Single-Molecule RNA Sizing Enables Quantitative Analysis of Alternative Transcription Termination

Authors:

Patiño-Guillén[1], G., Pešović, J.[2], Panic, M. [2,3], Savic-Pavicevic, D. [2], Bošković, F. [1,*], Keyser, U.F. [1,*]

Affiliations:

[1]Cavendish Laboratory, University of Cambridge, Cambridge, UK

[2]University of Belgrade – Faculty of Biology, Centre for Human Molecular Genetics, Belgrade, Serbia

[3]Institute of Virology, Vaccines and Sera "Torlak", Belgrade, Serbia.

*Corresponding authors: fnb24@cam.ac.uk , ufk20@cam.ac.uk

Summary:

Transcription, a critical process in molecular biology, has found many applications in RNA synthesis, including mRNA vaccines and RNA therapeutics. However, current RNA characterization technologies suffer from amplification and enzymatic biases that lead to loss of native information. Here, we introduce a strategy to quantitatively study both transcription and RNA polymerase behaviour by sizing RNA with RNA nanotechnology and nanopores. To begin, we utilized T7 RNA polymerase to transcribe linear DNA lacking termination sequences. Surprisingly, we discovered alternative transcription termination in the origin of replication sequence. Next, we employed circular DNA without transcription terminators to perform rolling circle transcription. This allowed us to gain valuable insights into the processivity and transcription behaviour of RNA polymerase at the single-molecule level. Our work demonstrates how RNA nanotechnology and nanopores may be used in tandem for the direct and quantitative analysis of RNA transcripts. This methodology provides a promising pathway for accurate RNA structural mapping by enabling the study of full-length RNA transcripts at the single-molecule level.

Key words:

Nanopores, RNA nanotechnology, T7 RNA polymerase, transcription

MAIN TEXT

The process of converting information from DNA into RNA is known as transcription and it is of fundamental need in gene expression[1]. Transcription is performed by the enzyme RNA polymerase (RNAP), which binds to a recognition sequence in DNA known as a promoter. RNAP unwinds double-stranded DNA and uses the template DNA strand and base pairing complementarity with the DNA template to create an RNA copy. Eventually, RNAP reaches a termination sequence, where it dissociates from the DNA template and the RNA transcript is released[2].

Transcription is not limited to the conversion of linear DNA templates. RNAPs may also transcribe circular DNA in a rolling circle manner[3–6]. Rolling circle transcription involves the continuous unidirectional synthesis of RNA from circular DNA, producing consecutive copies of the DNA sequence. The number of consecutive copies in RNA depends on how many times the polymerase goes around the DNA[7].

Besides their biological relevance, RNAPs have demonstrated immense technological impact due to their high fidelity and processivity[8], which enables the production of high quality and high yield *in vitro* transcribed RNA, used in RNA therapeutics and RNA vaccines[9,10], for example.

For the development of novel RNA technologies, in-depth knowledge of produced RNA is required. The main approaches to studying RNA are based on bulk techniques such as gel electrophoresis, quantitative reverse transcription–polymerase chain reaction (qRT-PCR) and RNA sequencing (RNA-seq)[11]. Nevertheless, these techniques may face limitations in understanding RNA diversity in its native form. qRT-PCR requires additional enzymatic steps and may encounter reverse transcription and polymerase biases[12]. The same happens with RNA-seq which may also face short-read limitations[13]. Despite its practical simplicity, accurate analysis of agarose gels can be complicated as no information about the sequence is obtained and electrophoretic mobility can be affected by the conformation of RNA[14]. Unexpected transcription species in gels may indicate transcription initiation in an alternative promoter sequence or they may correspond to the same RNA product with another conformation and therefore different electrophoretic mobility. It is therefore necessary to have a robust characterization platform that describes the role of RNAP in transcription and provides an accurate description of RNA products.

Single-molecule approaches have been employed to study both transcription and RNAPs themselves. Previously, the spatial distribution of RNAP has been characterized in growing bacteria using single-molecule fluorescence microscopy[15,16]. By combining optical tweezers and fluorescent microscopy, the dynamics of RNAP at each step of transcription has been investigated[17]; showing that single-molecule techniques can provide further insight into the characterization of transcription[18].

Nanopore sensing is another versatile single-molecule technology. It relies on the application of a potential through a nanoscale orifice in an electrolytic solution to produce an ionic current[19–21]. As analytes pass through the pore, driven by electrophoretic force, ions are depleted, causing ionic current blockage associated with the analyte's volume, conformation, or size[22]. The technique enables single-molecule detection and has facilitated the study of a wide range of biological species, including DNA[23,24], proteins[19,25,26] antibodies[19,27], RNAs[28–31], ribosomes[32], and viruses[33,34]. Recently, DNA-based labelling of RNA has been performed, enabling RNA structural mapping with nanopores[35]. This strategy involves hybridizing RNA with complementary DNA oligonucleotides to produce an RNA-DNA duplex. The short DNA strands may have overhangs to attach sequence-specific labels that can be distinguished by nanopore current signals.

In this work, we propose a method to study single-molecule transcription of T7 RNA polymerase (T7RNAP) *in vitro* with nanopores and RNA nanotechnology. Our strategy enables the quantitative description of the T7RNAP's ability to continuously transcribe different DNA templates and illustrates transcription termination while providing sequence-specific labelling. The strategy prevents reverse transcription biases, it circumvents the need for pre-amplification and allows to characterize transcripts with a length of 10s of kilobases.

We first transcribe a linear DNA that contains a T7RNAP promotor sequence and that lacks a terminator sequence. After transcription with T7RNAP, each RNA product is hybridized with complementary DNA oligos to produce an RNA-DNA duplex named RNA identifier (RNA ID) which enables parallel single-transcript mapping with nanopore sensors. Our strategy identified premature transcription termination of T7RNAP in the linear DNA construct within the origin of replication (OriC) sequence.

We next employed a circular DNA construct with the same backbone sequence as the linear DNA template to assess the rolling circle transcription of T7RNAP. Since no termination sequence was introduced to our construct, we anticipated transcription termination at random

positions and at random time points. However, after the identification of a premature transcription termination site in the OriC, we explore the enzyme's processivity in this DNA construct and study the correlation between transcript length and the premature termination sequence.

By coupling nanopore sensing with RNA nanotechnology, we demonstrated structural mapping of diverse populations of RNA transcripts, hence revealing premature terminator sequence in OriC. Our approach unlocks future study of the processivity of enzymes and the resulting sequence landscape of transcribed RNA.

R E S U L T S

RNA identifiers enable quantitative analysis of transcription termination in linear DNA

We start with the analysis of the transcripts from a linear DNA construct which contains a promoter sequence for T7RNAP (**Figure 1a**). The promoter enables transcription initiation by T7RNAP. The linear DNA construct was produced from a circular DNA plasmid by restriction digestion using DraIII. The restriction site is upstream of the T7RNAP promoter (**Figure 1a**; more details of the restriction digestion sequence position are shown in Supplementary **Table 1**). The construct is 3.1 kilobase pair-long DNA and it was engineered to have the OriC, 12 CTG tandem repeats downstream of the T7RNAP promoter, and importantly, no termination sequence. The CTG repeats are located downstream of the T7RNAP promoter, and the OriC is in the middle of the construct (**Figure 1a**; the exact distances are illustrated in Supplementary **Table 1**). Due to their vicinity to the promoter, CTG tandem repeats are used as a reference to indicate the position of transcription initiation and to indicate the 3'to 5' directionality of transcripts.

Our linear DNA construct was *in vitro* transcribed with T7RNAP. Since no termination sequence was introduced into the construct, the enzyme should transcribe the entire linear DNA and detach from it once it reaches its end (**Figure 1a**). After performing *in vitro* transcription, transcribed RNA was analysed using agarose gel electrophoresis (**Figure 1b**). DNA and RNA ladders (lanes 1 and 2, respectively) allowed us to assess the running speeds of DNA and RNA, respectively. In lanes 3 and 4, we run transcribed RNA without and with DNase I treatment, respectively.

