Baniulyte, G., Esnault, C., Chegireddy, K., Singh, N., Monge, M., et al. (2021). Regulatory roles of escherichia coli 5' utr and orf-internal rnas detected by 3' end mapping. *Elife* 10, 1–33. doi:10.7554/eLife.62438.

Begik, O., Lucas, M. C., Pryszcz, L. P., Ramirez, J. M., Medina, R., Milenkovic, I., et al. (2021). Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nat. Biotechnol.* doi:10.1038/s41587-021-00915-6.

Boldogkői, Z., Moldován, N., Balázs, Z., Snyder, M., and Tombácz, D. (2019). Long-Read Sequencing – A Powerful Tool in Viral Transcriptome Research. *Trends Microbiol.* 27, 578–592. doi:10.1016/j.tim.2019.01.010.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. doi:10.1093/bioinformatics/btu170.

Boratyn, G. M., Thierry-Mieg, J., Thierry-Mieg, D., Busby, B., and Madden, T. L. (2019). Magic-BLAST, an accurate RNA-seq aligner for long and short reads. *BMC Bioinformatics* 20, 405. doi:10.1186/s12859-019-2996-x.

Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., et al. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* 8, 16027. doi:10.1038/ncomms16027.

Byrne, A., Cole, C., Volden, R., and Vollmers, C. (2019). Realizing the potential of full-length transcriptome sequencing. *Philos. Trans. R. Soc. B Biol. Sci.* 374, 20190097. doi:10.1098/rstb.2019.0097.

Choi, S. C. (2016). On the study of microbial transcriptomes using second- and third-generation sequencing technologies. *J. Microbiol.* 54, 527–536. doi:10.1007/s12275-016-6233-2.

Croucher, N. J., and Thomson, N. R. (2010). Studying bacterial transcriptomes using RNA-seq. *Curr. Opin. Microbiol.* 13, 619–624. doi:10.1016/j.mib.2010.09.009.

Dar, D., and Sorek, R. (2018). High-resolution RNA 3-ends mapping of bacterial Rho-dependent transcripts. *Nucleic Acids Res.* 46, 6797–6805. doi:10.1093/nar/gky274.

Dong, X., Tian, L., Gouil, Q., Kariyawasam, H., Su, S., De Paoli-Iseppi, R., et al. (2021). The long and the short of it: unlocking nanopore long-read RNA sequencing data with short-read differential expression analysis tools. *NAR Genomics Bioinforma.* 3, 1–11. doi:10.1093/nargab/lqab028.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science (80-.).* 323, 133–138. doi:10.1126/science.1162986.

Escobar-Zepeda, A., Vera-Ponce de León, A., and Sanchez-Flores, A. (2015). The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. *Front. Genet.* 6. doi:10.3389/fgene.2015.00348.

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., et al. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* 15, 201–206. doi:10.1038/nmeth.4577.

Hajnsdorf, E., and Régnier, P. (1999). E. coli rpsO mRNA decay: RNase E processing at the beginning of the coding sequence stimulates Poly(A)-dependent degradation of the mRNA. *J. Mol. Biol.* 286, 1033–1043. doi:10.1006/jmbi.1999.2547.

Himeno, H., Kurita, D., and Muto, A. (2014). TmRNA-mediated trans-translation as the major ribosome rescue system in a bacterial cell. *Front. Genet.* 5, 1–13. doi:10.3389/fgene.2014.00066.

Hör, J., Gorski, S. A., and Vogel, J. (2018). Bacterial RNA Biology on a Genome Scale. *Mol. Cell* 70, 785–799. doi:10.1016/j.molcel.2017.12.023.

Irastortza-Olaziregi, M., and Amster-Choder, O. (2021). Coupled Transcription-Translation in Prokaryotes: An Old Couple With New Surprises. *Front. Microbiol.* 11. doi:10.3389/fmicb.2020.624830.

Jenjaroenpun, P., Wongsurawat, T., Wadley, T. D., Wassenaar, T. M., Liu, J., Dai, Q., et al. (2021). Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic Acids Res.* 49, e7. doi:10.1093/nar/gkaa620.

Ju, X., Li, D., and Liu, S. (2019). Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria. *Nat. Microbiol.* 4, 1907–1918. doi:10.1038/s41564-019-0500-z.

Keller, M. W., Rambo-Martin, B. L., Wilson, M. M., Ridenour, C. A., Shepard, S. S., Stark, T. J., et al. (2018). Direct RNA Sequencing of the Coding Complete Influenza A Virus Genome. *Sci. Rep.* 8, 14408. doi:10.1038/s41598-018-32615-8.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi:10.1038/nmeth.1923.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., et al. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.* 9, 1–10. doi:10.1371/journal.pcbi.1003118.

Levy, S. E., and Myers, R. M. (2016). Advancements in Next-Generation Sequencing. *Annu. Rev. Genomics Hum. Genet.* 17, 95–115. doi:10.1146/annurev-genom-083115-022413.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi:10.1093/bioinformatics/bty191.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352.

Lin-Chao, S., Wei, C.-L., and Lin, Y.-T. (1999). RNase E is required for the maturation of

ssrA RNA and normal ssrA RNA peptide-tagging activity. *Proc. Natl. Acad. Sci.* 96, 12406–12411. doi:10.1073/pnas.96.22.12406.

Liu, H., Begik, O., Lucas, M. C., Ramirez, J. M., Mason, C. E., Wiener, D., et al. (2019). Accurate detection of m6A RNA modifications in native RNA sequences. *Nat. Commun.*, 1–9. doi:10.1038/s41467-019-11713-9.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17, 10. doi:10.14806/ej.17.1.200.

Matz, M., Shagin, D., Bogdanova, E., Britanova, O., Lukyanov, S., Diatchenko, L., et al. (1999). Amplification of cDNA ends based on template-switching effect and step-out PCR. *Nucleic Acids Res.* 27, 1558–1560. doi:10.1093/nar/27.6.1558.

Mikheyev, A. S., and Tin, M. M. Y. (2014). A first look at the Oxford Nanopore MinION sequencer. *Mol. Ecol. Resour.* 14, 1097–1102. doi:10.1111/1755-0998.12324.

