SURVEY AND SUMMARY

Maximizing the potential of genomic and transcriptomic studies by nanopore sequencing

Daria Meyer [1,2,*], Winfried Göttsch [1,3,*], Jannes Spannenberg [1,*], Patrick Bohn [4], Bettina Stieber [2], Sebastian Krautwurst [1,5], Christian Höner zu Siederdissen [1], Akash Srivastava [1,3], Milena Zarkovic [1,3], Damian Wollny [1,3,6,#], and Manja Marz [1,3,5,7−9,#]

[1]RNA Bioinformatics and High-Throughput Analysis, Friedrich Schiller University, Leutragraben 1, Jena, 07743, Germany

[2]Oncgnostics GmbH, Löbstedter Str. 41, Jena, 07749, Germany

[3]Leibniz Institute for Age Research, Beutenbergstraße 11, Jena, 07743, Germany

[4]Helmholtz Institute for RNA-based Infection Research, HIRI, Josef-Schneider-Straße 2/D15, Würzburg, 97080, Germany

[5]European Virus Bioinformatics Center, Friedrich Schiller University, Leutragraben 1, Jena, 07743, Germany

[6]Max Planck Institute for Evolutionary Anthropology, Deutscher Pl. 6, Leipzig, 04103, Germany

[7]German Center for Integrative Biodiversity Research (iDiv), Puschstraße 4, Leipzig, 04103, Germany

[8]Michael Stifel Center Jena, Friedrich Schiller University, Ernst-Abbe-Platz 2, Jena, 07743, Germany

[9]Cluster of Excellence Balance of the Microverse, Friedrich Schiller University, Fürstengraben 1, Jena, 07743, Germany

*These authors contributed equally.

#These authors contributed equally.

ABSTRACT

Nucleic acid sequencing is the process of identifying the sequence of DNA or RNA, with DNA used for genomes and RNA for transcriptomes. Deciphering this information has the potential to greatly advance our understanding of genomic features and cellular functions. In comparison to other available sequencing methods, nanopore sequencing stands out due to its unique advantages of processing long nucleic acid strands in real time, within a small portable device, enabling the rapid analysis of samples in diverse settings. Evolving over the past decade, nanopore sequencing remains in a state of ongoing development and refinement, resulting in persistent challenges in protocols and technology. This article employs an interdisciplinary approach, evaluating experimental and computational methods to address critical gaps in our understanding in order to maximise the information gain from this advancing technology. We present a robust analysis of all aspects of nanopore sequencing by providing statistically supported insights, thus aiming to provide comprehensive guidelines for the diverse challenges that frequently impede optimal experimental outcomes.

Here we present a robust analysis, bridging the gap by providing statistically supported insights into genomic and transcriptomic studies, providing fresh perspectives on sequencing.

INTRODUCTION

In the era of rapid advancements in genomic technologies, engaging in the exploration of genomic and transcriptomic studies is very timely and important. The sequencing of RNA and DNA has a longstanding tradition in the field of biology and has undergone several significant advancements. A milestone in the field was the transition to high-throughput next-generation sequencing in the beginning of this century (1), allowing great improvements in both transcriptomics and genomics.

During the last decade, significant breakthroughs were achieved: (1) The most common sequencing method, Illumina sequencing, offers with its short-read technology a high accuracy but struggles with repetitive regions and long-range information due to read length limitations; (2) PacBio Sequencing is one of the first Single-Molecule Real-Time (SMRT) technologies and therefore known for long-reads adressing a major limitation of Illumia sequencing.

(3) After the debut of the MinION device of Oxford Nanopore Technologies (ONT), Nanopore sequencing emerges as a game-changer in understanding genomic and transcriptomic landscapes across different species. Ultra-long reads are produced in real-time (2). This allows for comprehensive coverage of repetitive regions and enables therewith the detection of structural variations. Notably, the usage of Nanopore sequencing for direct RNA analysis has seen a recent upswing, encompassing tasks such as *de novo* transcriptome assembly, isoform expression quantification (3), and the direct detection of RNA modifications (4). Today, it stands as a well-established method across various DNA and RNA sequencing applications (5) including diagnostics (6) and metagenomic assemblies (7). Comparative analyses of PacBio and nanopore sequencing are already available (3, 8).

Nanopore sequencing offers several distinct advantages over short-read sequencing technologies: (i) Nanopores sequencing allows for rapid data generation, facilitated by swift library preparation and real-time data acquisition during sequencing (9); (ii) it has the unique capability to directly sequence RNA without the need for reverse transcription or amplification (3); (iii) N anopore s equencing p ermits the sequencing of fragments up to 2 Mb in length in a single read (2), (iv) The portability and affordability of the MinION device have made nanopore sequencing a valuable tool for monitoring viral outbreaks (10).

However, it is important to mention the initial disadvantages of Nanopore sequencing, including lower accuracy and throughput, which have improved significantly over time (8). Nevertheless, it's worth noting that nanopore sequencing still exhibits a higher error rate compared to Illumina sequencing (5).

While numerous reviews provide comprehensive overviews of various tools for standard nanopore sequencing applications (9, 11, 12, 13), given the relative novelty of the field, t here i s a w ide a rray o f n on-standard approaches available to maximize sample utility.

The versatility of nanopore sequencing, as it finds applications in diverse fields, l ends i tself t o customization to meet specific r equirements, i ncluding t he u se o f various flow c ells, l ibrary p reparation k its, s equencing buffers, and an extensive set of computational analysis tools. This wealth of options can pose a significant challenge, particularly for newcomers to the field, who may grapple with decisions regarding protocols, libraries, tools, and parameters. Furthermore, the community often makes claims without experimental validation, rendering them unquotable.

Our study, based on knowledge gained from over 200 sequencing runs (including both RNA and DNA), as well as insights derived from courses we conducted on sequencing and data analysis during the last decade, presents a robust analysis bridging the gap between statistically supported insights into sequencing outcomes and best practices. The diversity of our data, encompassing various species and employing different preparation and sequencing methods, provides a unique opportunity to explore and compare a wide range of aspects. We demonstrate how flow cell performance depends more on sample type than average read length, flow cell age, or the number of active pores at the start of a sequencing run. Additionally, we illustrate how flow cell washing and adaptive sampling can enhance output, offer guidance on effectively calling modified bases in RNA and DNA, and provide insights into normalizing raw signal data. These findings allow us to gain a clearer understanding of optimizing sequencing for tailored experiments while rendering forum opinions accessible and citable, serving as a resource for researchers in different fields using nanopore sequencing.

Additional data used in this review

Samples. The statistics made in this review is grounded in an extensive dataset comprising over 500 MinION sequencing runs, with a specific focus on more than 200 sequencing runs conducted using R9 MinION flow cells. These encompass 171 DNA sequencing runs and 33 RNA sequencing runs, as detailed in Tab. S1. The scope of this research encompasses a wide range of organisms, including 12 bacteria, 5 protists, 4 plants, 47 insects, 15 mice, and 49 humans, alongside 41 metagenomic samples, 28 virus samples, and 6 synthetic RNA sequences created *in vitro*. The study also acknowledges instances of sequencing failures and runs conducted on R10 and R8 or lower flow cells, although these are not included in Tab. S1 and the accompanying statistical analyses.

Flow cells. The flow cell characteristics are quantified based on the number of active channels, as reported in ONT run reports, which range from 76 to 512 out of a possible 512 channels. The count of active pores varies between 93 and 1765 out of the total of 2048 pores available on the flow cell. Furthermore, the age of the flow cells is measured from the time of their arrival in our laboratory. It's important to note that we lacked access to information regarding the production date of the flow cells. Consequently, we used flow cells within a time frame spanning from 3 days after arrival to as long as 249 days after arrival. This extended range of usage duration is noteworthy, especially considering that flow cells stored at 2-8°C have a maximum recommended shelf life of 90 days.

Sample types. The details regarding the sequenced samples encompass a wide spectrum of sample types, ranging from prokaryotes to eukaryotes, and including viruses and metagenomic samples. Tab. S1 provides additional information, indicating whether each run was an RNA or DNA sequencing run and specifying the quantity of input material loaded in nanograms. Notably, in certain sequencing runs, we deliberately reduced the input material to as low as 30 ng, while in other cases, we loaded a higher amount of 1428 ng of the library onto a flow cell. This variability in input material is a notable aspect of our sequencing experiments.

Output of sequencing runs. The sequencing run output includes several crucial metrics. Firstly, it comprises the estimated sequence bases as determined by MinKNOW, spanning a broad range from $7 \cdot 10^6$ to $28 \cdot 10^9$ bases per sequencing run, excluding washing steps. Secondly, it encompasses the mean read length, varying from 302 to 19,632 nt. Additionally, we assess the half-life of the flow cell, which is defined as the duration of sequencing until only half of the total number of initially active nanopore channels remain functional. The observed flow cell half-life values span from 0 to 60 hours, providing valuable insights into flow cell longevity.

For a comprehensive exploration of the factors affecting flow cell longevity, we have conducted detailed analyses,

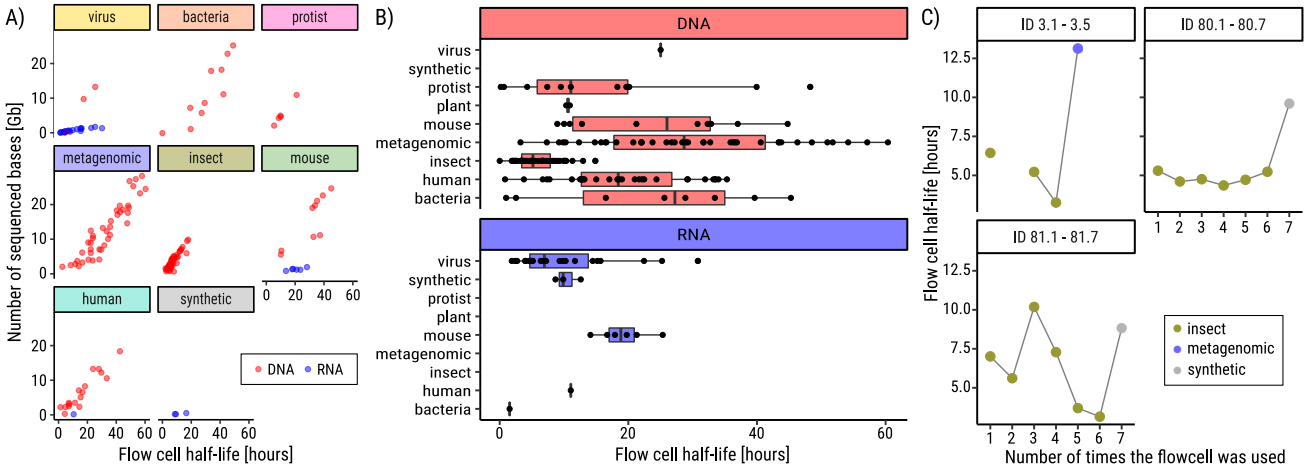

Figure (1). The sequencing yield depends on the flow cell half-life **A)**. This holds true for both DNA (red) and RNA (blue), though the effect is more pronounced for DNA. **B)** Flow cell half lives appear short especially for extremely high molecular weight (HMW) insect samples (cricket). **C)** The influence of sample type on flow cell half-life can be seen when different samples are sequenced on the same flow cell suggesting the differences not being flow cell specific.

which include the development of custom scripts. These analyses and scripts are available for reference and further investigation at the following `https://github.com/wollnylab/Nanopore_Seq_Metaanalysis/`. This resource serves as a valuable repository for those interested in delving deeper into the factors influencing flow cell performance and longevity.