Surprisingly, RNA that was not treated with DNase I (lane 3) produces three bands in the agarose gel. The topmost band is ascribed to linear DNA used as a template, which is ~3 kb long. The chemical nature of this band is confirmed by DNase I treatment (lane 4) which removes the band. DNase I treatment does not remove the middle and lower bands which suggests they correspond to RNA products. The middle band which is slightly lower than 3 kb RNA, could be ascribed to RNA where T7RNAP transcribes the entire DNA construct until it reaches its end, which should be 2.9 kb long (**Figure 1b**). The bottom band, however, located between 1 and 2 kb in the RNA ladder, is unexpected. This RNA could originate from transcription initiation in an alternative promoter sequence, it could correspond to RNA of an alternative transcript variant, or it could be the same RNA product with another conformation and therefore different electrophoretic mobility. Unravelling the identity of this RNA specie from the gel is challenging or impossible, since gel electrophoresis does not provide any information related to the sequence of the RNA molecule.

In order to elucidate the identity of these RNA species we use RNA nanotechnology and nanopore sensing. Initially, we hybridize RNA with short complementary single-stranded DNA (ssDNA) oligonucleotides (oligos) to produce RNA IDs (**Figure 1c**)[35]. RNA IDs reveal structural elements of RNA sequence, enabling us to determine where transcription begin, ends and any other structural arrangements. Through the inclusion of labelled oligos, RNA IDs enabled the characterization of RNA transcripts by nanopore sensing, which revealed two types of RNA after transcription. In the first transcript type, the polymerase reaches the end of the linear DNA template (END RNA), as expected. In the second RNA type, nanopore readout shows that premature transcription termination occurs in the middle of the DNA (PT RNA, **Figure 1c**) thus originating the unexpected RNA specie.

The RNA IDs were formed with oligos complementary to the linear DNA sequence (oligos sequences in Supplementary **Table 4**), to perform simultaneous sequence-specific labelling of all RNA transcripts in the sample, independently of the position where the polymerase began or stopped transcribing. Two types of labels have been designed to enable the mapping of individual transcripts. Given their proximity to the promoter, the CUG tandem repeats in RNA (originally CTG in DNA) were used as a label to identify the position where transcription was initiated. RNA IDs can pass through the nanopore in both 5'-3' and 3'-5' directions, the repeats can be used as a marker to identify the direction in which the molecule translocated through the pore. These repeat labels are named 'R' (**Figure 1c**; Supplementary **Figure 1**) and they are constituted of a $(CAG)_{10}$ oligo with an overhang sequence containing a biotin group at its 3'

end (docking strand). This oligo is reverse-complementary to the CUG repeats in target RNA. A complementary strand to the docking strand (imaging strand) also has a 3' biotin. The docking strand and imaging strand can bind two monovalent streptavidins[36] creating a single ionic current drop in nanopore signals.

Besides, the rest of the RNA is decorated with '1' bits (**Figure 1c**). These lack the biotin group in the docking strand, which only allows the binding of one streptavidin per bit and produces a distinguishable signal from the CUG repeats 'R' (Supplementary **Figure 1**). More details about the design and assembly of 'R' labels and '1' bits are shown in Supplementary **Figure 2**.

RNA ID assembly enabled single-molecule transcript analysis using glass nanopore sensors (**Figure 1d**). After applying a voltage bias, negatively charged RNA IDs translocate through the nanopore towards the positively charged electrode, blocking the pore partially and inducing a transient drop in ionic current. The current signal is directly associated with the RNA ID design and produces a unique molecular fingerprint of the RNA ID that passes through the nanopore[35]. The translocation of RNA IDs produces current signatures with downward spikes from the labelled repeats 'R' and '1' bits included in the design. The current spikes are generated by the streptavidin-biotin conjugate which blocks more ions than the RNA ID backbone on its own, producing a deeper ionic current drop[35]. 'R' labels contain two biotin-streptavidin conjugates thus blocking twice the current in comparison to '1' bits (**Figure 1e**; 'R' – red' and '1' bits – blue) that contain one streptavidin.

Two RNA IDs were detected, originating from two different RNAs. In the first one, T7RNAP reaches the end of DNA (END RNA ID). In the second, transcription termination is caused by a premature termination sequence (PT RNA ID). The RNA IDs for both PT and END transcripts are electrophoretically driven through the same nanopore, enabling characterization of the transcript products while preventing nanopore readout variability [37]. In END RNA IDs, the nanopore translocation shows the 'R' and '1' bit current spikes, followed by a prolonged plateau which is attributed to the unlabelled region of the RNA ID, only constituted by RNA-DNA duplex. The plateau suggests an unlabelled region spanning from the OriC sequence to the end of the linear DNA template. This readout corresponds to RNA where transcription terminated after the enzyme most likely detached at the end of the DNA template. The second type of current signature detected, PT RNA ID, also presents 'R' and '1' bit spikes, meaning it shares the same initial sequence as the full-length transcript, however, the translocation ends

right after the second '1' bit current spike and no plateau is observed. This '1' bit is located 0.2 kb into the OriC, indicating that this RNA specie originates from premature transcription termination within the OriC sequence.