Nowrousian, M. (2010). Next-Generation Sequencing Techniques for Eukaryotic Microorganisms: Sequencing-Based Solutions to Biological Problems. *Eukaryot. Cell* 9, 1300–1310. doi:10.1128/EC.00123-10.

Parker, M. T., Knop, K., Sherwood, A. V, Schurch, N. J., Mackinnon, K., Gould, P. D., et al. (2020). Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. *Elife* 9. doi:10.7554/eLife.49658.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14, 417–419. doi:10.1038/nmeth.4197.

Perocchi, F., Xu, Z., Clauder-Münster, S., and Steinmetz, L. M. (2007). Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. *Nucleic Acids Res.* 35. doi:10.1093/nar/gkm683.

Pitt, M. E., Nguyen, S. H., Duarte, T. P. S., Teng, H., Blaskovich, M. A. T., Cooper, M. A., et

al. (2020). Evaluating the genome and resistome of extensively drug-resistant Klebsiella pneumoniae using native DNA and RNA Nanopore sequencing. *Gigascience* 9, 1–14. doi:10.1093/GIGASCIENCE/GIAA002.

Proshkin, S., Rachid Rahmouni, A., Mironov, A., Nudler, E., Rahmouni, A. R., Mironov, A., et al. (2010). Cooperation Between Translating Ribosomes and RNA Polymerase in Transcription Elongation. *Science (80-.).* 328, 504–508. doi:10.1126/science.1184939.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi:10.1093/bioinformatics/btq033.

Régnier, P., and Hajnsdorf, E. (1991). Decay of mRNA encoding ribosomal protein S15 of Escherichia coli is initiated by an RNase E-dependent endonucleolytic cleavage that removes the 3′ stabilizing stem and loop structure. *J. Mol. Biol.* 217, 283–292. doi:10.1016/0022-2836(91)90542-E.

Régnier, P., and Portier, C. (1986). Initiation, attenuation and RNase III processing of transcripts from the Escherichia coli operon encoding ribosomal protein S15 and polynucleotide phosphorylase. *J. Mol. Biol.* 187, 23–32. doi:10.1016/0022-2836(86)90403-1.

Riley, M., Abe, T., Arnaud, M. B., Berlyn, M. K. B., Blattner, F. R., Chaudhuri, R. R., et al. (2006). Escherichia coli K-12: a cooperatively developed annotation snapshot--2005. *Nucleic Acids Res.* 34, 1–9. doi:10.1093/nar/gkj405.

Sahlin, K., Sipos, B., James, P. L., and Medvedev, P. (2021). Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. *Nat. Commun.* 12, 1–13. doi:10.1038/s41467-020-20340-8.

Saliba, A.-E., C Santos, S., and Vogel, J. (2017). New RNA-seq approaches for the study of bacterial pathogens. *Curr. Opin. Microbiol.* 35, 78–87. doi:10.1016/j.mib.2017.01.001.

Santos-Zavaleta, A., Salgado, H., Gama-Castro, S., Sánchez-Pérez, M., Gómez-Romero, L.,

Ledezma-Tejeida, D., et al. (2019). RegulonDB v 10.5: Tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. *Nucleic Acids Res.* 47, D212–D220. doi:10.1093/nar/gky1077.

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., et al. (2006). The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.* 7, 3. doi:10.1186/1471-2199-7-3.

Seki, M., Oka, M., Xu, L., Suzuki, A., and Suzuki, Y. (2021). “Transcript Identification Through Long-Read Sequencing,” in, 531–541. doi:10.1007/978-1-0716-1307-8_29.

Sessegolo, C., Cruaud, C., Da Silva, C., Cologne, A., Dubarry, M., Derrien, T., et al. (2019). Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules. *Sci. Rep.* 9, 1–12. doi:10.1038/s41598-019-51470-9.

Smith, A. M., Jain, M., Mulroney, L., Garalde, D. R., and Akeson, M. (2019). Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing. *PLoS One* 14, e0216709. doi:10.1371/journal.pone.0216709.

Smith, M. A., Ersavas, T., Ferguson, J. M., Liu, H., Lucas, M. C., Begik, O., et al. (2020). Molecular barcoding of native RNAs using nanopore sequencing and deep learning. *Genome Res.* 30, 1345–1353. doi:10.1101/GR.260836.120.

Soneson, C., Yao, Y., Bratus-neuenschwander, A., Patrignani, A., Robinson, M. D., and Hussain, S. (2019). A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat. Commun.* 10, 1–14. doi:10.1038/s41467-019-11272-z.

Sović, I., Šikić, M., Wilm, A., Fenlon, S. N., Chen, S., and Nagarajan, N. (2016). Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat. Commun.* 7. doi:10.1038/ncomms11307.

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. *Nat.*

Rev. Genet. doi:10.1038/s41576-019-0150-2.

Thomason, M. K., Bischler, T., Eisenbart, S. K., Förstner, K. U., Zhang, A., Herbig, A., et al. (2015). Global Transcriptional Start Site Mapping Using Differential RNA Sequencing Reveals Novel Antisense RNAs in Escherichia coli. *J. Bacteriol.* 197, 18–28. doi:10.1128/JB.02096-14.

Tilgner, H., Jahanbani, F., Blauwkamp, T., Moshrefi, A., Jaeger, E., Chen, F., et al. (2015). Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* 33, 736–742. doi:10.1038/nbt.3242.

Tombácz, D., Moldován, N., Balázs, Z., Gulyás, G., Csabai, Z., Boldogkői, M., et al. (2019). Multiple Long-Read Sequencing Survey of Herpes Simplex Virus Dynamic Transcriptome. *Front. Genet.* 10. doi:10.3389/fgene.2019.00834.

Tuiskunen, A., Leparc-Goffart, I., Boubis, L., Monteil, V., Klingström, J., Tolou, H. J., et al. (2010). Self-priming of reverse transcriptase impairs strand-specific detection of dengue virus RNA. *J. Gen. Virol.* 91, 1019–1027. doi:10.1099/vir.0.016667-0.

Viehweger, A., Krautwurst, S., Lamkiewicz, K., Madhugiri, R., Ziebuhr, J., Hölzer, M., et al. (2019). Direct RNA nanopore sequencing of full-length coron-avirus genomes provides novel insights into structural variants and enables modification analysis. *Genome Res.*, 483693. doi:10.1101/483693.