Perspective on sequencing yield

Over the course of several years we have performed more than 200 nanopore RNA and DNA sequencing experiments of a large variety of species, see Tab. S1 and Fig. S1A. While examining the quality and quantity of the sequencing outputs, we noticed that the sequencing yield per experiment varies greatly across experiments (Fig. S1B).

Thus, we aimed to investigate potential reasons for this disparity by comparing various parameters characterizing the different sequencing runs. We have noticed that one important variable is how long the flow cell stays active over time. In order to quantify the activity as a single parameter for each experiment, we estimated the flow cell half-life for each run. We define the flow cell half-life as the sequencing time until only half of the starting nanopore channels are still active. Over the course of many sequencing runs we noticed that the flow cell half-life varies considerably for RNA as well as DNA runs, Fig. S1C.

The sequencing yield strongly depends on the flow cell half-life, see Fig. 1A. We observed this correlation irrespective of which organisms were sequenced.

Flow cell half-life depends on sample type. Hence, we tried to find experimental or sample-specific parameters that could explain the differences in flow cell half-lives. We found the flow cell half-life to depend strongly on the type of sample that is being sequenced, see Fig. 1A. Most notably, sequencing runs where insect samples were sequenced consistently displayed a shorter flow cell half-life compared to other samples types, see Fig. 1B. This indicates that genomic features might be an important contributor to the differences in flow cell half-lives observed across sequencing runs.

The amount of loaded input material has almost no effect on the half-life and output of the flow cell, see Fig. S2.

Interestingly, when changing the sample type while washing the flow cell, even after 7 washing steps a higher output can be observed when loading a 'better' sample type, see Fig. 1C.

Reusage of flow cells. The number of active pores and channels decreases over time of sequencing, see Fig. S1C. When the number of active pores decreases, the pores can be recovered by washing the flow cell with the "Flow Cell Wash Kit" (EXP-WSH003, EXP-WSH004) provided by ONT. The wash kit contains DNase I, which digests the library loaded on the membrane. As shown in Fig. 2C, by washing and reactivating inactive pores and channels, the resulting overall yield of the sequencing run can be increased.

In total we gained experience from washing flow cells 70 times. Flow cell ID 80 and 81, see Tab. S1, were washed most often with 6 wash steps per flow cell, however, more wash steps could have been possible, see below. Although washing increases the yield and amount of active pores in total, mostly it can be seen that the flow cell half life is shorter after washing, see Fig. 2B. Wash steps were performed for sequencing *Acetobacter*, *Baccharis*, *Coronavirus*, Cricket, *D. discoideum*, human genome, *in vitro* RNA, metagenomic data, and *Planctomycetes*, see Tab. S1.

Fig. 1C and Fig. 2B give an overview of the different flow cell half-lives before and after washing, when sequencing the same (Fig. 2B), or a different (Fig. 1C) sample after washing.

As described in the ONT guidelines, pore blocking can reduce the flow cell output. When a pore is blocked, the pore state changes from "single pore" to "unavailable". This state is reversable (either by a voltage reversal mechanism, which is done automatically during sequencing, or by washing the flow cell). Until today no explanation is given in the ONT guidelines why for some samples more pore blocking can be observed than in others.

Reusability and flow cell half-life. We investigated to which extend the reusage of flow cells influences their flow cell half-life, and subsequently the sequencing output. We have

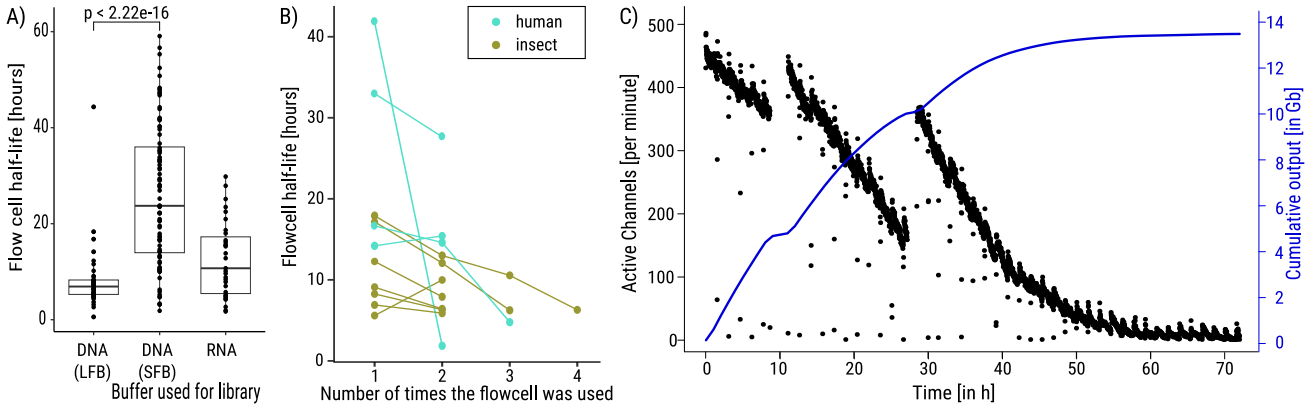

Figure (2). A) Samples for which LFB (long fragement buffer) has been used in the library preparation show a reduced flow cell half-life compared to those with SFB (short fragment buffer). During the library preparation of RNA a different buffer is used. T-test was used to calculate the p-value. **B)** The flow cell half-life depends on how often the flow cell has been washed and reused. Different sequencing runs of the same sample type (insect or human, DNA) are shown. **C)** Influence of washing the flow cell on the amount of active channels. In this arbitrary example, the human DNA Sample (ID 111.1-111.3) has been washed twice during sequencing (after 10 h and 30 h). After each wash steps the number of active channels increased compared to directly before washing. Blue – cumulative output in Gb. In this example we have performed the first washing step too early, whereas the time point of the second washing step has been chosen well.

analysed more than 70 wash steps for flow cells that were reused multiple times. We found that, in the majority of cases re-loading with the same type of sample flow cell half-lives decrease when flow cells were reused, see Fig. 1B. This is irrespective of the decrease of the number of active pores as a consequence of flow cell re-use. Interestingly, we found that this trend is also strongly sample type specific. Flow cells that showed steady decrease in flow cell half-life for insect samples show drastic changes in flow cell half-life once other sample types are loaded Fig. 1C. This finding supports the notion that sample-specific genomic features might to a large extent underlie the observed differences in flow cell half-lives across sequencing runs.

A higher flow cell half-life with SFB We observed a clearly higher half-life and therefore total output when using short fragment buffer (SFB) instead of long fragment buffer (LFB), see during Fig. 2A. Runs in which LFB was used during the library preparation show a significantly ($p < 2.22 \cdot 10^{-16}$ in two-sided t-test) shorter flow cell half-life compared to those samples for which SFB has been used.

Flow cell age, mean read length, active pores and channels are independent of flow cell half-life. In contrast, all other parameters that we investigated showed no clear correlation with the half-life of the flow cells. Notably, the age of the flow cell, Fig. 3A, defined as the time between flow cell delivery and usage, does not influence the half-life of the flow cell. We used flow cells older than 200 days (maximum 249 days) with output comparable to novel flow cells, especially when the number of active pores at the start of the sequencing run are high. This indicates that flow cells seem to be stable for up to 200 days, when stored according to manufacturer's specifications.

Interestingly, also the mean read length (Fig. 3B), the number of active channels (Fig. 3C) or pores (Fig. 3D) at start of the sequencing have no influence to the flow cell half-life.

Yet, we obviously cannot exclude hidden variables such as manufacturing differences between flow cells which could potentially also have a strong impact on the half-life of flow cells.

Perspective on library preparation

DNA and RNA: How to get long reads In order to get reads as long as possible, the DNA or RNA should not be sheared. The longest DNA read we obtained was 1.4 Mb (sample 73.1). This read originated from a metagenomic sample and classifies as *H. sapiens* by `kraken2` (14), with a GC-content of 41.4%. The longest RNA read contained 175 Kb (sample 103.1) and classifies as bacteria (no species associated) by `kraken2`, interestingly, with a GC-content of 68.6%.

Vortexing seems to be way too rough and results in shorter fragments (15). The application of suction force during pipetting can cause shearing of nucleic acids. To obtain longer reads, it is recommended to use tapping for mixing and employ cut-tips for pipetting, while also pipetting at a slow pace (16). Most of the DNA isolation kits are column or particle-based. It as well shears and fragments the nucleic acid strands. Hence, we advocate the use of the classic phenol-chloroform method. Recently, stress has been laid on extracting high molecular weight (HMW) DNA. Different brands developed innovative ideas like silica lamella surface topography and solid phase methods for extracting HMW DNA. Such kits can be preferred at a higher cost of reagents for the ease-of-use (17, 18).

In summary: *Based on our experience, we recommend 4 changes compared to a standard library preparation: (1) do not vortex the samples; (2) pipette as slowly as possible; (3) use cut tips; (4) use phenol-chloroform extraction.*

Low input material possible In general, the rule is: the more input material, the better. However, the yield of DNA/RNA is not always high in some samples. This part focuses on small sample quantities that should not be amplified.