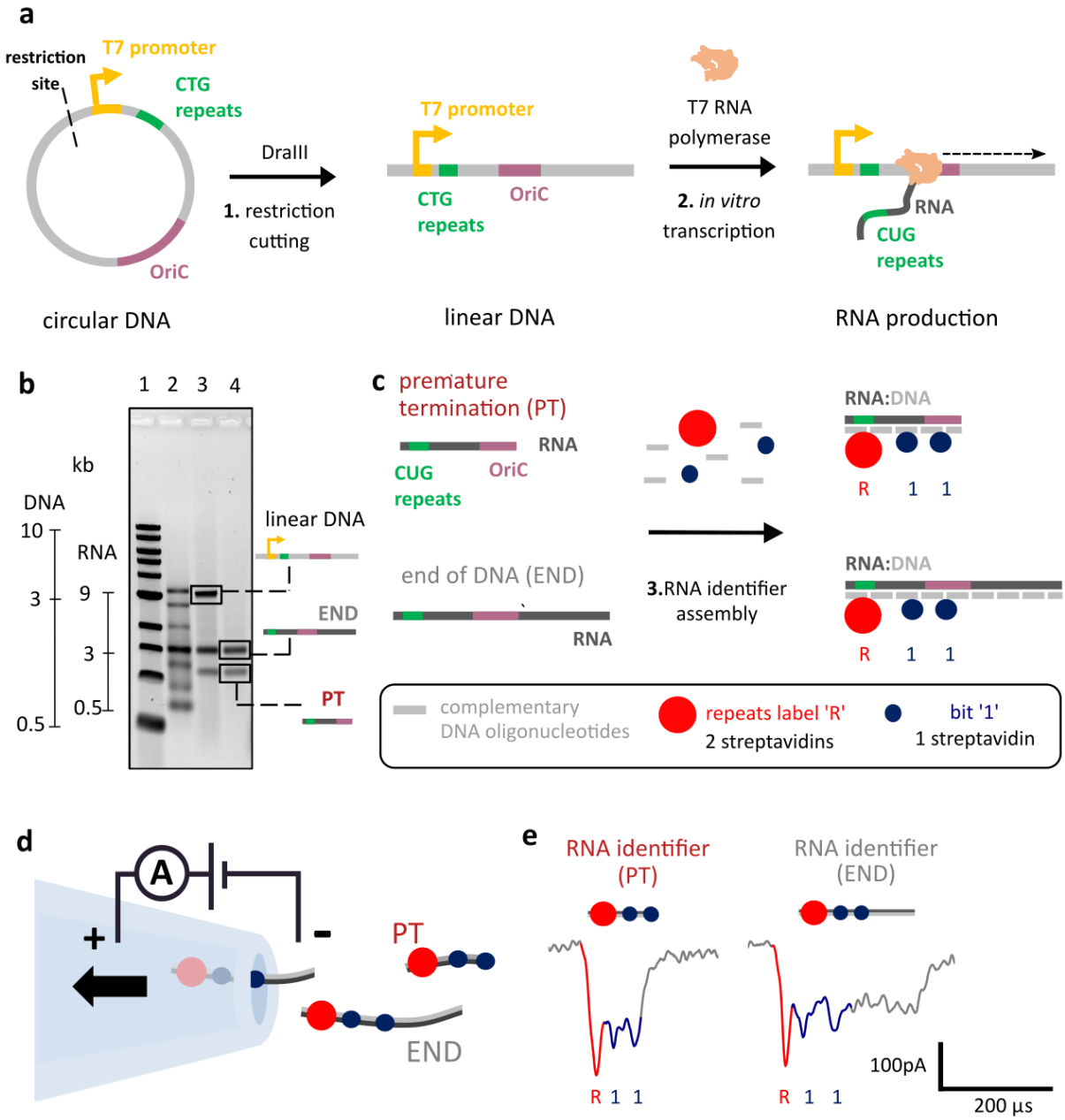

Figure 1. RNA identifier (ID) assembly enables single-molecule structural analysis of RNA transcripts using nanopore sensing. **a** *In vitro* transcription of linear DNA containing 12 CTG tandem repeats. - A circular DNA construct contains a single T7RNAP promoter, the OriC, 12 CTG tandem repeats, and a DraIII restriction site. The circular DNA construct was linearized using DraIII, by cutting upstream of the T7RNAP promoter. The linear DNA was *in vitro* transcribed using T7RNAP. **b** Agarose gel electrophoresis of transcribed linear DNA indicates the presence of two RNA species (lane 4), one in which the RNA polymerase transcribes the entire linear DNA (END) and RNA of unknown nature. With nanopore sensing, it is later revealed that the lower band corresponds to RNA where transcription is prematurely terminated within the OriC sequence (PT). Gel lanes: 1 – 1 kb ladder;

2 – single-stranded RNA ladder; 3 – transcribed RNA from linear DNA; 4 - transcribed RNA from linear DNA treated with DNaseI. **c** RNA transcripts were hybridized with short complementary DNA oligonucleotides (~40 nt), producing RNA IDs. CUG repeats in RNA were labelled with two streptavidins (repeats label 'R') indicating the beginning of transcription given their vicinity to the promoter. Further positioning along the transcript is achieved using bits '1' that have one streptavidin enabling their distinction from label 'R'. **d** RNA IDs made from RNA transcripts that have different termination points (PT or END) translocating through the nanopore. Nanopore readout is based on electrophoretically driven transport of negatively charged RNA IDs through the nanopore towards a positively charged electrode. **e** Nanopore RNA ID readouts for PT and END RNAs are presented left and right, respectively. RNA ID for both PT and END shows downward spikes associated to the labelled repeats 'R' (red) and '1' bits (blue). PT translocation finishes right after the second '1' bit hence making a clear distinction towards END translocations, which takes longer to translocate through the pore.

Having performed structural mapping of RNA to identify a premature transcription termination position, we next sought to conduct quantitative analysis of the transcripts. The RNA IDs of both termination points (PT and END, shown in **Figure 1e** and **2a**) are quantitively studied using nanopores by looking at two parameters for each RNA ID translocation. First, we compute the translocation time, *i.e.* the time it takes a molecule single RNA ID to pass through the pore, and second, the charge deficit of the translocation event, which represents the surface area of the nanopore event[38–40] (**Figure 2b**). Raw PT and END RNA ID events are shown in Supplementary **Figure 3** and Supplementary **Figure 4**, respectively. Translocation time and charge deficit of all translocation events detected are shown in Supplementary **Figure 5** and Supplementary **Figure 6**.

Translocation time and charge deficit of 70 unfolded (linear) events of each transcript type (140 events in total) are displayed in a scatter plot (**Figure 2c**). The histogram of translocation time and charge deficit are also plotted above the *x* and *y* axis, respectively. The plot shows that RNA IDs from END transcripts have longer translocation times and larger charge deficits than RNA IDs from PT transcripts. Longer RNA IDs, in this case, RNA IDs from END transcripts, require more time to translocate through the nanopore. As they take longer to translocate, more ions are depleted, which makes the charge deficit larger. Translocation time and charge deficit have a linear dependency (**Figure 2c**). Additionally, the histograms of charge deficit and translocation time exhibit distinct distributions for PT and END RNA IDs, demonstrating that both parameters are suitable for the identification of transcripts with different termination sites.

We next performed single-molecule sizing of the transcripts by calculating an estimated length of each transcript in base pairs from their respective nanopore readout (**Figure 2d**). In the RNA ID design, the labelled repeats 'R' are located 0.5 kb away from its adjacent '1' bit, and two

‘1’ bits are separated 0.6 kb from each other. A base pair-to-time conversion factor for each individual event is obtained by associating the base pair distance between each label to the distance between translocation times of their current spikes in the nanopore events. The overall event translocation time is then translated into a base pair estimate using the conversion factor for each RNA ID event.

RNA ID translocation times were converted to base pairs, revealing two distinct distributions, summarized in **Figure 2e**. The PT type RNA IDs had a mean length of 1.75 ± 0.17 kb. This finding is in agreement with the position of the OriC within the linear DNA construct (Supplementary **Table 1**). The RNA IDs from END transcripts had a mean length of 3.19 ± 0.27 kb, which is also in agreement with the length of the linear DNA available for transcription after digestion with DraIII. Transcription termination detected in nanopore sensors agrees with gel electrophoresis assays which show ~47% termination efficiency (Supplementary **Figure 7, 8**).