Vilfan, I. D., Tsai, Y.-C., Clark, T. A., Wegener, J., Dai, Q., Yi, C., et al. (2013). Analysis of RNA base modification and structural rearrangement by single-molecule real-time detection of reverse transcription. *J. Nanobiotechnology* 11, 8. doi:10.1186/1477-3155-11-8.

Wang, C., Molodtsov, V., Firlar, E., Kaelber, J. T., Blaha, G., Su, M., et al. (2020). Structural basis of transcription-translation coupling. *Science (80-.).*, eabb5317.

doi:10.1126/science.abb5317.

Wang, D., Jiang, A., Feng, J., Li, G., Guo, D., Sajid, M., et al. (2021). The SARS-CoV-2 subgenome landscape and its novel regulatory features. *Mol. Cell*, 1–13. doi:10.1016/j.molcel.2021.02.036.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi:10.1038/nrg2484.

Webster, M. W., Takacs, M., Zhu, C., Vidmar, V., Eduljee, A., Abdelkareem, M., et al. (2020). Structural basis of transcription-translation coupling and collision in bacteria. *Science (80-.).*, eabb5036. doi:10.1126/science.abb5036.

Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Razaghi, R., et al. (2019). Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* 16, 1297–1305. doi:10.1038/s41592-019-0617-2.

Yan, B., Boitano, M., Clark, T. A., and Ettwiller, L. (2018). SMRT-Cappable-seq reveals complex operon variants in bacteria. *Nat. Commun.* 9, 3676. doi:10.1038/s41467-018-05997-6.

Yang, M., Cousineau, A., Liu, X., Luo, Y., Sun, D., Li, S., et al. (2020). Direct Metatranscriptome RNA-seq and Multiplex RT-PCR Amplicon Sequencing on Nanopore MinION – Promising Strategies for Multiplex Identification of Viable Pathogens in Food. *Front. Microbiol.* 11, 1–14. doi:10.3389/fmicb.2020.00514.

Zhao, L., Zhang, H., Kohnen, M. V., Prasad, K. V. S. K., Gu, L., and Reddy, A. S. N. (2019). Analysis of Transcriptome and Epitranscriptome in Plants Using PacBio Iso-Seq and Nanopore-Based Direct RNA Sequencing. *Front. Genet.* 10. doi:10.3389/fgene.2019.00253.

TABLE AND FIGURES LEGENDS

Figures

Figure 1. Overview of generated datasets for comprehensively comparing Nanopore sequencing of RNA and cDNA molecules in *Escherichia coli.* A, Five replicates of the prokaryotic model organism *Escherichia coli* strain K-12 MG1655 were sequenced using currently available RNA-seq protocols from Oxford Nanopore, including direct RNA sequencing (DRS), direct cDNA sequencing (cDNA) and sequencing of PCR-amplified cDNAs (PCR-cDNA). Different rRNA-depletion, additional treatment strategy (Terminator 5´-Phosphate-Dependent Exonuclease, TEX), kit names used (RNA001, RNA002, DCS109, PCB109) and key steps of the library preparation are outlined in the graphic workflow summary. **B**, Principle of Nanopore sequencing: An ionic current drives the cDNA or the RNA strand of a RNA/cDNA hybrid through the membrane-embedded Nanopore. The motor protein, attached during library preparation, unzips the double strands, and controls the translocation speed. Translocation of the strand alters the electric signal, which is used to determine the sequence. **C**, Basic workflow for analysing Nanopore reads including basecalling and demultiplexing using ONT-developed guppy, custom scripts to perform quality control of runs/reads, minimap2 (Li, 2018) to align the reads to the reference genome and salmon (Patro et al., 2017) in alignment-based mode for gene quantification. **D**, Genome browser view of the 30 longest untrimmed reads per selected library sorted by read start position in the genomic region of the rpsP-rimM-trmD-rplS operon. The longest read of each ONT protocol is highlighted in red.

Figure 2. ONT sequencing of RNA and cDNA molecules is suitable for quantitative measurements. A, Correlation between counts measured in transcripts per million (TPM) of cDNA replicate 2 and DRS replicate 2. Each point represents one gene, colour-coded by the density at the plot position. Gene lengths are indicated by circle size. Pearson correlation is given at the top left. **B**, Correlation between TPM counts of cDNA replicate 2 and the publicly available short-read Illumina dataset (ILL). **C**, Correlation matrix between all ONT, ILL and

SMRT-Cappable-seq (SMRT) samples (Yan et al., 2018). Pearson correlation coefficients calculated from pairwise-complete observations are depicted and colour- and square-size coded. Additionally, the number of available pairwise comparisons is shown by circle size and colour in the upper right half of the plot.

Figure 3. **Analysis of transcript boundaries detected using ONT RNA-seq methods. A,** Exemplary region in *E. coli* containing the non-coding RNA *tff* and 5 other genes. Coverage profiles of raw reads for the different library protocols have been normalised to 100, which refers to the highest coverage of each analysed sample. 5´ read ends of raw reads (light-blue) and 3´ read ends of trimmed reads (dark-blue) are shown as line plots with the number of reads starting or ending at the positions shown on the right scale. Following the analysis pipeline depicted on the right we identified 5´ and 3´ enriched positions. Primary 5´ (solid lines) and 3´ends (dashed lines) derived from the analysis are shown for each dataset in the coverage plot. **B**, Accuracy of 5´end detection using raw reads assessed by comparison of distances between primary ONT 5´ends to differential RNA-seq primary transcription start sites (TSS) (Thomason et al., 2015) and SMRT-Cap TSS (Yan et al., 2018). **C**, Accuracy of 3´end detection using trimmed reads assessed by comparison of distances between primary ONT 3´ends to short-read Term-seq primary transcription termination sites (TTS) (Dar and Sorek, 2018) and SMRT-Cap TTS (Yan et al., 2018).