For DNA 1000 ng (100-200 fmol, SQK-LSK114) or if working with long DNA fragments, 100 ng of HMW genomic DNA is recommended. Usually one gains after the library preparation about 50 % of the input DNA for loading onto the flow cell. It has been reported that 1 ng of DNA input yielded 6118 reads with N50 of 3907 nt (19). The approach of carrier sequencing has allowed Mojarro *et al.* to detect down to 0.2 ng of *B. subtilis* DNA prepared with 1000 ng of Lambda DNA using MinION without amplification (20). By use of

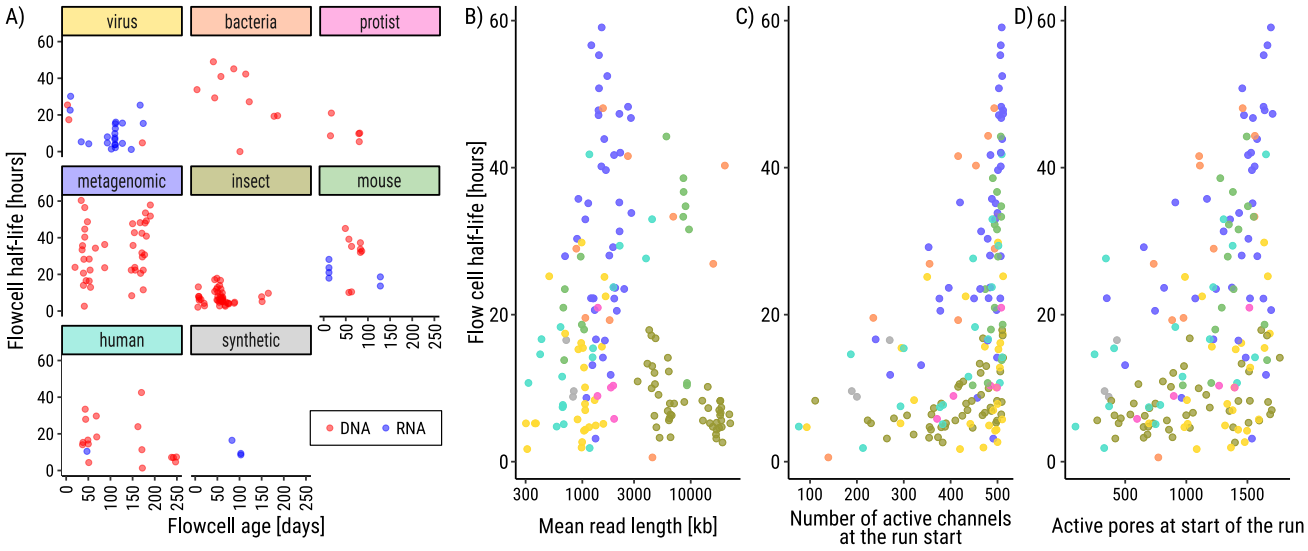

Figure (3). A) The age of the flow cell does not influence the flow cell half-life. The mean read length **B)**, number of active channels **C)**, and active pores **D)** do not influence the flow cell half-life. Color corresponds to sample type as in A).

10X Genomics Chromium Controller input as low as 50 pg has been reported to be usable to realize long read nanopore sequencing with ultralow input of genomic DNA. (21)

In our hands, we were able to succesfully load 30 ng DNA onto a flow cell (before library preparation 435 ng, ID 36.6) and we produced 648 MB data. In another run we loaded 50 ng onto the flow cell (ID 85.2), which resulted luckily in even 1,5 GB data output.

When sequencing direct RNA, a poly-adenylated 3' end is necessary for adapter ligation. The vast majority of RNAs in an arbitrary cell consists of rRNAs (e.g. in *E. coli* 85% (22), which typically have no poly-A tail. Therefore, a poly-A-based sequencing approach would naturally discard the usually unpleassant rRNAs during sequencing. However, not all mRNAs have adenines added to their 3' end as performed typically by e.g. polymerase II, as for example histone-coding mRNAs (23). On the other hand the majority of polymerase III based transcripts (most of ncRNAs, non-canonical derived miRNAs and some lncRNAs) do not have a poly-A tail (24).

ONT suggests to start with 500 ng total RNA, or – in agreement with the ratio of rRNAs in the cell – about 50 ng poly-A-tailed RNA to obtain the poly-adenylated RNA fraction of the total RNAs. However, if you are interested all RNAs but rRNAs, you should ligate poly-A to any RNA in your sample. In case you are particularly not interested in rRNAs, then an rRNA depletion step (25) should be performed before the ligation of poly-A.

For total RNA a method has been reported, which needs only 10 ng input material (26). In our hands we could produce 157 mb with only 42 ng RNA (ID 108.1) input.

In summary: *It is possible to gain reasonable output with 30 ng DNA or 40 ng RNA or lower.*

Abstain from Spike-in The DNA ligation kit (SQK-LSK114) and the direct RNA kit (SQK-RNA002) contain positive control strands called DNA control strand (DCS) and RNA calibrate strand (RCS), respectively. ONT describes DCS as a 3.6 kb standard amplicon mapping the 3' end of the Lambda virus genome at 10 ng/μL, see Fig. 4, whereas the

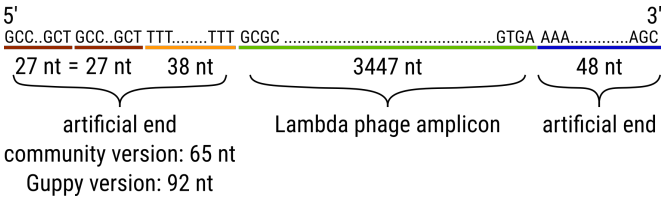

Figure (4). Structure of the ONT DNA spike-in: The positive control spike-in for DNA sequencing is a 3.6 kb amplicon, which maps to the 3' end of the Lambda phage genome. The 5' artificial end consists of 65 nts in the community version; in `Guppy` version v6.4.2+97a7f06 the first 27 nts are repeated.

RNA CS (RCS) is the Enolase II (ENO2, YHR174W) from *Saccharomyces* genome of strain S288C, at a concentration of 50 ng/μL. They both serve as a positive control in respective library preparations and are not used for normalization.

We could find about 1 % of DCS/RCS in our samples. We opted to exclude the positive control entirely for the majority of samples. Our rationale behind this decision was to maximize the utilization of the flow cell's sequencing capacity solely for our target sample. ***In summary:*** *Reducing the amount of DCS/RCS leads to an absolutely higher number of sequenced sample of interest.*

Adapter ligation The last step of any library preparation is adapter ligation. Adapters are oligos attached to motor proteins that unwind the double helix. Thus, it is crucial to ligate sample material to the adapters with high efficiency. The protocol states, use of double the amount of T4 quick ligase compared to any other ordinary reaction along with 5 μl of adapter mix (AMX-F or LA or 6 μl RMX or AMII or NA). The concentration of the adapter mix is proprietary information. The protocol suggests 1 μg (or 100-200 fmol) HMW DNA or 50 ng of poly(A)-tailed RNA or 500 ng of total RNA. Thus the amount of adapters in the mix is standardized for a reaction with 200 fmol. We observed whenever we had very HMW DNA less adapters were needed due to fewer fragment ends in the sample. We think the increased number of free adapters

might compete for active pores and thus less output could be observed.

NEB suggests a 10:1 ratio in their general protocol for adapter ligation, an excess of adapters is needed but should not be too high. ONT describes no issues with high proportions of adapters, as long as the majority of pores are actively sequencing. For RNA fragments a potential influence of RNA modifications on the efficiency of adapter ligation has been reported (3).

In summary: *For very HMW DNA, we recommend the usage of different amounts of adapter mix, based on the observed amount of free adapters during sequencing.*

Loading amount of DNA based on the quantity of available pores Depending on the quantity of DNA, the choice of the amount of material to be loaded can be significant. Hence, one should load everything but an appropriate amount of DNA on the flow cell. For flow cells with more than 1500 active pores, we observed 350 ng of DNA was sufficient. Similarly, for about 1200 active pores or 800 active pores, we load 300 ng or 250 ng, respectively. From our data in Tab. S1, one can see that the amount of data generated is directly proportional to the number of active pores at the start and increases to a certain threshold with the increase in loading amount, after which it reaches a level of saturation.

In summary: *If the amount of available pores is low, the sequencing run is not expected to be bad, but produces linearly less output (Fig. S3) — a bacterial or virus sample can be optimal for such a run instead of a flongle. One can load linearly less material, however one can not load too much.*

Amplicon sequencing If you aim to build a consensus genome, e.g. for viruses if the viral titer is very low, PCR amplicon sequencing can be an optimal solution. First, efficient primers for the amplicons that span through the whole genome need to be designed. Identification of the specific virus strain is crucial, as show exemplarily at Plum pox virus (PPV) and its diversity in Fig. 5A.

We recommend to use amplicons, longer than 1000 bases and with an overlap of 200 bases between consecutive amplicons. The `ARTIC` protocol for SARS-CoV-2 produces amplicons of 392 bp in length and with 90 bp overlap, however we had good experience with longer fragments (1200 bp and 117 bp overlap) for SARS-CoV-2 (27).

After multiplex PCR amplification and sequencing, a uniform coverage of the whole genome can be observed, see Fig. 5B. A characteristic feature to note is the increased representation of the amplicons overlapping parts.

The community and published literature suggest mixing the amplicon in equal nanomolar concentrations (28, 29, 30). Since all the amplicons were 1000 bp in size, we performed size selection. After PCR, different samples could be equally pooled via mass concentration (nanogram per microliter) instead of molar concentration of solution. With this method the coverage is high enough to pinpoint mutations.

In summary: *Amplicon sequencing can be performed with amplicons longer than 1000 nt. We suggest mixing the amplicons in equal nanomolar concentrations.*

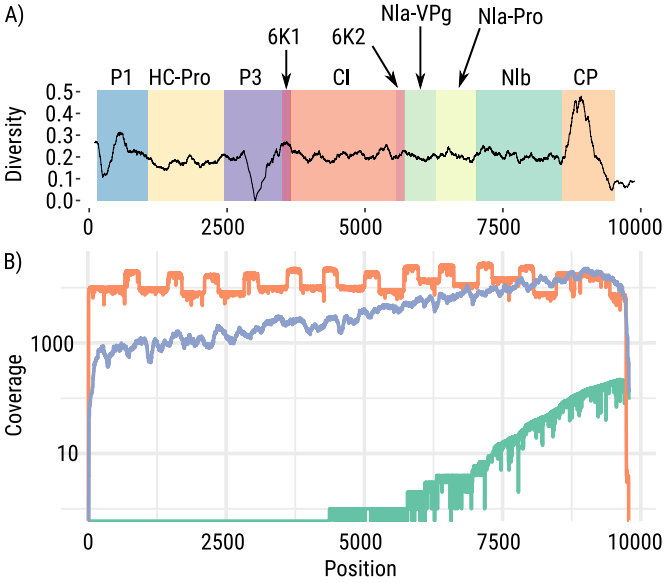

Figure (5). A) Genetic diversity for each position in the PPV genome across 109 PPV isolates of all known PPV strains. Note that the capsid protein (CP) encoding region of the plum pox virus (PPV) genome displays a significantly higher diversity than other genomic regions. The drop of diversity within the P3 region is caused by an insertion in a single PPV sequence. **B)** Coverage comparison of three sequencing methods: direct RNA sequencing using nanopore (green), Illumina RNA-Seq (purple), and nanopore-based PCR amplicon sequencing (orange) at the example of PPV. The overrepresented 3' ends are based on sequencing only RNA fragments with intact poly(A) tails for the direct RNA sequencing approach.