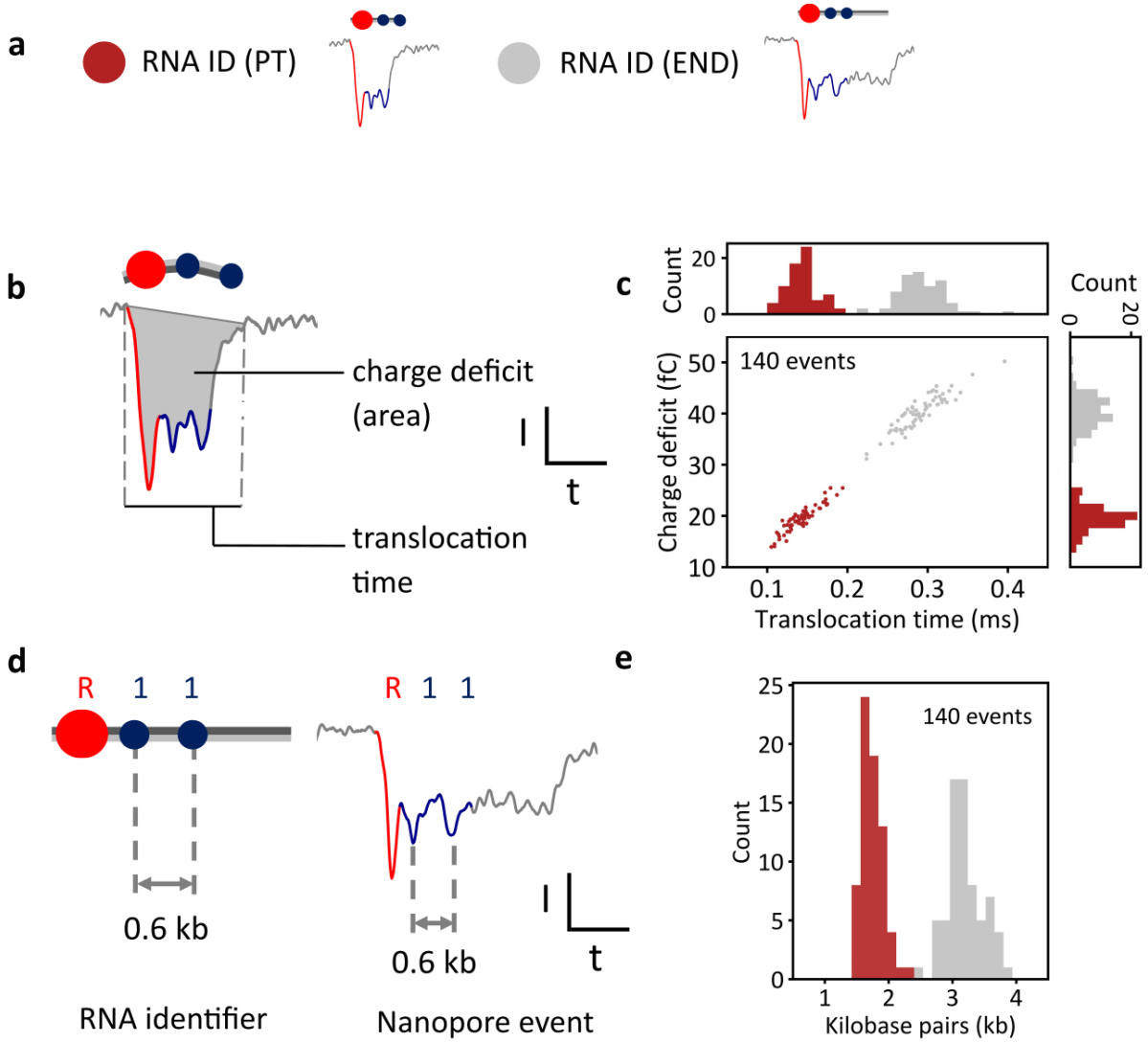

Figure 2. Quantitative analysis of single-molecular sizing of RNA determines the position of the alternative terminator **a** Example nanopore events showing translocations of RNA IDs from PT RNAs (red), where transcription terminated prematurely, and RNA IDs from long END RNAs (gray), where transcription terminated at the end of the linear DNA. **b** Physical parameters that are used to characterize nanopore translocation events, including event charge deficit that represents the area of an event, and translocation time. **c** Scatter plot of charge deficit against translocation time for RNA IDs of PT (red) and END (gray) RNA transcripts, which shows the linear dependence of both parameters. END RNA IDs require more time to translocate through nanopores. These block the ionic current for a longer time, depleting more ions, and producing a larger charge deficit. Histograms of translocation time and charge deficit show distinct distributions between both transcripts. Datapoints correspond to the translocation of RNA IDs measured within the same nanopore. The sample size was 140. **d** The ID design is used to convert translocation time into an estimate of the RNA length in base pairs, by knowing the base pair distance between labels and associating that distance to the time difference between the current downward spikes of the labels. **e** Base pair length of all molecules converted from translocation time (in *c*), which shows two distinct distributions. PT distribution has a mean length of (1.75 ± 0.16) kb and END transcripts have a mean length of (3.19 ± 0.27) kb. Errors correspond to standard deviation.

RNA identifiers characterize T7 RNA polymerase processivity in rolling circle transcription

Thanks to the molecular profiling capabilities of our strategy, we perform single-molecule analysis of T7RNAP transcription in a more complex scenario, using the circular DNA construct introduced in the previous section. The construct is not linearized, and preserves one T7RNAP promoter, the OriC sequence, and CTG tandem repeats downstream of the T7RNAP promoter.

Since supercoiled DNA can present a roadblock for *in vitro* transcription (Supplementary **Figure 9**), we performed relaxation to the circular DNA to facilitate transcription[41,42]. The relaxation was performed with *Escherichia coli* Topoisomerase I and subsequently *in vitro* transcribed with T7RNAP (**Figure 3a**). DNase I treatment is shown in Supplementary **Figure 10**. As the DNA construct is circular and lacks a termination sequence, the enzyme should not necessarily fall off after the transcription of the whole DNA sequence. Instead, the polymerase can transcribe the circular construct several times continuously, in a rolling circle manner and transcription termination is governed solely by T7RNAP stochastic dissociation or by potential alternative termination sequences, such as previously mentioned OriC (**Figure 1**). After the polymerase transcribes the entire construct once, it can continue to transcribe the construct's sequence on numerous occasions without stopping[3], producing an RNA molecule with multiple copies of the construct's sequence (**Figure 3b**).

In our DNA construct, the CTG repeats (CUG repeats in RNA) are located only 74 nucleotides away from the T7RNAP promoter. By labelling the CUG repeats in RNA, we use the 'R' repeats label to identify every time T7RNAP goes past the promoter. If T7RNAP starts transcribing the circular DNA for the first time, the 'R' repeats label will be detected once, however, if T7RNAP transcribes the whole sequence of the circular DNA and transcribes beyond the promoter, a second 'R' repeats label will appear. Corresponding numbers of labels will be detected if T7RNAP transcribes past the promoter on a third or fourth occasion. In this work, we refer to each instance we see the 'R' repeats label, as a transcription cycle 'n'. A transcription cycle 'n' corresponds to every time T7RNAP starts transcribing the circular DNA construct from the promoter. For the first transcription cycle (n =1), only one repeats label 'R' will be detected, and it will indicate the position where transcription is initiated. The subsequent appearance of the repeats label 'R' indicates the beginning of a new transcription cycle (n =2). Additionally, '1' bits were incorporated as done for RNA IDs of the linear construct, to facilitate nanopore readout and perform single-molecule mapping.

Example nanopore events of RNA IDs with n = 1 and n = 2 are shown in **Figure 3b**. RNA IDs from one transcription cycle (n = 1), in which the whole circular DNA sequence was transcribed once, produce a nanopore current signature with the downward spikes from labelled repeats 'R' and two current spikes from two '1' bits. When two full transcription cycles (n = 2) are completed by T7RNAP, the nanopore readout represents double the current signature of n = 1. Hence two sets of 'R' peaks and bits can be observed (**Figure 3b**) in the event with twice the duration. The sequence-specific labels of RNA IDs enable the distinction of the position within the RNA ID in which the second transcription cycle began and the position where transcription terminated.

With our strategy, we observed RNA IDs originating from one (blue), two (orange), three (green), four (red), or five (purple) transcription cycles (**Figure 3c**), showcasing the high processivity of T7RNAP[43]. Each nanopore readout shows its corresponding 'R' and '1' current downward spikes which can be used to infer the number of transcription cycles. Nanopore-identified RNA transcripts with multiple transcription cycles were also visualized by agarose gel electrophoresis (**Figure 3d**, lane 1), using an ssRNA ladder as a reference to infer transcript length (lane 2). The multiple bands in lane 1 are ascribed to RNA products from one (blue) to five (purple) transcription cycles. RNA IDs from produced transcripts were also characterized using agarose electrophoresis and presented the multiple bands associated with the different transcription cycles (Supplementary **Figure 11**).