Figure 4. **Gene body coverage of raw and full-length enriched Nanopore reads. A**, Meta-analysis of gene body coverage profiles for genes that have an ONT-annotated 5´and 3´end. The gene bodies were scaled between the TSS and TTS to adjust for different transcript lengths. Coverage profiles show the mean values for each position after adjusting the coverage to values between 0 and 100 for each individual gene. Coverages are shown as area plots for calculated coverages based on raw (brown, dotted line) and trimmed (mint, solid line) reads. **B**, The decay from 5´ or 3´ ends and the overall coverage profiles were evaluated based on the quartile coefficient of variation (interquartile range/median, QCoV) and by comparing the mean values of the first (CoV5) and last 10% (CoV3) of the coverage

profiles to the median. Analysis of **C**, QCoV calculated from trimmed (mint) and raw (brown) coverage profiles, **D**, CoV5 (trimmed)**, E**, CoV3 (trimmed), and **F**, QCoV (trimmed) values grouped by gene lengths that are indicated by transparency of the respective library type colour.

Figure 5. Capability of ONT sequencing to capture complex bacterial transcriptional units. A, Workflow, and visualisation of transcriptional unit analysis using long-read Nanopore data. After finding overlaps between single reads (compare single-read track sorted with increasing length of the read) and gene positions, transcriptional units are defined by the unique combination of genes that are covered by a read. The total number of reads assigned to a transcriptional unit is depicted on the left. The distribution of contexts per gene was calculated from the number of reads assigned to a gene in the respective context divided by the total number of reads per gene and is visualised using a colour code. **B**, UpSet-plot showing that the comparability and number of identically detected transcriptional units in the different library preparation methods is sequencing-depth dependent. **C**, Distribution of transcriptional contexts, which is defined as the number of transcriptional units a gene is part of.

Figure 6. Advantages, disadvantages and application of Nanopore RNA-seq in prokaryotes. A, Advantages and disadvantages of the three ONT library preparation protocols for RNA sequencing are shown divided into different aspects that can be considered when setting up an experiment. Significant Pros or cons are indicated by a double sign. Efficient polyadenylation of all transcripts is critical for all protocols. **B**, Applications of Nanopore RNA-seq in prokaryotes (left), the suggested library protocol (middle) and the suggested workflow (right).

Supplementary Tables

Supplementary Table 1. Sequencing information and mapping summary statistics.

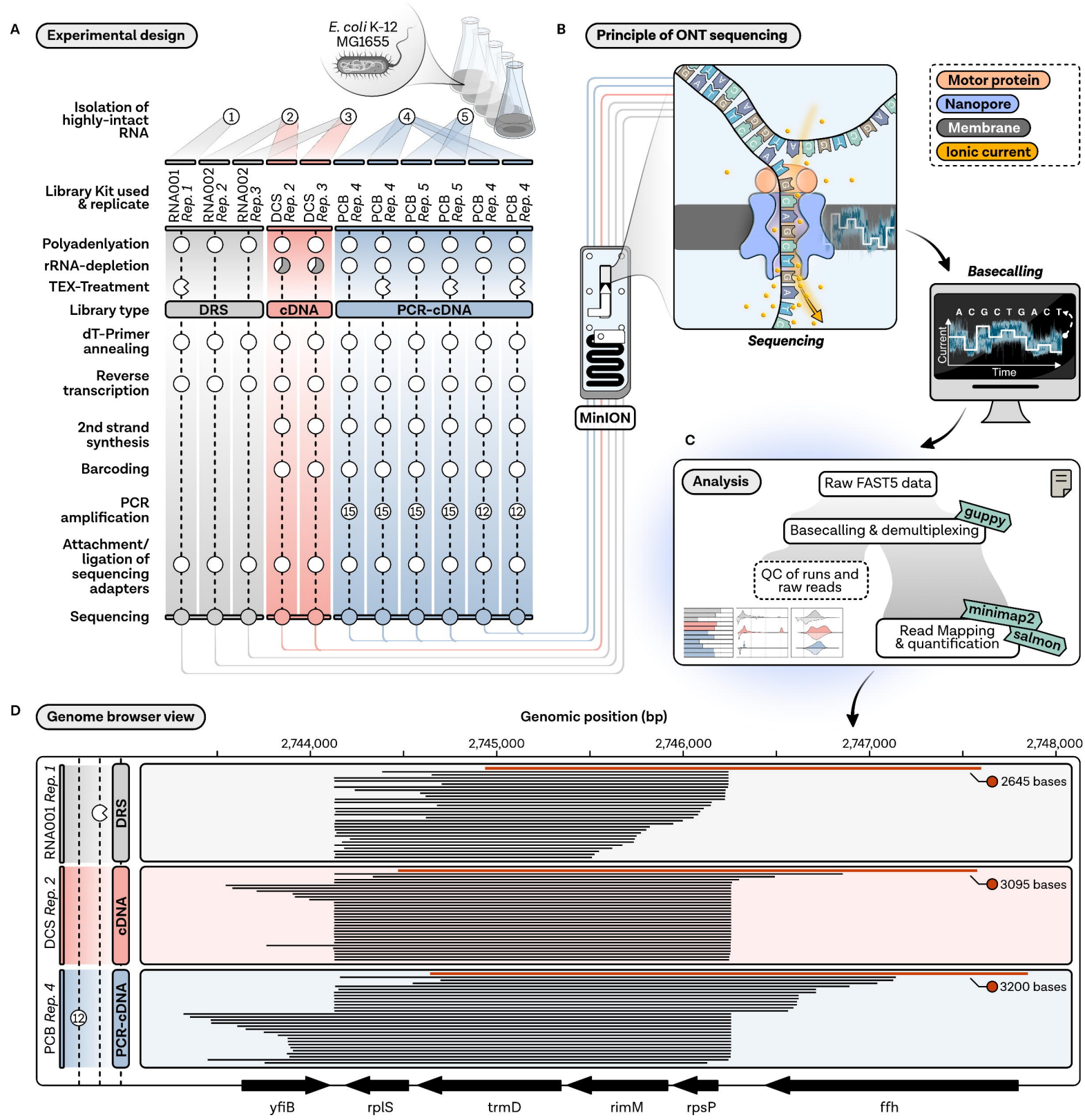
Supplementary Table 2. List of figures with link to data and code to reproduce the results.