Perspective on ONT sequencing

Short reads are sequenced first? To answer the question if short reads are sequenced first, we analyzed all of our samples as shown examplarily for four samples in Fig. 6A. When analyzing read lengths over 72 h run time, we observe very different patterns. For ID 25.1 (virome), and ID 38.1 and ID 40.1 (fungi on LFB and SFB) we observe almost equal length distributions over time. For sequencing run ID 96.1 (mouse genome) we can see clearly a preference for slightly shorter fragments in the beginning and towards the end of the sequencing run. Further details can be seen, when analyzing the first hours of a sequencing run. We observe different patterns: We sequenced *Dictyostelium discoideum* with LFB (Fig. 6B) and SFB (Fig. 6C) library preparation, respectively, to analyze the effect. As expected we sequenced in the first 30 min shorter fragments with SFB, however we can also observe a clear preference for shorter reads to be sequenced first.

This effect can not be drawn back to the buffer only, as the sample properties play also an important role: The DNA virome sequencing run (ID 25.1, Fig. 6D) shows no significant read length bias towards the first hour, whereas for the mouse genome (ID 96.1, Fig. 6E) a clear preference for shorter reads to be sequenced first can be observed. Notably, Fig. 6D and E, both contain DCS 1:10 diluted.

Towards the end of the sequencing run, the variance of read length is higher, because fewer reads are sequenced.

In summary: *The community reported small fragments to be sequenced mainly in the beginning. This observation can not be generalized.*

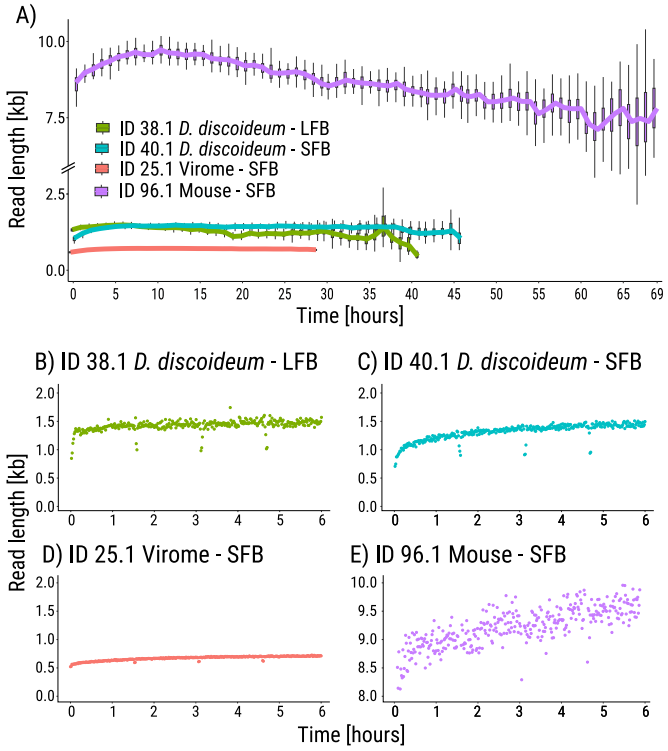

Figure (6). Short reads are sequenced first? **A)** Read length distribution of four samples over 72 h: ID 38.1 (DNA *Dictyostelium discoideum* – LFB), ID 40.1 (DNA *Dictyostelium discoideum* – SFB), ID 25.1 (DNA virome), and ID 96.1 (DNA mouse). **B-E)** Read length distribution of the same four samples over the first 6 h. In B), D) no significant changes are observable, whereas in C) and E) a clear observation for a preference of small reads can be detected. Note the different scales on the y-axes.

For washing flow cells, we consider the most important parameters to be (i) the time point of washing; (ii) the amount of wash steps; and (iii) the splitting of the library.

Time point of washing. We found a good measure for the time point is the amount of active pores. We have discovered a useful guideline regarding when to perform the next wash step: when approximately 35% to 40% of the pores are still active from the beginning of the run or after the last wash step.

Alternatively, the cumulative output over time plot from MinKNOW can be used as a guideline: when the curve starts to flatten, see Fig. 2C - blue line, and the increase of data is not linear anymore, this can be a sign for the next washing step.

In summary: *Perform the next wash step when approximately 35% to 40% of the pores remain active, either from the beginning of the run or after the last wash step.*

Amount of wash steps. The advice from ONT is to wash a flow cell up to 4 times. Here we show that flow cells can be washed at least six times, still producing considerable amount of data. The ideal number of wash steps depends on the materials (DNA or RNA), the flow cell performance (flow cell half-life) and is limited by the amount of starting material, which is available. The cost of washing plays a negligible role in this context (∼15 € per wash step; currently the Wash Kit EXP-WSH004 costs 95 € and can be used for six wash steps). We consider a wash step successful, if at least 50% of pores are active after washing (compared to the beginning of the run or wash step).

In summary: *Although ONT recommends to wash a flow cell up to four times, we observed very good results even after six wash steps (assuming sufficient input material).*

Splitting the library. If enough input material exists, it is possible to perform one library preparation. To keep the molarity of adapters for more than 1.5 μg input material, several libraries can be prepared following the same procedure. The obtained input material for sequencing should then be split for reloading the flow cell. As the number of active pores decreases over time, see Fig. **??**, it is advisable to load the highest amount of library in the beginning, and decrease the amount of loaded library after each wash step. If planning to wash twice, a good distribution might be 50%, 30% and 20%. Nevertheless the number of wash steps also depends on the amount of DNA in the library. To our experience it is not advisable to load less than 50 ng due to the very low obtained yield (see sample IDs 108.1 or 36.6 with 42 and 30 ng input material and 157 and 648 bases yield, respectively).

Noteworthy, before washing the run can be paused or stopped. If the run is stopped a new file will be generated, whereas in case of pause the output of the next load will be written into the same file. That becomes important if the washing method is used to sequence different samples on one flow cell without barcoding.

In summary: *We recommend applying the library in a logarithmic scale across the various wash steps, rather than in equal portions.*

Flongle dRNA sequencing As of 01.07.2023, the Flongle costs 1/7 and produces 1/17 output compared to MinION Flow cell. Flongles are particularly useful, when barcoding strategies fail, e.g. for direct RNA sequencing or when small genomic units, such as RNA viruses are sequenced. A community-developed protocol is available at protocols.io. Noteworthy, several community barcoding strategies for RNA sequencing have been developed, but their performance is not great yet (relatively high data loss and misclassification rates. Despite we lost 30% of the data in our hands, we still misclassified 5-10% between different barcodes (data not shown).

The most critical step for a successful Flongle sequencing run is priming, as accidentally introduced air bubbles rapidly results in pore loss, and the combined priming and loading port is located directly above the pore array. To lower the risk of pore loss an alternative loading method that uses pressure differential to slowly push out air bubbles and then pull the priming liquid into the flow cell was developed by Graham Wiley.

The direct RNA sequencing protocol includes an ”optional” cDNA synthesis step that may increase throughput and read quality by unfolding RNA structures that could have stalled the motor protein. However, the long incubation at an elevated temperature in the presence of divalent cations, and potential residual RNase H activity of retrovirus-derived reverse transcriptases such as SuperScript IV may adversely lead to RNA cleavage (31), thus reducing the number of

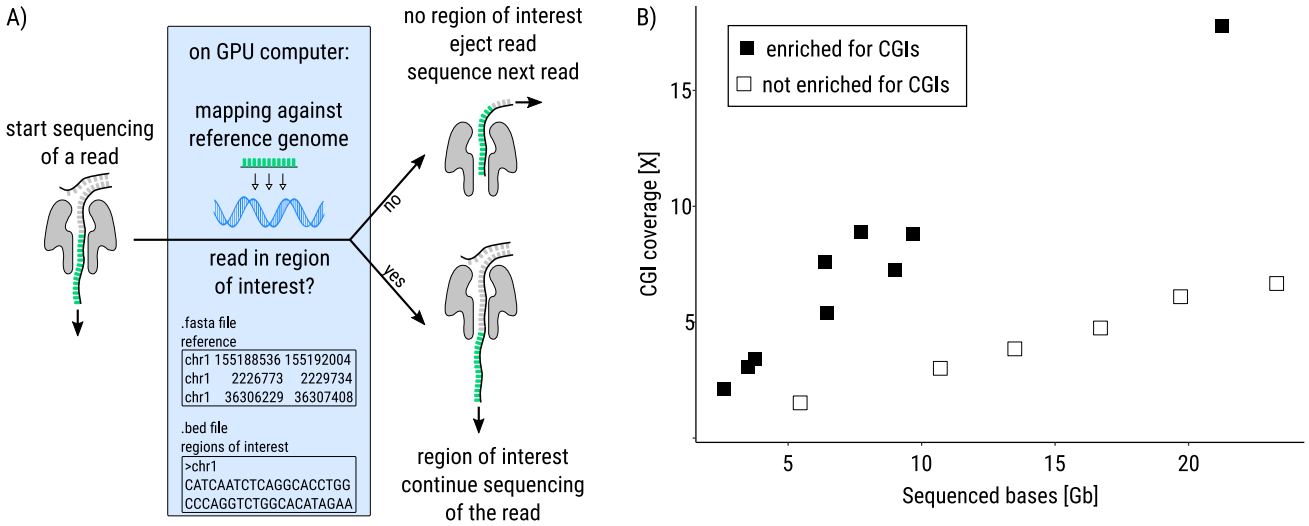

Figure (7). The basic principle of adaptive sampling **A)**. For each read, the first ∼180 bases are mapped against a given reference and are continuously sequenced only if they align to a region of interest (32). Rejecting the read takes about 0.5 s additional time is needed to capture the next read . About 450 bases are sequenced per second for R9.4 Nanopore flow cells (5). About 450 bases are sequenced per second for R9.4 Nanopore flow cells (5). Every rejection takes about 0.5s plus the additional time the pore needs to capture another read . **B)** Using adaptive sampling to enrich for more than 30.000 CpG islands (CGIs, 0.74 % of the human genome) on the human genome increases the coverage on the CGIs. For the enrichment, the CGIs were extended for 2000 nt on each side.

full length RNA molecules to be sequenced. Therefore, other reverse transcriptases with the ability to unfold RNA structures at lower incubation temperatures and lacking all RNase H activity, for example the group II intron-derived reverse transcriptases Induro RT and Marathon-RT, may be preferred when performing reverse transcription for dRNA sequencing.

In summary: *Flongles are usefull especially for sequencing RNA viruses. To sequence full length RNA transcripts, we propose not to convert the RNA to cDNA.*

The pros and cons of adaptive sampling Adaptive Sampling can be used, to enrich the yield of predefined input material, e.g. increase the coverage for a specific genomic area, or increase defined species in metagenomic samples (32). Fig. 7A shows the basic principle of adaptive sampling, during which the first ∼180 nt of each read are sequenced to make decision whether the sequencing should be continued or not (32). During adaptive sampling, many of the long fragments are only partially sequenced, resulting in a high number of short reads, a lower N50 and median read length (only about 500 nt). The average read length of reads mapping to the region of interest is not affected and keeps long sequenced reads. To utilize adaptive sampling, a GPU is required for the sequencing computer to enable real-time basecalling during the sequencing process. It is essential to use a fast basecalling model specifically designed for adaptive sampling. Once the sequencing run is completed, the data can be basecalled again using a higher accuracy model to enhance the quality.