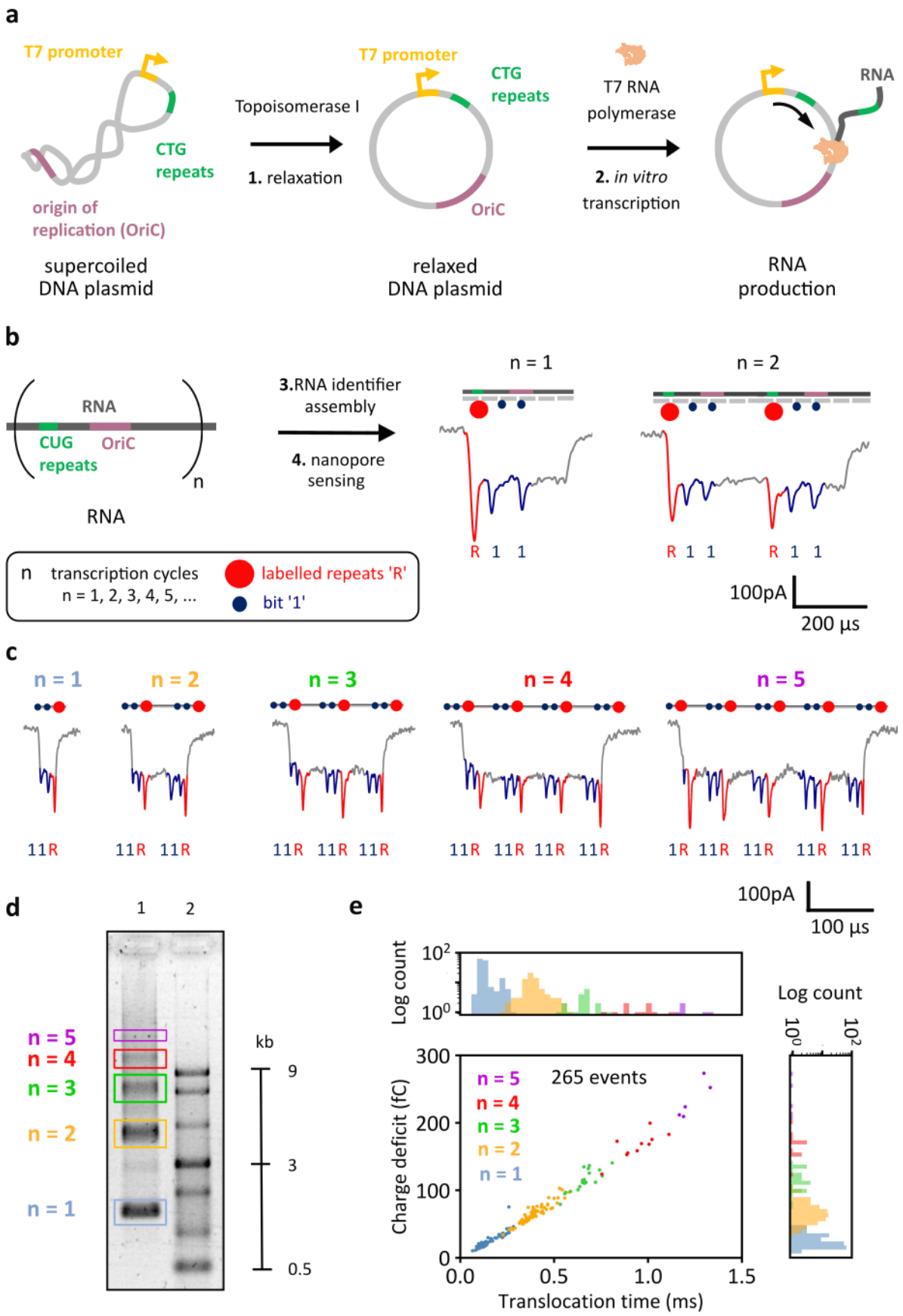

Figure 3. RNA identifier (ID) assembly enables single-molecule characterization of transcription of circular DNA derived form DM1 patients. **a** *In vitro* transcription of the DNA plasmid containing CTG tandem repeats. The

supercoiled plasmid contains the T7RNAP promoter, the OriC and 12 CTG tandem repeats. The DNA was relaxed with Topoisomerase I and *in vitro* transcribed using T7RNAP, allowing RNA transcription in a rolling-circle manner. **b** T7RNAP produces long RNA that contains multiple ('n') copies of the plasmid sequence. Transcribed RNA was hybridized with short complementary DNA oligonucleotides (~40 nt), producing RNA IDs. CUG repeats in RNA were labelled (repeats label 'R') indicating the beginning of a transcription cycle given their vicinity to the T7RNAP promoter and '1' bits were included to facilitate transcript identification. Nanopore RNA ID readouts (n = 1 and n = 2) show downward spikes associated to the labelled repeats 'R' (red) and '1' bits (blue). **c** Nanopore readouts for RNA IDs of transcripts produced from 1 to 5 transcription cycles 'n', each showing the labelled repeats 'R' and '1' bits. RNA IDs can translocate both 5'-3' and 3'-5' directions through the nanopore. These events show translocations in the 3'-5' direction. More example events are shown in Supplementary **Figure 12-16**. **d** Electrophoretic mobility shift assay shows single-stranded RNA ladder (ssRNA) on lane 2. Lane 1 shows transcription products of the relaxed plasmid, confirming nanopore readout of various transcript lengths. Gel: 1 % (w/v) agarose, 1 × TBE, 0.02% sodium hypochlorite. **e** Scatter plot of charge deficit against translocation time for RNA IDs of multiple transcription cycles, which shows the linear dependence of both parameters. Events with more cycles correspond to longer RNA molecules that need more time to translocate through nanopores, which blocks the ionic current for longer. Translocation time and charge deficit distributions show multiple transcripts with the same length at different transcription cycles. Datapoints correspond to the translocation of RNA IDs measured within the same nanopore. Histograms plotted in linear scale are presented in Supplementary **Figure 17**. The sample size was 265. Exemplary measurement comparing charge deficit of unfolded (linear) and folded RNA IDs is presented in Supplementary **Figure 18**.

For an exemplary nanopore measurement, the obtained translocation events were categorized by the number of transcription cycles (the sample size was 265 unfolded RNA ID events). As can be observed from **Figure 3e**, translocation time increases linearly with the number of transcription cycles because longer molecules take more time to translocate through the pore. For the same reason, the charge deficit of the events also increases with a higher number of transcription cycles (**Figure 3e**). Both parameters are seen to be linearly correlated. Single-molecule quantification of RNA with different transcription cycles reveals that the transcript count decreases with the number of transcription cycles, meaning their length. This can also be observed in histograms of charge deficit and translocation time (**Figure 3e**). Both histograms show localized time distributions, each corresponding to a transcription cycle, which indicates that there are RNA IDs with a similar length for each of the transcription cycles.

Up to this point, two experimental procedures have been discussed. First, the study of premature transcription termination of T7RNAP in a linear DNA template through single-molecule sizing of the transcripts using the RNA ID design. Second, the assembly of RNA IDs from transcripts with multiple lengths, produced from the rolling circle transcription of a circular DNA, and their identification using nanopore sensing.

Both procedures are now combined to elucidate T7RNAP's premature transcription termination in the circular plasmid and to obtain further understanding of the enzyme's capability to continuously synthesise RNA without releasing the template strand. During transcription of the plasmid, premature termination can occur at the OriC as previously discussed, or the polymerase can continue transcribing past this sequence (**Figure 4a**). If the polymerase continues transcribing, it can perform another transcription cycle and can go around the plasmid until it encounters the premature termination sequence again or stochastically falls off after multiple transcription cycles. It is then expected that most of the RNA products are the result of termination at the OriC despite the number of transcription cycles that the polymerase performs. Representative nanopore translocation events of RNA IDs with termination within the OriC sequence are shown in **Figure 4b**, where the translocation finishes after the second '1' bit regardless of the number of transcription cycles.