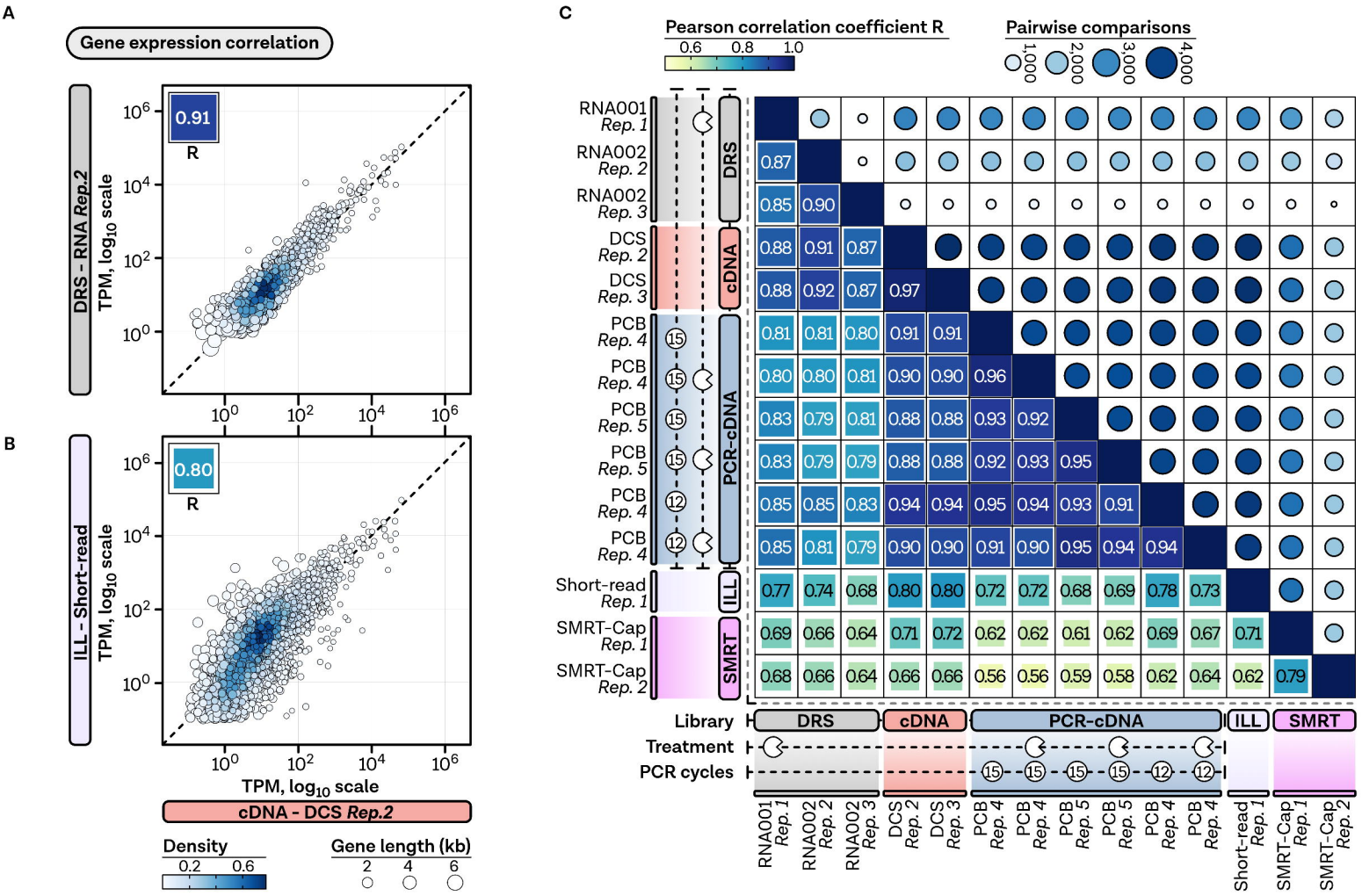
Supplementary Table 3. Salmon results in transcripts per million (TPM) for all of the ONT and used publicly available datasets.

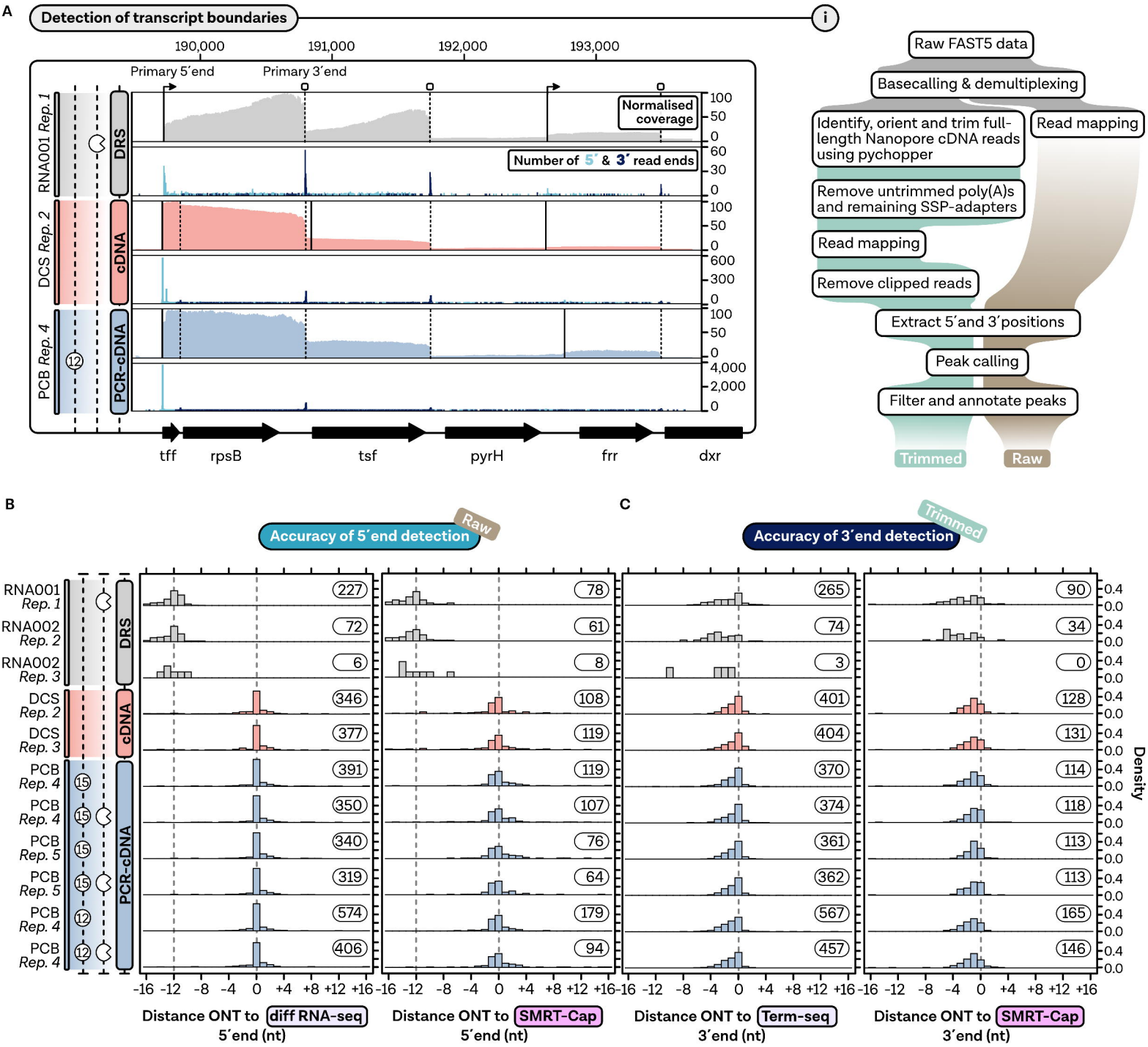
Supplementary Table 4. 5´end table for all ONT datasets and 5´end types.