As described by the ONT community, the coverage can be enriched 5-10 fold, however, we show based on CpG islands in the human genome up to ∼24 X coverage on CpG islands on a MinION flow cell, especially when less than 10 % of the genome are targeted, with up to 6000 target regions and with extending the regions of interest by a so called buffer region. When we tried to increase the coverage on ∼30.000 CpG islands in the human genome (sample ID 109.1 - 113.3, 116.1 - 122.1, 124.1, 127.1, 129.1 - 133.2), we found that the coverage could be increased, when extending the on average 777 nt long CpG islands for another 2.000 nt on both sides, see Fig. 7B.

In summary: *Adaptive sampling can also be used on the MinION for human genome, even for more than 30.000 target regions of 775 nt in length. We recommend to add 2000 nt on each side for an effective process.*

Perspective on computational analysis

Established standard methods for the bioinformatic analysis of ONT sequencing data, as e.g. depicted in Fig. 8 A), are often modular and can be replaced with different components that can be operated by non-informaticians.

In a first step the raw signals, called squiggles, have to be converted into read sequences. The raw signals are subjected to basecalling using tools, commonly the tool `Guppy` from ONT. However, alternative tools can be used: `Dorado` from ONT is still under development but already works on `pod5`, the new data format, `Remora` from ONT can base call for modified/methylated bases and prepare datasets to train basecalling models for modifications. And `bonito` from ONT can be used to train novel basecalling models to call canonical bases. A quality control step should always be performed after the basecalling phase. Tools such as `pycoQC` (33), `NanoPlot` (34), `nanoQC`, and `longQC` (35) are ready for usage. `PycoQC` provides the broadest overview with an interactive and comprehensive view of the data, including read quality, length, coverage, active channels, quality over sequencing time, and much more all in one HTML file. If for a certain experiment specific length or quality requirements exist, then you may apply the tools `Filtlong` or `chopper`, both part of the `NanoPack` (34).

The sequenced reads are then further used for mapping, assembly or other applications. For mapping commonly the tool `minimap2` (36) is used, but may also be

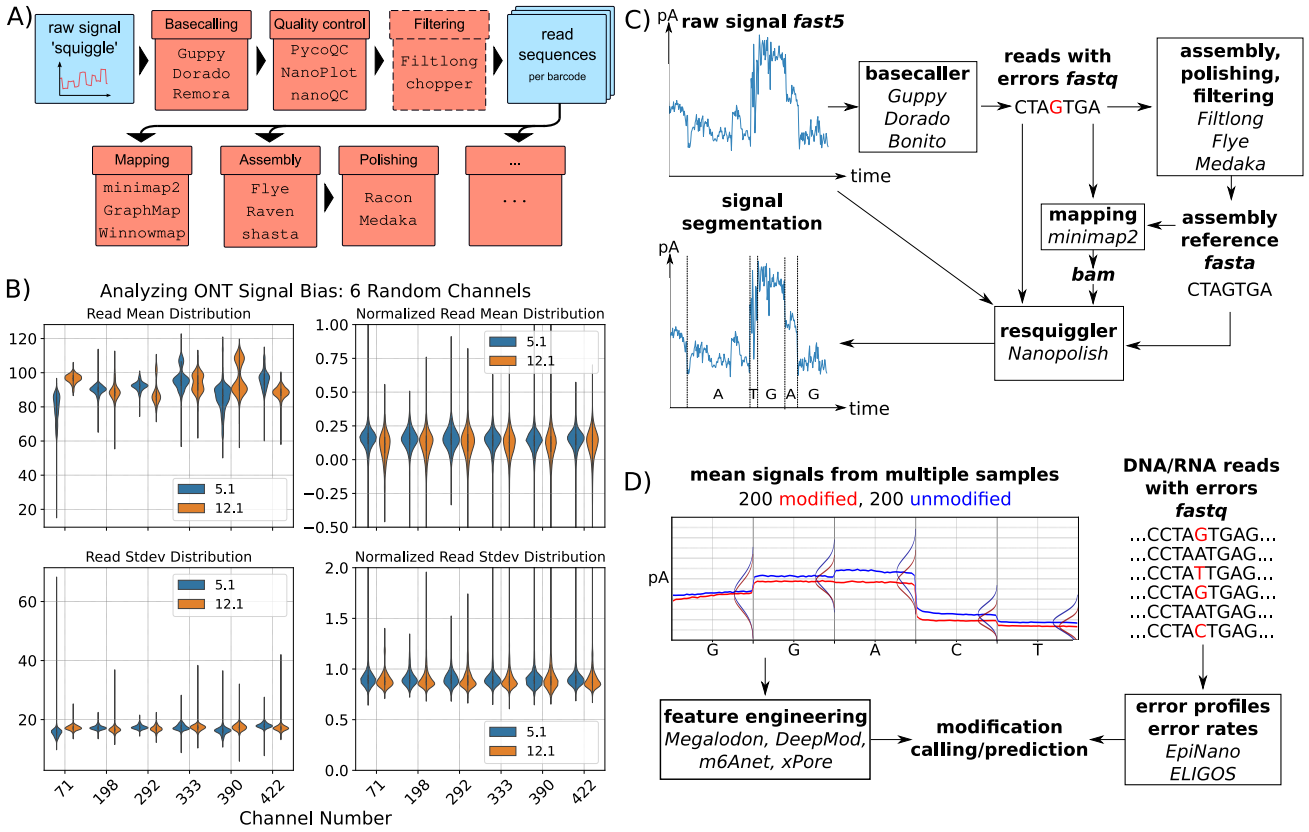

Figure (8). A) A potential workflow for nanopore sequencing data involves basecalling of the raw data, followed by quality control and filtering. The reads can then be analyzed with specialized long-read tools for mapping, assembly, or other workflows. **B)** Read mean distribution (top) and read standard deviation (bottom) before (left) and after (right) normalization of six random picked channels from flow cells of sequencing runs 5.1 and 12.1. Bias caused by pores and sensors, or flow cells vanish after normalization, see Eq. **2**. This bias is majorly observed in the signal mean. **C)** Overview of Signal Processing and Modification Calling. The raw ONT signal can be segmented and corrected for basecalling errors by resquigglers using basecalled (.fastq), mapped (.bam) reads, and a reference sequence (.fasta). **D)** Modifications shift the ONT signal, which machine learning models can detect. The red and blue lines represent the mean signal of 200 reads respectively, while the vertical bulges indicate the signal distribution per base following a normal distribution. These shifts can lead to error patterns in the reads which indicate the presence of a modification.

replaced by `GraphMap` (37) or the 2020 published tool `Winnowmap` (38). Chirag Jain *et al.* (38) show that `Winnomap` performs better than `minimap2` when mapping simulated PacBio and real ONT Human data. For assembly the common tool is `Flye` (39), which can also be substituted by `NextDenovo` (40) or `shasta` (41). `Shasta` allows to assemble long reads very fast (42). `Flye` takes longer to assemble long reads, but shows an overall good performance to create long contigs with low errors compared to many other assemblers in case of bacteria (43). `NextDenovo` (40) is a newer assembler that is able to assemble longer contigs than `Flye` in case of Mollusca (42).

Afterwards, a polishing step by `Racon` (44) or `Medaka` from ONT can be used to reduce errors in assemblies. `Racon` shows a good performance in reducing single nucleotide variants and insertion errors, while `Medaka` is better in reducing deletion errors (43). Both can also be used in combination.

The following section aims to describe, beyond the standard approaches, the theoretical possibilities ('hacks') of what can be achieved with the MinION output.

In summary: *Instead of relying solely on a standard pipeline, it is advisable to carefully select the tool that best matches specific needs.*

Working with raw data The internal data structure of `fast5` raw signal data is not officially documented. Gamaarachchi *et al.* recently published a detailed document that explains the details of the format. Here we give an overview for advanced users that want to access the raw signal data and metadata directly.

To access the raw `fast5` files, we recommend to use the Python package `h5py`. Since `fast5` v2.3 ONT changed the data compression algorithm to `vbz` compression. The plugin `ont-vbz-hdf-plugin` is needed to access the data. Reads are stored in batches of usually 4 000 per `fast5` file. The `sequencing_summary.txt` file provides information on which `fast5` file contains which reads. A `fast5` file is structured like a directory. Data groups are separated by `'/'`. Data is stored as datasets and attributes within the `fast5` file. Reads can be found in the root directory of each `fast5` file. The `fast5` structure follows the following scheme:

```
- 'read_<readid>'
    - Raw/Signal
    - channel_id
    - context_tags
    - tracking_id
```

The raw signals can be found as numpy arrays under `read_<readid>/Raw/Signal`. Additional metadata can be found in `channel_id`, `context_tags` and `tracking_id`. The `channel_id` contains information about the channel number of the sequenced read or parameters like digitisation, offset and range. In `context_tags` data about the sequencing run, like the used sequencing kit or the use of barcodes, is stored. The `tracking_id` contains technical data about the used devices and software, like the flowcell i d, s tart t ime o f t he s equencing r un a nd software versions. To save space, the signal is stored as integer values and must to be converted back to the pico Ampere (pA) signal with the following equation:

$$\text{pA\_signal} = (\text{signal} + \text{offset}) * \text{range} / \text{digitization} \quad (1)$$

The offset, range and digitization values can be found as attributes in `read_<readid>/channel_id`.

Other file formats such as `slow5` or `pod5` have different data structures than `fast5`, so the access paths and methods do not match between them and their programming interfaces (APIs). With our wrapper `read5`, we standardized and unified the access to the data. Methods to extract and normalize the raw data are also provided within the package. For further analysis and comparison between different sequencing runs, normalization of the pA_signal is needed.

In summary: *Analyze the raw ONT data in your own way to gain more information than just FASTQs, such as channel number, normalization parameters or the sequencing start time of the read. Investigate for significant signals that hint to e.g. specific motifs, barcodes, or modifications (see below).*

Sequencing Biases Require Normalization. Each read must be normalized, as various factors influence and bias the signal. Tools from ONT like Guppy normalize the data for you before processing it. When working with the raw ONT data, make sure to normalize for biases from the pore and the sensor. The most prominent factors are the flowcell itself and the pores together with the sensor within each sequencing channel on the flowcell, Fig. 8 B). We need to normalize for the sensors and pores to be able to compare reads and the raw signal from different experiments. For the signal normalization, ONT uses the median and the median absolute deviation (mad):

$$\text{norm\_signal} = \frac{\text{pA\_signal} - \text{median}(\text{pA\_signal})}{\text{mad}(\text{pA\_signal})} \quad (2)$$

After normalization, the distributions of the mean sequencing signals across different channels equalize, and biases disappear, see Fig. 8 B).