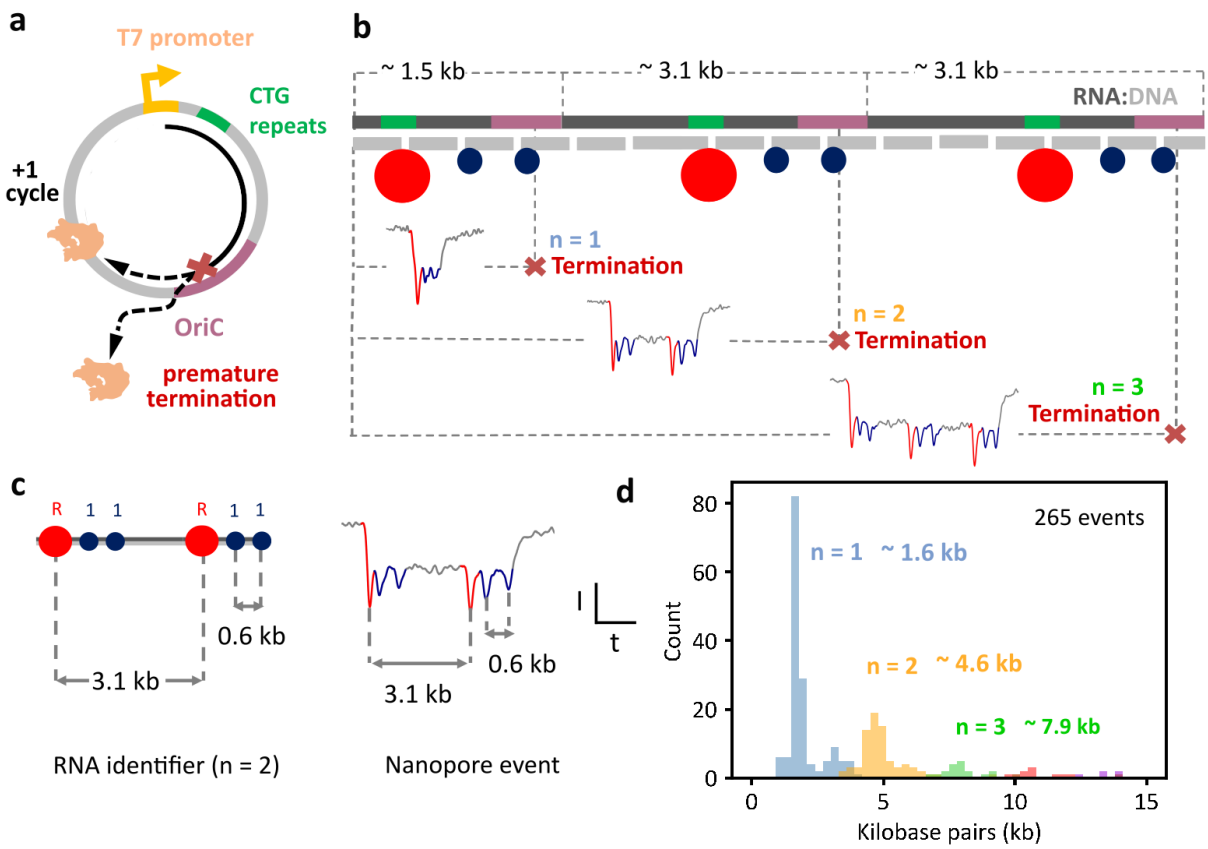

Figure 4. T7 RNA Polymerase's transcription capability and premature termination analysis in circular DNA at the single-molecule level. **a** Schematic of DNA plasmid, showing how termination can occur at the OriC sequence or the polymerase can continue transcribing the circular DNA for another cycle. **b** Schematic of RNA ID showing example events of termination after 1, 2 or 3 transcription cycles. **c** The ID design is used to convert translocation time into an estimate of the RNA length in base pairs. The distance between current spikes of 'R' and '1' bits is translated to base pairs to obtain a conversion factor to obtain the length in base pairs for the

whole event. **d** Histogram of normalized length (base pairs) of RNA IDs produced from the transcription of circular DNA, showing premature termination at ~1.6 kb, ~4.6 kb, and ~7.9 kb in the OriC sequence, after transcription of 1 to 3 cycles, respectively. The sample size was 265 nanopore events. Raw translocation events are presented in Supplementary **Figure 19-21**.

The translocation time of the nanopore events was normalized into base pairs using the ID design (**Figure 4c**), which showed localized distributions attributed to premature termination within the OriC for multiple transcription cycles (**Figure 4d**). n = 1 events show RNA IDs of ~1.6 kb, suggesting premature termination within the OriC right after initiation of the first transcription cycle. n = 2 transcript size is ~4.6 kb, which corresponds to the transcription of the full plasmid plus the transcription from the T7RNAP promoter to the OriC in the second transcription cycle. The same occurs with n = 3 events, these have a transcript size of ~7.9 kb, which corresponds to two full transcription cycles and the transcription from the T7RNAP promoter to the OriC in the third cycle. Premature transcription termination at the multiple transcription cycles was further confirmed *via* agarose gel electrophoresis in DNA constructs engineered with the OriC at different positions from the T7RNAP promoter. Supplementary **Figures 22,23** show transcription of linear and circular DNA with insertions of different length between T7RNAP promoter and OriC, and **S**upplementary **Figure 24** shows transcription of a different linear DNA template that also contains a T7RNAP promoter and OriC.

The normalized length distribution of the RNA IDs after performing molecular sizing clearly shows a decay-like behaviour, meaning that there are fewer RNA molecules resulting from a higher number of transcription cycles. Our results indicate the expected behaviour, as the RNA polymerase must overcome the premature termination sequence in each of the cycles. Hence, the RNA ID length distribution describes T7RNAP's capability to continuously transcribe circular DNA and the efficiency of alternative termination.

D I S C U S S I O N

We showcase the utilization of RNA IDs to analyse complex transcript populations at the single-molecule level. This approach offers a means to directly label RNA, enabling the concurrent analysis of multiple transcripts and the visualization of sequence-specific biomarkers. By doing so, we address the shortcomings of commonly employed transcript characterization methods like gel electrophoresis, which can be challenging to interpret and lack sequence information. Furthermore, our method overcomes reverse transcription and DNA polymerase biases and is not restricted to the analysis of short transcripts.

Transcription of linear DNA and subsequent RNA ID assembly are utilized to investigate premature transcription termination. The translocation time and the charge deficit are reliable parameters for describing the transcript length. Moreover, an estimated base pair length for different transcripts was computed without additional reference molecules. This enabled the identification of a premature transcription termination site within the OriC (~47% termination efficiency) in our linear DNA construct. Previously, termination efficiency of up to 92% has been shown for bacterial RNA polymerases[44,45], indicating termination in the OriC is not as efficient for T7RNAP. The OriC also affected T7RNAP's capability to continuously transcribe the circular DNA construct, where we observed that termination within the OriC dominates at different transcription cycles.

The refined understanding we provide here is essential for downstream applications, including the synthesis and characterization of therapeutic RNAs and messenger RNA vaccines, as well as for *in vitro* and *in vivo* production of RNAs and proteins. Moreover, the comprehensive tools here presented are useful for the study of RNA produced from rolling circle transcription, which finds therapeutic applications in the enhancement of RNA delivery in cells and the prolongation of protein expression[5,46,47].

Finally, the strategy presented is also well-suited for describing transcription processes with relevance in fundamental research, such as analysis of gene expression and profiling or for the identification of transcription regulation factors for DNA-based enzymes. Our work also indicates that single-molecule characterization of biomarkers in RNA, like labelled CUG tandem repeats are suitable for the identification of transcription initiation. Successful transcription and recognition of these repeats, which have mutation-induced capabilities[48,49], opens a door towards in-depth characterization of tandem repeats, which may have clinical relevance[48,50].

M E T H O D S

Circular DNA plasmid production

The circular plasmid was chosen from the existing plasmid collection from Centre for Human Molecular Genetics, University of Belgrade-Faculty of Biology which was previously produced by several reactions of cloning and subcloning of CTG repeats originating from human *DMPK* locus using pJet1.2/blunt cloning vector from CloneJET PCR Cloning Kit (Thermo Fisher Scientific). The subcloned region sequence was verified with Sanger

sequencing before prior use. The whole plasmid sequencing was performed by the DNA sequencing facility, Department of Biochemistry, University of Cambridge. The plasmid was propagated in *Escherichia coli* JM110 strain upon chemical transformation. The plasmid DNA was purified using GeneJET Plasmid Miniprep Kit (Thermo Fisher Scientific) and eluted in nuclease-free water. The quality of plasmid DNA was checked using agarose gel electrophoresis, while the concentration was measured using Qubit 2.0 fluorometer (Thermo Fisher Scientific). More information is available in Supplementary Methods.

Experimental design

The circular DNA plasmid was digested or relaxed depending on the type of DNA template required. The plasmid was digested to obtain linear DNA using DraIII-HF or treated with *Escherichia coli* Topoisomerase I to only relax the circular plasmid. Once the plasmid was cut or relaxed, it was purified using DNA purification kit. The purified DNA was then transcribed with T7RNAP. After transcription, the reaction was treated with DNase I to remove the DNA templates, and RNA was purified using RNA purification kit. Finally, purified RNA was hybridized with DNA oligonucleotides to produce RNA ID which were then characterized using nanopore sensing.

Circular DNA plasmid digestion

DraIII-HF (New England Biolabs (NEB), R3510S) was used to linearize the circular DNA plasmid. Reaction contained 2,000 ng of DNA, according to the manufacturer's recommendations. After DraIII-HF treatment, DNA was purified using Monarch PCR & DNA Cleanup Kit (5 μg) (NEB, T1030S) following the manufacturer's instructions. More information is available in Supplementary Methods.