Supplementary Table 5. 3´end table for all ONT datasets and 3´end types.

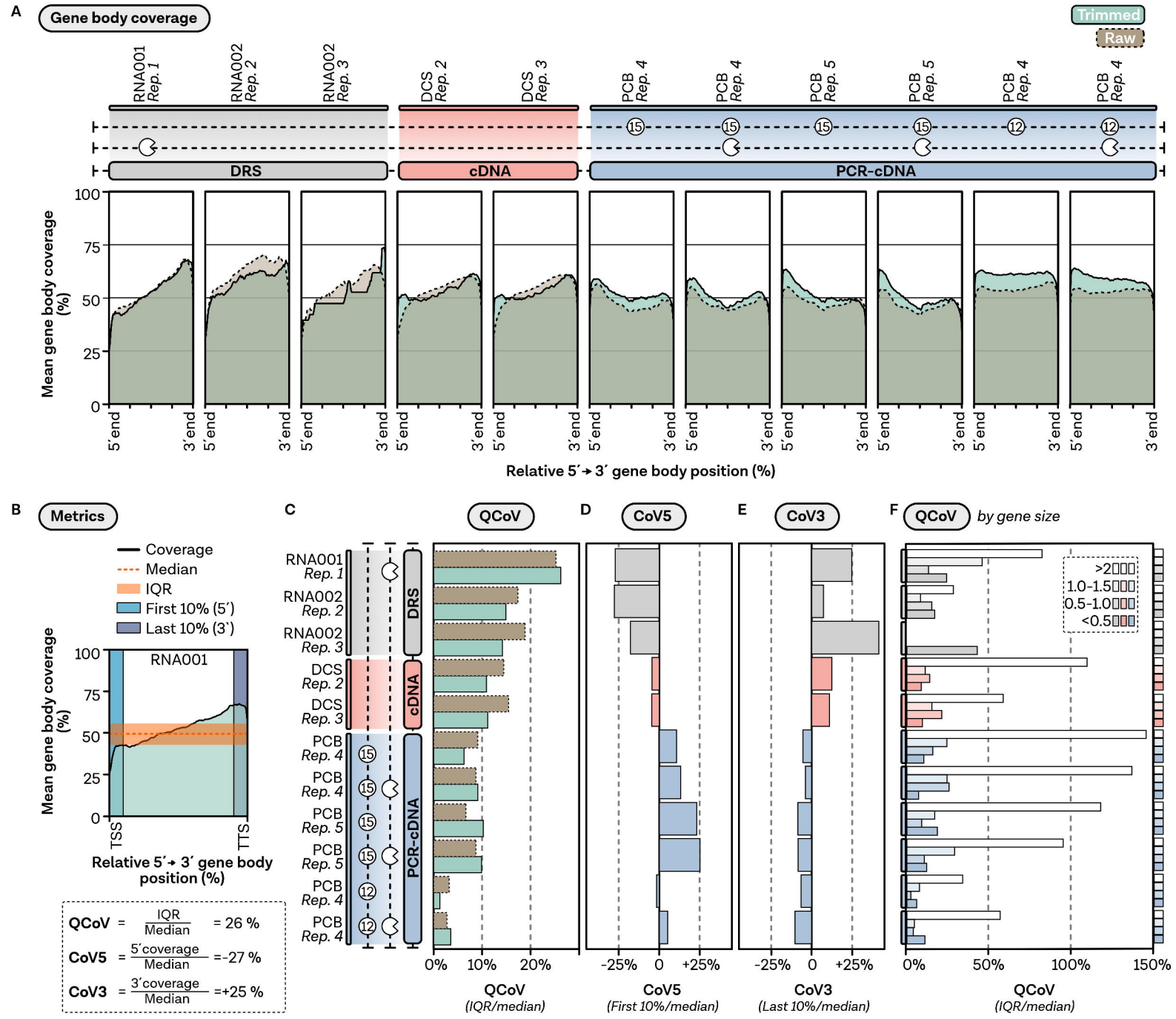
Supplementary Table 6. Transcription units table for three datasets (RNA: RNA001_Ecoli_TEX_replicate1, DCS: DCS109_Ecoli_NOTEX_replicate2, PCB: PCB109_PCR12_Ecoli_NOTEX_replicate4).