In summary: *Always normalize each ONT signal individually to reduce pore, sensor, and flowcell biases.*

ONT Signal Processing and Segmentation When the motor protein at the pore rotates then one nucleotide is released on the lower side of the membrane from the pore and one nucleotide is inserted into the pore from the upper side of the membrane. The time between two rotations is called dwell time. The motor protein does not rotate at constant time points, but rather processes nucleotides irregularly depending on the specific nucleotides at the motor protein (45). Therefore, for the raw signal processing, signal segmentation is necessary. Previous version of `Guppy` (up to v6.3.2) provided additional data when using the `--fast5_out` parameter. Since version v6.3.2 `--fast5_out` is not available anymore. ONT states in the release notes for this version that the move table which contains segmentation information from `Guppy` can be found in BAM output files. More information is available via the `ont-pyguppy-client-lib` interface. Generally, resquiggler `Nanopolish Eventalign` (46), `Tombo` from ONT or `f5c` (47) allow a more detailed analysis of the raw ONT signal, as they provide the signal segmentation and a base to signal assignment, Fig. 8 C). They allocate signal datapoints to a sequence of bases that most likely produced these datapoints (46). Note, although identical content in the pores can produce slightly different signal intensities (pico Amperes), whereas different content in the pores in most cases produce significant different signal intensities.

Required inputs are the raw ONT signal (`fast5`), the basecalls (`fastq`), a reference sequence (`fasta`), and a mapping (`bam`). The reference sequence is either known or assembled from the data. `Nanopolish eventalign` produces a Tab-separated values (TSV) output that is written to stdout in the command line. The stdout should be redirected to a TSV file. This file can be of several 100 gigabytes in size. A Coronavirus run (ID 5.1, 32,000 bp) with 1.3 Gb across 857 000 reads (75 gigabytes) results in an uncompressed and large `nanopolish eventalign` output of 181.5 Gigabytes. The user must be sure, to reserve enough storage.

Required parameters are `--reads <reads.fa>`, which specifies the input reads in `fasta` or `fastq`; the mapping file with `--bam <alignments.bam>` of the provided reads to reference sequences; and the reference sequence `--genome <genome.fa>` used for the mapping.

Additional, common parameters are `--summary <file>` which produces a file to write additional data like a read index to read name mapping or the path of the `fast5` file; and `--signal-index` will add the segment start and end indices to the TSV output.

ONT developed `Tombo`, a resquiggling tool that uses dynamic time warping to segment the raw signal, again according to a given reference sequence. Since 2020, `Tombo` is no longer developed by ONT and they do not recommend to use it, as they state on the Tombo Github page. Therefore, `Tombo` does only work with the old single-read fast5 format, whereas multi-read fast5 and in the future pod5 are the standard formats.

The resquiggler `nanopolish eventalign` and `Tombo` do not support the new R10 pores. Data from r10 pores can be resquiggled with `f5c`. `F5c` is an optimized re-implementation of, among others, `nanopolish eventalign` and enables the usage of GPU (graphics processing unit) acceleration and multi-threading. Both decrease the runtime. The output format and most of the parameters stay the same.

In summary: *Beyond the standard ONT pipeline a more accurate segmentation procedure can reduce errors. Resquigglers are needed during the preprocessing step for detailed signal analysis and investigation.*

Detection of DNA modifications For DNA about 17 modifications are known (48): e.g. oxidation of cytosine (5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC),

5-formylcytosine (5fC) and 5-carboxycytosine (5caC)) (49) as well as N6-methyladenine(6mA) (50). As these modifications change the signal generated by Nanopore sequencing, theoretically all modifications can be detected with Nanopore sequencing. Multiple established tools for the detection of DNA modifications have been developed over the last years, mainly for calling 5mC and 6mA. An (incomplete) overview of known modification calling tools is given in Tab. 1. All these tools use change of the electrical current produced by the modified base compared to a non-modified base passing the nanopore. Most tools fail to detect modified bases in close proximity of a unmodified bases (e.g. `Nanopolish` cannot make a call if two CpGs are in close proximity with only one of them being methylated (51)). Methylation calling is usually done after basecalling, Fig. 8 D). The analysis of per base modifications can be performed either by a hidden markov models (`Nanopolish` (51)), by statistical tests (`Tombo` (52)), or by neuronal networks (`Guppy` ONT, `Megalodon` ONT, `DeepSignal` (53), `DeepMod` (54), `DeepMP` (55), `m6Anet` (56). Additionally, `METEORE` (13) combines the results from up to six other tools (Nanopolish, Tombo, DeepSignal, Guppy, Megalodon, DeepMod) using a random forest model and thus shows increased accuracy compared to using the single tools, but also increased runtime. In the last years multiple review paper comparing the different modification calling tools were published (13, 68, 69, 70). `Nanopolish`, the currently most widely used tool, tends to overpredict methylation values, while `DeepMod` and `Guppy` tend to underpredict methylation values (13), compared to whole genome bisulfite sequencing (WGBS), which is still the "gold-standard" used for validation. In general `Nanopolish`, `Megalodon`, `DeepSignal`, and `Guppy` show a high correlation with BS-seq results (68).

In summary: *Choose a methylation calling algorithm based on the underlying biological question, as some methylation callers tend to underpredict methylations, while others overpredict methylations, and others cannot resolve methylations for all positions.*

Custom basecalling for RNA modification detection While the modification detection of DNA is relatively established, the modification analysis of RNAs is still under development. For DNA, only about 17 modifications are known (48), while for RNA more than 170 different modifications have been seen (71).

Invoking `Guppy` from the command line interface allows for full control of parameters to configure basecalling to specific user needs, e.g. selecting non-standard basecalling models to improve basecalling quality, disabling quality filters to maximize data yield or disabling barcode trimming for downstream use. The executable is called `guppy_basecaller`. All required parameters and important optional parameters are explained above. The parameter `--config <config file>` will be most interesting for advanced users. It specifies the exact basecalling model to be used.

In addition to the choice of model accuracy level, this allows the selection of special models, e.g. for 5mC methylation detection or special research release models (e.g. 'Rerio' models), as well as custom models self-trained with the Taiyaki framework. The more recent frameworks of the Bonito basecaller and Remora modification caller from ONT meanwhile replaced Taiyaki and Rerio. Custom basecallers can be trained from scratch or by fine-tuning pre-trained models. These models can make binary predictions about modified or unmodified bases, or they call the base directly. Note that if the custom model to be loaded is not in the default data folder of the `Guppy` installation, the path with the `--data_path <folder>` parameter needs to be supplied.

These custom models could be used to directly detect RNA probing modifications for secondary structure prediction without the need of transcription to cDNA (72).

In summary: *Train your own basecalling model to predict RNA modifications or to improve basecalling quality for your special kind of data (special nucleotide distribution, many repeat regions, etc) using the provided frameworks by ONT.*

Pitfalls when using summary statistics over multiple flowcells To detect modifications, the raw ONT signal can be compared between knockout and wild-type or other samples. Nucleotide modifications, such as RNA modifications or isotopic labeling with deuterium, shift the signal compared to the unmodified bases. Therefore, they show differential signals, that can be detected using statistics. Especially summary statistics can give insights into differential signals between modified and unmodified bases. However, summary statistics can be influenced by sources other than the modification, such as different flowcells and sensors.

An example of this type of problem was discovered in (73, Sec. 3.3). Statistical regression models designed to detect nucleotide modifications exhibited perfect classification performance with an AUC of 1.0 also shown in Fig. 3.3 of (73). The classifier in question did learn the difference in features between the two flow cells, instead of between modification and canonical nucleotide. A repetition of the statistical experiment with both datasets prepared on a single flow cell showed a much more modest and expected response of the model.

Statistical analysis of the ONT signal can reveal nucleotide modifications and mutations. A range of important summary statistics (we refer to the 1^{st} to 4^{th} moment: mean, variance, skewness, and kurtosis) can be calculated for nanopore reads and are available using just one pass over the data (i.e. Welfords algorithm (74), or extensions to higher moments). These statistics can be calculated both, for full-length reads and, if segmentation data is available, also for individual nucleotides or k-mer signals. Such statistics are thus cheap to compute and a potential source for distinguishing different kinds of events, for example distinguishing a canonical nucleotide from a modified nucleotide.

The ONT signal is biased and influenced by multiple variables. Caution must be exercised to ensure that the statistics being calculated are actually relevant and useful for their intended purpose. A number of experimental setups should be considered, and there are certainly more not mentioned here. First consider a single experiment run on a single flow cell. The summary statistics for each read will depend on a number of factors. Among them the initial quality and health of the pore that generated the read, as well as the nucleotide composition of the read. These factors contribute to biases in the signal, as previously mentioned. If these variables are not properly taken into account, there is a risk that the

Table (1). Overview of methylation calling tools (adapted from Liu *et al.*), sorted by publication date. 5mC – 5-methylcytosine in DNA; m5C – 5-methylcytosine in RNA; 6mA – N6-methyladenosine in DNA; m6A – N6-methyladenosine in RNA; 5hmC – 5-hydroxymethylcytosine; psU – pseudouridine; m1A – N1-methyladenosine; 5moU – 5-methoxyuridine; m7G – N7-methylguanosine; Ino – inosine; f5C – 5-formylcytidine;

Tool	DNA or RNA	Modification	year	citation
`Nanopolish`	DNA	5mC	2017	(51)
`Tombo`	DNA, RNA	4mC, 5mC, m5C, 6mA	2017	(52)
`SignalAlign`	DNA	5mC, 5hmC, 6mA	2017	(57)
`Guppy`	DNA	5mC, 6mA		ONT
`NanoMod`	DNA	5mC	2019	(58)
`mCaller`	DNA	6mA	2019	(59)
`DeepSignal`	DNA	5mC, 6mA	2019	(53)
`DeepMod`	DNA	5mC, 6mA	2019	(54)
`Megalodon`	DNA	5mC, 6mA		ONT
`f5c`	DNA	5mC	2020	(47)
`methBERT`	DNA	5mC, 6mA	2021	(60)
`METEORE`	DNA	5mC, 6mA	2021	(13)
`DeepMP`	DNA	5mC, 6mA	2021	(55)
`MINES`	RNA	m6A	2019	(61)
`EpiNano`	RNA	m6A	2019	(62)
`xPore`	RNA	m6A	2021	(63)
`Nanom6A`	RNA	m6A	2021	(64)
`nanoRMS`	RNA	psU	2021	(65)
`ELIGOS`	RNA	m6A, m1A, 5moU, psU, m7G, Ino, hm5C, f5C	2021	(66)
`Nanocompore`	RNA	m6A	2021	(67)
`m6Anet`	RNA	m6A	2022	(56)

statistics or a statistical model may inadvertently capture these more prominent biases in the signal instead of focusing on the actual object of interest, such as modification signals.