Circular DNA plasmid relaxation

Escherichia coli Topoisomerase I (NEB, M0301) was used to relax a supercoiled circular DNA plasmid. Reaction of 2,000 ng of DNA was prepared according to the manufacturer's recommendation. After treatment, the enzyme was inactivated in a thermocycler (65 °C, 20 min). Topoisomerase I treated DNA was purified using Monarch PCR & DNA Cleanup Kit (5 μg) (NEB, T1030S) following the manufacturer's instructions. See Supplementary Methods for additional details.

T7 RNA polymerase *in vitro* transcription of DNA

Purified linear and circular DNA were both *in vitro* transcribed using HiScribe™ T7 Quick High Yield RNA Synthesis Kit (NEB, E2050S). 240 ng reaction was prepared according to the manufacturer's recommendation (see details in Supplementary Methods). After 4 hours of incubation (required for transcription) RNA was directly treated with 4 units of DNase I (NEB, M0303S) to remove linear or circular DNA used as a template for transcription (see details in Supplementary Methods). After DNA removal, RNA was purified using Monarch RNA Clean-up Kit (50 μg) (NEB, T2040S).

RNA identifier fabrication

To fabricate an RNA ID, purified RNA was mixed with complementary oligos, repeats label 'R' oligos, and '1' bit oligos. A 40 μl reaction included 800 fmol of target RNA (or 20 nM in final volume) and 2400 fmol of the complementary oligos (or 60 nM in final volume). For the inclusion of repeats label 'R' and '1' bits, 2400 fmol of the oligos containing the docking strand were added. The sequences of the oligos used for RNA ID assembly of transcripts produced from linear DNA templates are found in Supplementary **Table 4** and oligos implemented for RNA ID assembly of transcripts produced from rolling circle transcription are found in Supplementary **Table 5**. The imaging strand was added in 1.5 times excess (10.8 pmol or 270 nM) to the three available docking strands; one from the repeats label 'R' and two from the two '1' bits in the design. The mixture was done in 100 mM LiCl, 1× TE buffer (10 mM Tris-HCl buffer, 1 mM ethylenediaminetetraacetic acid, pH 8.0) and nuclease-free water was added to achieve the final reaction volume[35]. LiCl was used to prevent magnesium fragmentation[51] and nuclease-degradation of RNA that relies on magnesium ions[52]. Prior to use, the nuclease-free water was filtered with MF-Millipore membrane filters (0.22 μm pore size) and irradiated with UV light for 10 minutes. The reaction components were mixed by pipetting and spinning down briefly. Assembly of the components was done by heating to 70 °C for 30 s and gradually cooling over 45 minutes to room temperature (90 cycles of 30 s where temperature decreases by 0.5 °C in each cycle). RNA IDs were filtered twice in 100 kDa cut-off Amicon filters to remove excess oligos. 10 mM Tris-HCl (pH 8.0) with 0.5 mM $MgCl_2$ was used as a washing buffer for filtration. After purification, RNA IDs were kept at 4 °C before nanopore and agarose gel characterization.

Nanopore fabrication

Glass nanopores were produced from quartz glass capillaries with 0.5 mm outer diameter and 0.2 mm inner diameter (Sutter Instruments, US) using a laser-assisted capillary puller P-2000 (Sutter Instruments, US). The nanopores had diameters of 8 to 15 nm[53], achieved by using the following heating parameters of the instrument: HEAT, 470; VEL, 25; DEL, 170; and PUL, 200.

Nanopore measurements

For nanopore measurements, RNA ID was diluted to 400 picomolar (pM) in 4 M LiCl, 1× TE, pH 9.4. A fixed potential of 600 mV was used for all measurements and data was recorded using an Axopatch 200B. Single translocation events were isolated from the ionic current by setting thresholds in some signal parameters, which included charge deficit (area of the event), peak current, and translocation time, using home-built LabVIEW codes. In its lower limit, charge deficit and translocation time thresholds were set close to our detection limit and the upper limit was set significantly higher than the value that we would expect from any translocation associated with RNA IDs in our samples. This enabled us to visualize events from RNA ID translocations while filtering out random RNA blobs or loose streptavidin. For further study of the RNA IDs, unfolded RNA ID (linear RNA ID) was selected. We identified the number of downward current spikes in the events and used the distance between spikes to associate them with our initial designs and categorize them accordingly.

A C K N O W L E D G M E N T S

U.F.K. was supported by funding from an European Research Council (ERC) consolidator grant (DesignerPores no. 647144) and an ERC-2019-PoC grant (PoreDetect no. 899538). G.P.G. acknowledges funding from EPSRC CDT MRes/PhD Studentship in Nanoscience and Nanotechnology (NanoDTC Cambridge EP/S022953/1) and Trinity-Henry Barlow Scholarship. F.B. acknowledges research funding from the George and Lilian Schiff Foundation Studentship, the Winton Programme for the Physics of Sustainability PhD Scholarship and St John's College Benefactors' Scholarship. J.P., M.P., and D.S.P. acknowledge funding from the Science Fund of the Republic of Serbia, Grant No. 7754217, READ-DM1. We thank Dr. Casey Platnich and Dr. Roger Rubio-Sánchez for critically reading the manuscript. We thank Lorenzo Peri for facilitating python-based graphic user interfaces used for analysis of data.

COMPETING INTERESTS

F.B. and U.F.K. are inventors of two patents related to RNA analysis with nanopores (UK patent application no. 2113935.7, in process; UK Patent application nos. 2112088.6 and PCT/GB2022/052171, in process) submitted by Cambridge Enterprise on the behalf of the University of Cambridge. U.F.K. is a co-founder of Cambridge Nucleomics. All other authors have no competing interests.

REFERENCES

1. Engstrom, M. D. & Pfleger, B. F. Transcription control engineering and applications in synthetic biology. *Synth Syst Biotechnol* **2**, 176–191 (2017).

2. Sousa, R. & Mukherjee, S. T7 RNA Polymerase. in *Progress in Nucleic Acid Research and Molecular Biology* vol. 73 1–41 (2003).

3. Furukawa, K. *et al.* Fluorescence generation from tandem repeats of a malachite green RNA aptamer using rolling circle transcription. *Bioorg Med Chem Lett* **18**, 4562–4565 (2008).

4. Dang, Y. & Green, B. R. Long transcripts from dinoflagellate chloroplast minicircles suggest ‘rolling circle’ transcription. *Journal of Biological Chemistry* **285**, 5196–5203 (2010).

5. Jang, M., Kim, J. H., Nam, H. Y., Kwon, I. C. & Ahn, H. J. Design of a platform technology for systemic delivery of siRNA to tumours using rolling circle transcription. *Nat Commun* **6**, 7930 (2015).

6. Beyer, S., Nickels, P. & Simmel, F. C. Periodic DNA nanotemplates synthesized by rolling circle amplification. *Nano Lett* **5**, 719–722 (2005).

7. Mohsen, M. G. & Kool, E. T. The Discovery of Rolling Circle Amplification and Rolling Circle Transcription. *Acc Chem Res* **49**, 2540–2550 (2016).

8. Borkotoky, S. & Murali, A. The highly efficient T7 RNA polymerase: A wonder macromolecule in biological realm. *Int J Biol Macromol* **118**, 49–56 (2018).

9. Bloom, K., van den Berg, F. & Arbuthnot, P. Self-amplifying RNA vaccines for infectious diseases. *Gene Ther* **28**, 117–129 (2021).

10. Zhu, Y., Zhu, L., Wang, X. & Jin, H. RNA-based therapeutics: an overview and prospectus. *Cell Death Dis* **13**, (2022).

11. Costa, C., Giménez-Capitán, A., Karachaliou, N. & Rosell, R. Comprehensive molecular screening: From the RT-PCR to the RNA-seq. *Transl Lung Cancer Res* **2**, 87–91 (2013).

12. Conn, V. & Conn, S. J. SplintQuant: a method for accurately quantifying circular RNA transcript abundance without reverse transcription bias. *RNA* **25**, 1202–1210 (2019).

13. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* **15**, 201–206 (2018).

14. Abe, B. T., Wesselhoeft, R. A., Chen, R., Anderson, D. G. & Chang, H. Y. Circular RNA migration in agarose gel electrophoresis. *Mol Cell* **82**, 1768–1777 (2022).

15. Bakshi, S., Choi, H. & Weisshaar, J. C. The spatial biology of transcription and translation in rapidly growing Escherichia coli. *Front Microbiol* **6**, 636 (2015).

16. Janissen, R. *et al.* Global DNA Compaction in Stationary-Phase Bacteria Does Not Affect Transcription. *Cell* **174**, 1188-1199.e14 (2018).

17. Skinner, G. M., Baumann, C. G., Quinn, D. M., Molloy, J. E. & Hoggett, J. G. Promoter binding, initiation, and elongation by bacteriophage T7 RNA polymerase: A single-molecule view of the transcription cycle. *Journal of Biological Chemistry* **279**, 3239–3244 (2004).

18. Janissen, R., Eslami-Mossallam, B., Artsimovitch, I., Depken, M. & Dekker, N. H. High-throughput single-molecule experiments reveal heterogeneity, state switching, and three interconnected pause states in transcription. *Cell Rep* **39**, 110749 (2022).

19. Xue, L. *et al.* Solid-state nanopore sensors. *Nat Rev Mater* **5**, 931–951 (2020).

20. Smeets, R. M. M. *et al.* Salt dependence of ion transport and DMA translocation through solid-state nanopores. *Nano Lett* **6**, 89–95 (2006).

21. Shi, W., Friedman, A. K. & Baker, L. A. Nanopore Sensing. *Anal Chem* **89**, 157–188 (2017).

22. Ying, Y. L. *et al.* Nanopore-based technologies beyond DNA sequencing. *Nat Nanotechnol* **17**, 1136–1146 (2022).

23. Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Biophysics (Oxf)* **93**, 13770–13773 (1996).

24. Kang, X. *et al.* Nanopore-Based Fingerprint Immunoassay Based on Rolling Circle Amplification and DNA Fragmentation. *ACS Nano* **17**, 5412–5420 (2022).

25. Houghtaling, J. *et al.* Estimation of Shape, Volume, and Dipole Moment of Individual Proteins Freely Transiting a Synthetic Nanopore. *ACS Nano* **13**, 5231–5242 (2019).

26. Liu, K. *et al.* Detecting topological variations of DNA at single-molecule level. *Nat Commun* **10**, 3 (2019).

27. Ying, Y. L., Yu, R. J., Hu, Y. X., Gao, R. & Long, Y. T. Single antibody-antigen interactions monitored: Via transient ionic current recording using nanopore sensors. *Chemical Communications* **53**, 8620–8623 (2017).

28. Van Den Hout, M., Vilfan, I. D., Hage, S. & Dekker, N. H. Direct force measurements on double-stranded RNA in solid-state nanopores. *Nano Lett* **10**, 701–707 (2010).

29. Zahid, O. K., Wang, F., Ruzicka, J. A., Taylor, E. W. & Hall, A. R. Sequence-Specific Recognition of MicroRNAs and Other Short Nucleic Acids with Solid-State Nanopores. *Nano Lett* **16**, 2033–2039 (2016).

30. Chau, C. *et al.* Probing RNA Conformations Using a Polymer-Electrolyte Solid-State Nanopore. *ACS Nano* **16**, 20075–20085 (2022).

31. Tripathi, P. *et al.* Discrimination of RNA fiber structures using solid-state nanopores. *Nanoscale* **14**, 6866–6875 (2022).

32. Raveendran, M., Leach, A. R., Hopes, T., Aspden, J. L. & Actis, P. Ribosome Fingerprinting with a Solid-State Nanopore. *ACS Sens* **5**, 3533–3539 (2020).

33. Taniguchi, M. *et al.* Combining machine learning and nanopore construction creates an artificial intelligence nanopore for coronavirus detection. *Nat Commun* **12**, 3726 (2021).

34. Bošković, F. *et al.* Simultaneous identification of viruses and viral variants with programmable DNA nanobait. *Nat Nanotechnol* **18**, 290–298 (2023).

35. Bošković, F. & Keyser, U. F. Nanopore microscope identifies RNA isoforms with structural colours. *Nat Chem* **14**, 1258–1264 (2022).

36. Howarth, M. *et al.* A monovalent streptavidin with a single femtomolar biotin binding site. *Nat Methods* **3**, 267–273 (2006).

37. Tang, W., Fried, J. P., Tilley, R. D. & Gooding, J. J. Understanding and modelling the magnitude of the change in current of nanopore sensors. *Chem Soc Rev* **51**, 5757–5776 (2022).

38. Meller, A., Nivon, L., Brandin, E., Golovchenko, J. & Branton, D. Rapid nanopore discrimination between single polynucleotide molecules. *PNAS* **97**, 1079–1084 (2000).

39. Kumar Sharma, R., Agrawal, I., Dai, L., Doyle, P. S. & Garaj, S. Complex DNA knots detected with a nanopore sensor. *Nat Commun* **10**, 4473 (2019).

40. Rivas, F. *et al.* Label-free analysis of physiological hyaluronan size distribution with a solid-state nanopore sensor. *Nat Commun* **9**, 1037 (2018).

41. Pommier, Y., Sun, Y., Huang, S. Y. N. & Nitiss, J. L. Roles of eukaryotic topoisomerases in transcription, replication and genomic stability. *Nat Rev Mol Cell Biol* **17**, 703–721 (2016).

42. Liu, L. F. & Wangtt, J. C. Supercoiling of the DNA template during transcription. *Proceedings of the National Academy of Sciences* **84**, 7024–7027 (1987).

43. Steen, R., Dahlberg, A. E., Lade, B. N., Studier, F. W. & Dunn, J. J. T7 RNA polymerase directed expression of the Escherichia coli rrnB operon. *EMBO J* **5**, 1099–1103 (1986).

44. Junker Jr, D. E., Rokeach, L. A., Ganea, D., Chiaramello, A. & Zyskind, J. W. Transcription termination within the Escherichia coil origin of DNA replication, oriC. *Molecular and General Genetics* **203**, 101–109 (1986).

45. Rokeachsb, L. A., Kassavetisll, G. A. & Zyskinds, J. W. RNA Polymerase Pauses in Vitro within the Escherichia coli Origin of Replication at the Same Sites Where Termination Occurs in Vivo. *Journal of Biological Chemistry* **262**, 7264–7272 (1957).

46. Kim, H., Park, Y. & Lee, J. B. Self-assembled Messenger RNA Nanoparticles (mRNA-NPs) for Efficient Gene Expression. *Sci Rep* **5**, 12737 (2015).

47. Lee, J. H. *et al.* Rolling circle transcription-based polymeric siRNA nanoparticles for tumor-targeted delivery. *Journal of Controlled Release* **263**, 29–38 (2017).

48. Hannan, A. J. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet* **19**, 286–298 (2018).

49. Pešović, J. *et al.* Repeat interruptions modify age at onset in myotonic dystrophy type 1 by stabilizing DMPK expansions in somatic cells. *Front Genet* **9**, 1–14 (2018).

50. Chi, L. M. & Lam, S. L. Structural roles of CTG repeats in slippage expansion during DNA replication. *Nucleic Acids Res* **33**, 1604–1617 (2005).

51. Matsuo, M., Yokogawa, T., Nishikawa, K., Watanabe, K. & Okada, N. Highly specific and efficient cleavage of squid tRNA(Lys) catalyzed by magnesium ions. *Journal of Biological Chemistry* **270**, 10097–10104 (1995).

52. Yang, W. Nucleases: Diversity of structure, function and mechanism. *Q Rev Biophys* **44**, 1–93 (2011).

53. Bell, N. A. W. & Keyser, U. F. Digitally encoded DNA nanostructures for multiplexed, single-molecule protein sensing with nanopores. *Nat Nanotechnol* **11**, 645–651 (2016).