A
Experimental design
E. coli K-12 MG1655
Isolation of highly-intact RNA
Library Kit used & replicate
RNA001 Rep. 1
RNA002 Rep. 2
RNA002 Rep.3
DCS Rep. 2
DCS Rep. 3
PCB Rep. 4
PCB Rep. 4
PCB Rep. 5
PCB Rep. 5
PCB Rep. 4
PCB Rep. 4
Polyadenlyation
rRNA-depletion
TEX-Treatment
Library type
DRS
cDNA
PCR-cDNA
dT-Primer annealing
Reverse transcription
2nd strand synthesis
Barcoding
PCR amplification
Attachment/ ligation of sequencing adapters
Sequencing
B
Principle of ONT sequencing
Motor protein
Nanopore
Membrane
Ionic current
Sequencing
Basecalling
Current
Time
MinION
C
Analysis
Raw FAST5 data
Basecalling & demultiplexing
guppy
QC of runs and raw reads
Read Mapping & quantification
minimap2
salmon
D
Genome browser view
Genomic position (bp)
2,744,000
2,745,000
2,746,000
2,747,000
2,748,000
RNA001 Rep. 1
DRS
2645 bases
DCS Rep. 2
cDNA
3095 bases
PCB Rep. 4
PCR-cDNA
3200 bases
yfiB
rplS
trmD
rimM
rpsP
ffh


A
Gene expression correlation
DRS - RNA Rep.2
TPM, log10 scale
0.91
R
B
ILL - Short-read
TPM, log10 scale
0.80
R
TPM, log10 scale
cDNA - DCS Rep.2
Density
0.2 0.6
Gene length (kb)
2 4 6
C
Pearson correlation coefficient R
0.6 0.8 1.0
Pairwise comparisons
1,000 2,000 3,000 4,000
RNA001 Rep. 1
RNA002 Rep. 2
RNA002 Rep. 3
DCS Rep. 2
DCS Rep. 3
PCB Rep. 4
PCB Rep. 4
PCB Rep. 5
PCB Rep. 5
PCB Rep. 4
PCB Rep. 4
Short-read Rep. 1
SMRT-Cap Rep. 1
SMRT-Cap Rep. 2
DRS
cDNA
PCR-cDNA
ILL
SMRT
Library
Treatment
PCR cycles


A
Detection of transcript boundaries
190,000
191,000
192,000
193,000
Primary 5′end
Primary 3′end
RNA001 Rep. 1
DCS Rep. 2
PCB Rep. 4
DRS
cDNA
PCR-cDNA
Normalised coverage
Number of 5′ & 3′ read ends
tff
rpsB
tsf
pyrH
frr
dxr
Raw FASTQ data
Basecalling & demultiplexing
Identify, orient and trim full-length Nanopore cDNA reads using pychopper
Read mapping
Remove untrimmed poly(A)s and remaining SSP-adapters
Read mapping
Remove clipped reads
Extract 5′and 3′positions
Peak calling
Filter and annotate peaks
Trimmed
Raw

B
Accuracy of 5′end detection
Raw
C
Accuracy of 3′end detection
Trimmed
RNA001 Rep. 1
RNA002 Rep. 2
RNA002 Rep. 3
DCS Rep. 2
DCS Rep. 3
PCB Rep. 4
PCB Rep. 4
PCB Rep. 5
PCB Rep. 5
PCB Rep. 4
PCB Rep. 4
DRS
cDNA
PCR-cDNA
227
72
6
346
377
391
350
340
319
574
406
78
61
8
108
119
119
107
76
64
179
94
265
74
3
401
404
370
374
361
362
567
457
90
34
0
128
131
114
118
113
113
165
146
Density
-16 -12 -8 -4 0 +4 +8 +12 +16
Distance ONT to diff RNA-seq 5′end (nt)
Distance ONT to SMRT-Cap 5′end (nt)
Distance ONT to Term-seq 3′end (nt)
Distance ONT to SMRT-Cap 3′end (nt)


A
Gene body coverage
Trimmed
Raw
RNA001 Rep. 1
RNA002 Rep. 2
RNA002 Rep. 3
DCS Rep. 2
DCS Rep. 3
PCB Rep. 4
PCB Rep. 4
PCB Rep. 5
PCB Rep. 5
PCB Rep. 4
PCB Rep. 4
DRS
cDNA
PCR-cDNA
Mean gene body coverage (%)
5'end
3'end
Relative 5′→3′ gene body position (%)

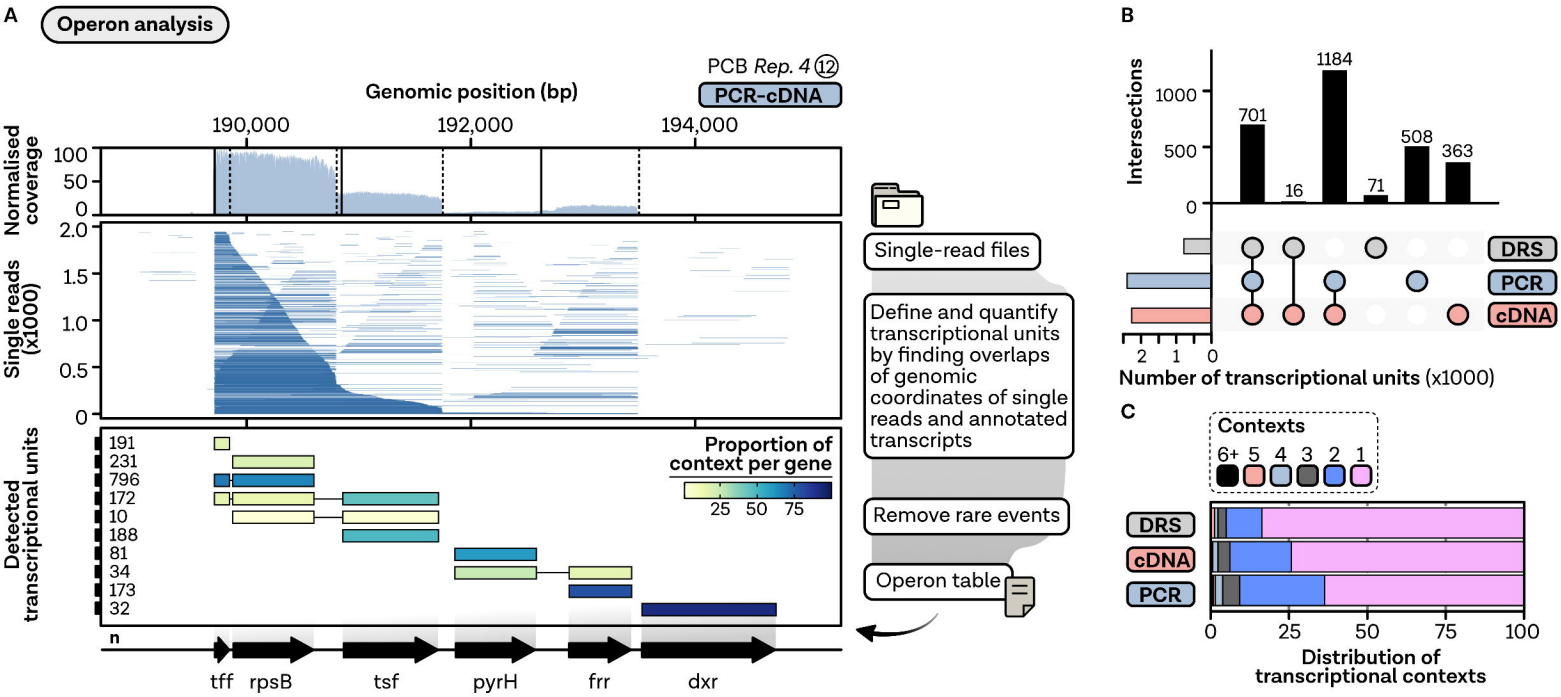
B
Metrics
Coverage
Median
IQR
First 10% (5′)
Last 10% (3`)
RNA001
Mean gene body coverage (%)
TSS
TTS
Relative 5′→3′ gene body position (%)
QCoV = IQR / Median = 26 %
CoV5 = 5′coverage / Median =-27 %
CoV3 = 3′coverage / Median =+25 %