Normalization of mean and standard deviation can remove desired biases. This problem becomes a bit more insidious when experimental designs over multiple flow cells, or with intermediate washing cycles are considered. In a typical statistical setup, data from two flow cells would be normalized similar to what happens in preparation for basecalling, see Equation **2**. This will remove any information that could be gained from the actual change in observed current, but the background noise – for example a constant shift in current in one of the flow cells – is likely to drown out small changes in current that could be of interest. Similarly, variance (on a per-read level) is also not available after normalization.

Higher moments like skew and kurtosis are also biased by pores, channels, and flow cells. It is tempting to just resort to any remaining and easy to calculate statistics, such as the next moments (skew and kurtosis), or changes in observed read length. However, systematic changes in read quality from one of the two flow cells are still leading to changes in those statistics that "survive" normalization. If the experimental setup aims to detect a particular modification it is necessary to try to quantify the effect of signal change between flow cells in a controlled test case before running the full experiment.

In summary: *If you are interested in modifications in the backbone, then the standard normalization for background distribution may destroy the signal to be detected. Hence, comparative analysis between two such samples should be performed on the same flow cell. For comparing samples across several flow cells, alternative (not established) normalization steps are required, which account for e.g. flow cell specific signal patterns.*

CONCLUSION

In summary, this study has yielded critical insights into refining nanopore sequencing using the MinION platform, with far-reaching implications for research and application. We displayed 19 suggestions for the user of nanopore sequencing in the three categories (1) library preparation guidelines; (2) sequencing guidelines; and (3) computational guidelines. Beside these hints we show data about smaller fragments sequenced first, a topic strongly discussed in the community.

We showed general statistics for the number of sequenced bases depending on the flow cell half-life, which itself surprisingly depends rather on the sample type than obvious properties such as number of active pores, amount of input material, or flow cell age. Additionally, we showed, that the flow cell half-life depends for DNA samples on the buffer used (LFB vs. SFB) and that flow cells can perform well after they have been expired. Meanwhile, we could also verify most of the observations on R10.4.1. However, for the current time point the sample size is too small to make statistically valid statements.

With this publication, correlations and assertions from the community have been statistically substantiated and are now citable for future work.

Here, we aim to describe an additional observation: within the community, it is often posited claimed that theoretically, one could sequence reads of unlimited length if the DNA fragments were arbitrarily long. We

extracted fragments of 500.000 nt from algae *Chlamydomonas euryale*, *C. reinhardtii*; however, we encountered significant challenges during the sequencing process. Despite trying various methods, we consistently observed good sequencing performance only in the initial minutes before the sequencing had to be aborted. Regarding this observation, we currently have a few potential explanations: (1) the molarity of the adaptors with respect to the HMW fragments might have been incorrect; or (2) algae may contain a high amount of polysaccharides that could influence pore activity. As seen for this example many of the observations can still not be entirely explained and call systematic analysis.

ACKNOWLEDGEMENTS

We thank Marie Lataretu for her contribution to the ONT DNA spike-in, Franziska Aron for her support and help in the lab and Carolin Dippmann for for her contribution to adaptive sampling and flow cell reusage. Competing interests: DM, CD and BS are employees of oncgnostics GmbH, a company that aims to commercialize DNA methylation markers. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2051 – Project-ID 390713860, the German DFG Collaborative Research Centre AquaDiva (CRC 1076 AquaDiva), the German state of Thuringia via the Thüringer Aufbaubank (2021 FGI 0009), and the Carl-Zeiss-Stiftung within the program Scientific Breakthroughs in Artificial Intelligence (project "Interactive Inference").

Conflict of interest statement. None declared.

REFERENCES

1. Pareek, C. S., Smoczynski, R., and Tretyn, A. (2011) Sequencing technologies and genome sequencing. *Journal of applied genetics,* **52**, 413–435.
2. Loose, M., Rakyan, V., Holmes, N., and Payne, A. (2019) Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. *Bioinformatics,* **35**(13).
3. Oikonomopoulos, S., Bayega, A., Fahiminiya, S., Djambazian, H., Berube, P., and Ragoussis, J. (2020) Methodologies for transcript profiling using long-read technologies. *Frontiers in genetics,* **11**, 606.
4. Jain, M., Abu-Shumays, R., Olsen, H. E., and Akeson, M. (2022) Advances in nanopore direct RNA sequencing. *Nature Methods,* **19**(10), 1160–1164.
5. Wang, Y., Zhao, Y., Bollas, A., Wang, Y., and Au, K. F. (2021) Nanopore sequencing technology, bioinformatics and applications. *Nature biotechnology,* **39**(11), 1348–1365.
6. Petersen, L. M., Martin, I. W., Moschetti, W. E., Kershaw, C. M., and Tsongalis, G. J. (2019) Third-generation sequencing in the clinical laboratory: exploring the advantages and challenges of nanopore sequencing. *Journal of clinical microbiology,* **58**(1), e01315–19.
7. Nicholls, S. M., Quick, J. C., Tang, S., and Loman, N. J. (2019) Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience,* **8**(5), giz043.
8. Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020) Opportunities and challenges in long-read sequencing data analysis. *Genome biology,* **21**(1), 1–16.
9. Rang, F. J., Kloosterman, W. P., and de Ridder, J. (2018) From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome biology,* **19**(1), 90.
10. Faria, N. R., Sabino, E. C., Nunes, M. R., Alcantara, L. C. J., Loman, N. J., and Pybus, O. G. (2016) Mobile real-time surveillance of Zika virus in Brazil. *Genome medicine,* **8**(1), 1–4.
11. Magi, A., Semeraro, R., Mingrino, A., Giusti, B., and D'aurizio, R. (2018) Nanopore sequencing data analysis: state of the art, applications and challenges. *Briefings in bioinformatics,* **19**(6), 1256–1272.
12. Leggett, R. M. and Clark, M. D. (2017) A world of opportunities with nanopore sequencing. *Journal of experimental botany,* **68**(20), 5419–5429.
13. Yuen, Z. W.-S., Srivastava, A., Daniel, R., McNevin, D., Jack, C., and Eyras, E. (2021) Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. *Nature communications,* **12**(1), 3438.
14. Wood, D. E., Lu, J., and Langmead, B. (2019) Improved metagenomic analysis with Kraken 2. *Genome biology,* **20**, 1–13.
15. 'Giron'Koetsier, P. A. and Cantor, E. J. (2021) A simple approach for effective shearing and reliable concentration measurement of ultra-high-molecular-weight DNA. *BioTechniques,* (2), 439–444.
16. Prall, T. M., Neumann, E. K., Karl, J. A., Shortreed, C. G., Baker, D. A., Bussan, H. E., Wiseman, R. W., and O'Connor, D. H. (2021) Consistent ultra-long DNA sequencing with automated slow pipetting. *BMC genomics,* **22**(1), 1–12.
17. Zhang, Y., Zhang, Y., Burke, J. M., Gleitsman, K., Friedrich, S. M., Liu, K. J., and Wang, T.-H. (2016) A Simple Thermoplastic Substrate Containing Hierarchical Silica Lamellae for High-Molecular-Weight DNA Extraction. *Advanced Materials,* **28**(48), 10630–10636.
18. Jaudou, S., Tran, M.-L., Vorimore, F., Fach, P., and Delannoy, S. (2022) Evaluation of high molecular weight DNA extraction methods for long-read sequencing of Shiga toxin-producing Escherichia coli. *Plos one,* **17**(7), e0270751.
19. Heavens, D., Chooneea, D., Giolai, M., Cuber, P., Aanstad, P., Martin, S., Alston, M., Misra, R., Clark, M. D., and Leggett, R. M. (October, 2021) How low can you go? Driving down the DNA input requirements for nanopore sequencing. *biorxiv,*.
20. Mojarro, A., Hachey, J., Ruvkun, G., Zuber, M. T., and Carr, C. E. (March, 2018) CarrierSeq: a sequence analysis workflow for low-input nanopore sequencing. *BMC Bioinformatics,* **19**(1), 108.
21. Arakawa, K. (2023) Ultralow-input genome library preparation for nanopore sequencing with droplet MDA. *Methods Mol. Biol.,* **2632**, 91–100.
22. Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z., and Hwa, T. (November, 2010) Interdependence of cell growth and gene expression: origins and consequences. *Science,* **330**(6007), 1099–1102.
23. Marzluff, W. F., Wagner, E. J., and Duronio, R. J. (November, 2008) Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nat. Rev. Genet.,* **9**(11), 843–854.
24. Turowski, T. W. and Boguta, M. (May, 2021) Specific features of RNA polymerases I and III: Structure and assembly. *Front. Mol. Biosci.,* **8**, 680090.
25. O'Neil, D., Glowatz, H., and Schlumpberger, M. (Jul, 2013) Ribosomal RNA depletion for efficient use of RNA-seq capacity. *Curr Protoc Mol Biol,* **Chapter 4**, Unit 4.19.
26. Fang, Y., Changavi, A., Yang, M., Sun, L., Zhang, A., Sun, D., Sun, Z., Zhang, B., and Xu, M. Q. (Jan, 2022) Nanopore Whole Transcriptome Analysis and Pathogen Surveillance by a Novel Solid-Phase Catalysis Approach. *Adv Sci (Weinh),* **9**(3), e2103373.
27. Brandt, C., Krautwurst, S., Spott, R., Lohde, M., Jundzill, M., Marquet, M., and Hölzer, M. (July, 2021) PoreCov-an easy to use, fast, and robust workflow for SARS-CoV-2 genome reconstruction via nanopore sequencing. *Front. Genet.,* **12**, 711437.
28. McCrone, J. T., Woods, R. J., Martin, E. T., Malosh, R. E., Monto, A. S., and Lauring, A. S. (May, 2018) Stochastic processes constrain the within and between host evolution of influenza virus. *Elife,* **7**.
29. Gilbert, K. B., Fahlgren, N., Kasschau, K. D., Chapman, E. J., Carrington, J. C., and Carbonell, A. (November, 2014) Preparation of multiplexed small RNA libraries from plants. *Bio Protoc.,* **4**(21).
30. Taylor, B. C., Lejzerowicz, F., Poirel, M., Shaffer, J. P., Jiang, L., Aksenov, A., Litwin, N., Humphrey, G., Martino, C., Miller-Montgomery, S., Dorrestein, P. C., Veiga, P., Song, S. J., McDonald, D., Derrien, M., and Knight, R. (March, 2020) Consumption of fermented foods is associated with systematic differences in the gut microbiome and metabolome. *mSystems,* **5**(2).
31. Maguire, S. and Guan, S. (2022) Rolling circle reverse transcription enables high fidelity nanopore sequencing of small RNA. *PLoS One,* **17**(10), e0275471.
32. Martin, S., Heavens, D., Lan, Y., Horsfield, S., Clark, M. D., and Leggett, R. M. (2022) Nanopore adaptive sampling: a tool for enrichment of low