C
QCoV
RNA001 Rep. 1
RNA002 Rep. 2
RNA002 Rep. 3
DCS Rep. 2
DCS Rep. 3
PCB Rep. 4
PCB Rep. 4
PCB Rep. 5
PCB Rep. 5
PCB Rep. 4
PCB Rep. 4
DRS
cDNA
PCR-cDNA
0%
10%
20%
QCoV
(IQR/median)

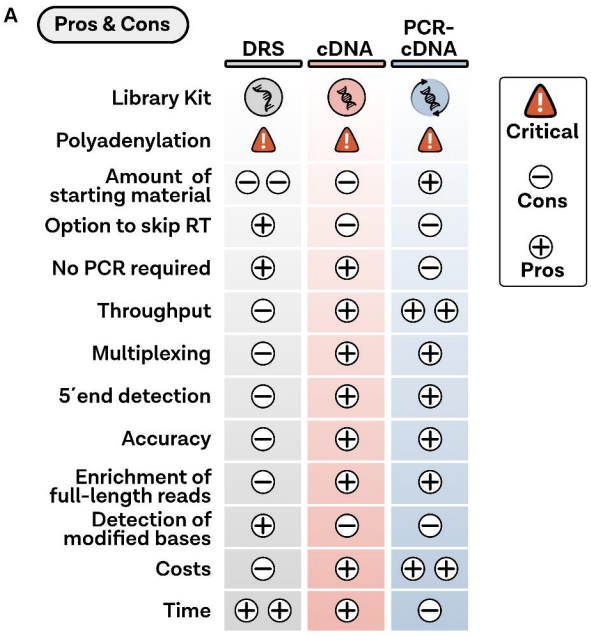
D
CoV5
-25%
0
+25%
CoV5
(First 10%/median)

E
CoV3
-25%
0
+25%
CoV3
(Last 10%/median)

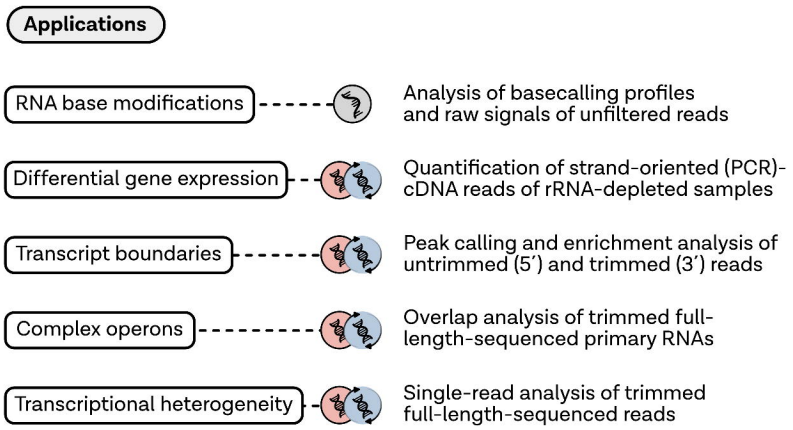
F
QCoV
by gene size
>2
1.0-1.5
0.5-1.0
<0.5
0%
50%
100%
150%
QCoV
(IQR/median)


A
Operon analysis
PCB Rep. 4 (12)
PCR-cDNA
Genomic position (bp)
190,000
192,000
194,000
Normalised coverage
100
50
0
Single reads (x1000)
2.0
1.5
1.0
0.5
0
Detected transcriptional units
191
231
796
172
10
188
81
34
173
32
n
Proportion of context per gene
25 50 75
tff rpsB tsf pyrH frr dxr
Single-read files
Define and quantify transcriptional units by finding overlaps of genomic coordinates of single reads and annotated transcripts
Remove rare events
Operon table
B
Intersections
1000
500
0
701
16
1184
71
508
363
DRS
PCR
cDNA
2 1 0
Number of transcriptional units (x1000)
C
Contexts
6+ 5 4 3 2 1
DRS
cDNA
PCR
0 25 50 75 100
Distribution of transcriptional contexts

A


Pros & Cons
DRS
cDNA
PCR-cDNA
Library Kit
Polyadenylation
Amount of starting material
Option to skip RT
No PCR required
Throughput
Multiplexing
5′end detection
Accuracy
Enrichment of full-length reads
Detection of modified bases
Costs
Time
Critical
Cons
Pros

B


Applications
RNA base modifications
Analysis of basecalling profiles and raw signals of unfiltered reads
Differential gene expression
Quantification of strand-oriented (PCR)-cDNA reads of rRNA-depleted samples
Transcript boundaries
Peak calling and enrichment analysis of untrimmed (5′) and trimmed (3′) reads
Complex operons
Overlap analysis of trimmed full-length-sequenced primary RNAs
Transcriptional heterogeneity
Single-read analysis of trimmed full-length-sequenced reads