abundance species in metagenomic samples. *Genome biology,* **23**(1), 1–27.
33. Leger, A. and Leonardi, T. (February, 2019) pycoQC, interactive quality control for Oxford Nanopore Sequencing. *Journal of Open Source Software,* **4**(34), 1236.
34. De Coster, W. and Rademakers, R. (05, 2023) NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics,* **39**(5) btad311.
35. Fukasawa, Y., Ermini, L., Wang, H., Carty, K., and Cheung, M.-S. (April, 2020) LongQC: A Quality Control Tool for Third Generation Sequencing Long Read Data. *G3 Genes|Genomes|Genetics,* **10**(4), 1193–1196.
36. Li, H. (May, 2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics,* **34**(18), 3094–3100.
37. Sović, I., Šikić, M., Wilm, A., Fenlon, S. N., Chen, S., and Nagarajan, N. (April, 2016) Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nature Communications,* **7**(1).
38. Jain, C., Rhie, A., Zhang, H., Chu, C., Walenz, B. P., Koren, S., and Phillippy, A. M. (July, 2020) Weighted minimizer sampling improves long read mapping. *Bioinformatics,* **36**(Supplement_1), i111–i118.
39. Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (April, 2019) Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology,* **37**(5), 540–546.
40. Hu, J., Wang, Z., Sun, Z., Hu, B., Ayoola, A. O., Liang, F., Li, J., Sandoval, J. R., Cooper, D. N., Ye, K., Ruan, J., Xiao, C.-L., Wang, D.-P., Wu, D.-D., and Wang, S. (March, 2023) An efficient error correction and accurate assembly tool for noisy long reads. *biorxiv,*.
41. Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., Sedlazeck, F. J., Marschall, T., Mayes, S., Costa, V., Zook, J. M., Liu, K. J., Kilburn, D., Sorensen, M., Munson, K. M., Vollger, M. R., Monlong, J., Garrison, E., Eichler, E. E., Salama, S., Haussler, D., Green, R. E., Akeson, M., Phillippy, A., Miga, K. H., Carnevali, P., Jain, M., and Paten, B. (May, 2020) Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology,* **38**(9), 1044–1053.
42. Sun, J., Li, R., Chen, C., Sigwart, J. D., and Kocot, K. M. (April, 2021) Benchmarking Oxford Nanopore read assemblers for high-quality molluscan genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences,* **376**(1825), 20200160.
43. Boostrom, I., Portal, E. A. R., Spiller, O. B., Walsh, T. R., and Sands, K. (March, 2022) Comparing Long-Read Assemblers to Explore the Potential of a Sustainable Low-Cost, Low-Infrastructure Approach to Sequence Antimicrobial Resistant Bacteria With Oxford Nanopore Sequencing. *Frontiers in Microbiology,* **13**.
44. Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (January, 2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research,* **27**(5), 737–746.
45. Fleming, A. M., Mathewson, N. J., Manage, S. A. H., and Burrows, C. J. (September, 2021) Nanopore Dwell Time Analysis Permits Sequencing and Conformational Assignment of Pseudouridine in SARS-CoV-2. *ACS Central Science,* **7**(10), 1707–1717.
46. Loman, N. J., Quick, J., and Simpson, J. T. (June, 2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods,* **12**(8), 733–735.
47. Gamaarachchi, H., Lam, C. W., Jayatilaka, G., Samarakoon, H., Simpson, J. T., Smith, M. A., and Parameswaran, S. (2020) GPU accelerated adaptive banded event alignment for rapid comparative nanopore signal analysis. *BMC bioinformatics,* **21**(1), 1–13.
48. Raiber, E.-A., Hardisty, R., van Delft, P., and Balasubramanian, S. (2017) Mapping and elucidating the function of modified bases in DNA. *Nature Reviews Chemistry,* **1**(9), 0069.
49. Klungland, A. and Robertson, A. B. (2017) Oxidized C5-methyl cytosine bases in DNA: 5-Hydroxymethylcytosine; 5-formylcytosine; and 5-carboxycytosine. *Free Radical Biology and Medicine,* **107**, 62–68.
50. Heyn, H. and Esteller, M. (2015) An adenine code for DNA: a second life for N6-methyladenine. *Cell,* **161**(4), 710–713.
51. Simpson, J. T., Workman, R. E., Zuzarte, P., David, M., Dursi, L., and Timp, W. (2017) Detecting DNA cytosine methylation using nanopore sequencing. *Nature methods,* **14**(4), 407–410.
52. Stoiber, M., Quick, J., Egan, R., Eun Lee, J., Celniker, S., Neely, R. K., Loman, N., Pennacchio, L. A., and Brown, J. (2016) De novo identification of DNA modifications enabled by genome-guided nanopore signal processing. *BioRxiv,* p. 094672.
53. Ni, P., Huang, N., Zhang, Z., Wang, D.-P., Liang, F., Miao, Y., Xiao, C.-L., Luo, F., and Wang, J. (2019) DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics,* **35**(22), 4586–4595.
54. Liu, Q., Fang, L., Yu, G., Wang, D., Xiao, C.-L., and Wang, K. (2019) Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nature communications,* **10**(1), 2449.
55. Bonet, J., Chen, M., Dabad, M., Heath, S., Gonzalez-Perez, A., Lopez-Bigas, N., and Lagergren, J. (2022) DeepMP: a deep learning tool to detect DNA base modifications on Nanopore sequencing data. *Bioinformatics,* **38**(5), 1235–1243.
56. Hendra, C., Pratanwanich, P. N., Wan, Y. K., Goh, W. S., Thiery, A., and Göke, J. (2022) Detection of m6A from direct RNA sequencing using a multiple instance learning framework. *Nature Methods,* **19**(12), 1590–1598.
57. Rand, A. C., Jain, M., Eizenga, J. M., Musselman-Brown, A., Olsen, H. E., Akeson, M., and Paten, B. (2017) Mapping DNA methylation with high-throughput nanopore sequencing. *Nature methods,* **14**(4), 411–413.
58. Liu, Q., Georgieva, D. C., Egli, D., and Wang, K. (2019) NanoMod: a computational tool to detect DNA modifications using Nanopore long-read sequencing data. *BMC genomics,* **20**(1), 31–42.
59. McIntyre, A. B., Alexander, N., Grigorev, K., Bezdan, D., Sichtig, H., Chiu, C. Y., and Mason, C. E. (2019) Single-molecule sequencing detection of N 6-methyladenine in microbial reference materials. *Nature communications,* **10**(1), 579.
60. Zhang, Y.-z., Yamaguchi, K., Hatakeyama, S., Furukawa, Y., Miyano, S., Yamaguchi, R., and Imoto, S. (2021) On the application of BERT models for nanopore methylation detection. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* IEEE pp. 320–327.
61. Lorenz, D. A., Sathe, S., Einstein, J. M., and Yeo, G. W. (2020) Direct RNA sequencing enables m6A detection in endogenous transcript isoforms at base-specific resolution. *Rna,* **26**(1), 19–28.
62. Liu, H., Begik, O., and Novoa, E. M. (2021) EpiNano: detection of m 6 A RNA modifications using oxford nanopore direct RNA sequencing. *RNA Modifications: Methods and Protocols,* pp. 31–52.
63. Pratanwanich, P. N., Yao, F., Chen, Y., Koh, C. W., Wan, Y. K., Hendra, C., Poon, P., Goh, Y. T., Yap, P. M., Chooi, J. Y., et al. (2021) Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. *Nature biotechnology,* **39**(11), 1394–1402.
64. Gao, Y., Liu, X., Wu, B., Wang, H., Xi, F., Kohnen, M. V., Reddy, A. S., and Gu, L. (2021) Quantitative profiling of N 6-methyladenosine at single-base resolution in stem-differentiating xylem of Populus trichocarpa using Nanopore direct RNA sequencing. *Genome biology,* **22**, 1–17.
65. Begik, O., Lucas, M. C., Pryszcz, L. P., Ramirez, J. M., Medina, R., Milenkovic, I., Cruciani, S., Liu, H., Vieira, H. G. S., Sas-Chen, A., et al. (2021) Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nature biotechnology,* **39**(10), 1278–1291.
66. Jenjaroenpun, P., Wongsurawat, T., Wadley, T. D., Wassenaar, T. M., Liu, J., Dai, Q., Wanchai, V., Akel, N. S., Jamshidi-Parsian, A., Franco, A. T., et al. (2021) Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic acids research,* **49**(2), e7–e7.
67. Leger, A., Amaral, P. P., Pandolfini, L., Capitanchik, C., Capraro, F., Miano, V., Migliori, V., Toolan-Kerr, P., Sideri, T., Enright, A. J., et al. (2021) RNA modifications detection by comparative Nanopore direct RNA sequencing. *Nature communications,* **12**(1), 7198.
68. Liu, Y., Rosikiewicz, W., Pan, Z., Jillette, N., Wang, P., Taghbalout, A., Foox, J., Mason, C., Carroll, M., Cheng, A., et al. (2021) DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome biology,* **22**(1), 1–33.
69. Yao, B., Hsu, C., Goldner, G., Michaeli, Y., Ebenstein, Y., and Listgarten, J. (2021) Nanopore callers for epigenetics from limited supervised data. *bioRxiv,* pp. 2021–06.
70. Akbari, V., Garant, J.-M., O'Neill, K., Pandoh, P., Moore, R., Marra, M. A., Hirst, M., and Jones, S. J. (2021) Megabase-scale methylation phasing using nanopore long reads and NanoMethPhase. *Genome biology,* **22**(1), 1–21.
71. Chen, H., Yao, J., Bao, R., Dong, Y., Zhang, T., Du, Y., Wang, G., Ni, D., Xun, Z., Niu, X., Ye, Y., and Li, H.-B. (February, 2021) Cross-talk of four types of RNA modification writers defines tumor microenvironment and pharmacogenomic landscape in colorectal cancer. *Molecular Cancer,* **20**(1).
72. Bohn, P., Gribling-Burrer, A.-S., Ambi, U. B., and Smyth, R. P.

(April, 2023) Nano-DMS-MaP allows isoform-specific RNA structure determination. *Nature Methods,* **20**(6), 849–859.

73. Lingl, T. K. Predicting RNA Modifications with Logistic Regression and Random Forests using Summary Statistics. Master's thesis Friedrich Schiller University Jena Fürstengraben 1, 07743 Jena, Germany (2022).
74. Welford, B. (1962) Note on a method for calculating corrected sums of squares and products. *Technometrics,* **4**(3), 419–420.