Nanopore SimulatION – a raw data simulator for Nanopore Sequencing

Christian Rohrandt
Institute for Communications Technologies and Embedded Systems
Kiel University of Applied Sciences
Kiel, Germany
christian.rohrandt@fh-kiel.de

Nadine Kraft
Institute for Communications Technologies and Embedded Systems
Kiel University of Applied Sciences
Kiel, Germany
nadine.kraft@student.fh-kiel.de

Pay Gießelmann
Max Planck Institute for Molecular Genetics
Berlin, Germany
giesselmann@molgen.mpg.de

Björn Brändl
Max Planck Institute for Molecular Genetics
Berlin, Germany
braendl@molgen.mpg.de

Bernhard M. Schuldt
Zentrum für Integrative Psychiatrie gGmbH
Universitätsklinikum Schleswig-Holstein Campus Kiel
Kiel, Germany
bernhardmichael.schuldt@uksh.de

Prof. Dr. Ulrich Jetzek
Institute for Communications Technologies and Embedded Systems
Kiel University of Applied Sciences
Kiel, Germany
ulrich.jetzek@fh-kiel.de

PD Dr. Franz-Josef Müller
Zentrum für Integrative Psychiatrie gGmbH
Universitätsklinikum Schleswig-Holstein Campus Kiel
Kiel, Germany

Max Planck Institute for Molecular Genetics
Berlin, Germany
franz-josef.mueller@uksh.de

Abstract— Nanopore DNA sequencing enables the sequence determination of single DNA molecules up to 10,000 times longer than currently permitted by second-generation sequencing platforms. Nanopore sequencing gives real-time access to sequencing data and enables the detection of epigenetic modifications. This unique feature set is poised to foster the development of novel biomedical applications previously deemed unfeasible.

Nanopore sequencing is based on picoampere scale measurement of current modulated by DNA or RNA polymers traveling through a nanometer opening between two compartments. Each of the five canonical nucleobases (A, T, G, C, U) has a characteristic electrical resistance, which ultimately enables the determination of the precise base sequence. However, a substantial computational effort is required to resolve the underlying sequence from a time-warped and noisy stream of digitized current measurements.

Recently, a wide range of digital signal analysis and machine learning methods have been developed for Nanopore sequencing applications. Clinically relevant questions, such as the quantification of short repetitive DNA sequences remain an unresolved challenge to current generic, state-of-the-art nanopore data analysis methods. We believe realistic simulation of the signal stream can be instrumental in the development of tailored algorithms for such novel biomedical applications.

Based on our work with the Oxford Nanopore Technologies MinION and PromethION platform, we have developed Nanopore SimulatION, a software package for the *in silico* generation of realistic, raw-signal-level data. Nanopore SimulatION starts from a reference genome in conjunction with a configuration and model file derived from real-world nanopore sequencing experiments as input. To validate our algorithm, we have sequenced custom synthetic DNA, and in so doing have generated a "ground-truth" data set potentially useful for downstream algorithm development. Additionally, we demonstrate Nanopore SimulatION's utility for method development in typical clinical use cases.

Supplementary examples, raw data obtained from our synthetic DNA sequencing experiments and the software are available under the open Mozilla Public license at https://github.com/crohrandt/.

Keywords— DNA, Nanopore Sequencing, Simulation

I. INTRODUCTION

DNA sequencing with nanopores has emerged as an accessible technology that can be deployed nearly everywhere[1]. Data generated with this method has characteristic and systematic differences to that obtained with the short read sequencing platforms (illumina/454/solid). The raw nanopore signals derived from stretches of single DNA molecules (further referred to as 'reads') can span more than 2 million bases under optimized conditions[2] while second generation technologies typically generate reads reflecting only few hundred base pairs. Nanopore reads can also span highly repetitive regions inaccessible to conventional sequencing approaches[3]. Nanopore sequencing has sparked the development of new methods to enable DNA sequencing applications, such as point-of-care testing of pathogens[4, 5], real-time selective sequencing[6] or the exact quantification of pathological short tandem repeat expansions[7].

In 2014 Oxford Nanopore Technologies (ONT) released the MinION, a portable first to market product for nanopore sequencing. It can be used for nanopore sequencing of DNA

and RNA molecules at a rate of currently up to 450 base pairs per second (R9.4 or R.9.5 pore, DNA sequencing). A MinION flow cell consists of a two-compartment chamber with an encased application-specific integrated circuit (ASIC) featuring a 512 channel analog-digital converter (ADC) each channel of which addresses four nanopores in a multiplexed fashion. A membrane separates the two compartments, and a single protein pore functions as the only connection between the two chambers filled with an electrolyte solution. Across this membrane, a voltage of typically 180 millivolts is applied. As the electrically charged, single-strand DNA nucleotide chain is traversing the pore, the three-dimensional conformation of each nucleotide in the pore influences the ionic current from one chamber into the other through the nanopore. Each nanopore measurement currently integrates the current across the pore modulated by six DNA bases (6-mer) present in the pore at each point in time. This current is in the range of picoampere, and the signal is sampled by the ASIC with a frequency of typically 4 kHz for DNA and 3 kHz for RNA. The resulting raw data consists of those ADC measurements, which typically show an analyte modulated amplitude between 20 – 40 pA with a step-wise pattern with most steps reflecting a distinct DNA 6-mer (Fig. 4). The subsequent conversion of the raw signal into the letters of the DNA alphabet commonly referred to as 'base-calling', remains challenging[8], resulting in a typical sequencing error rate in the range of 15% for a base called sequence of a single read[9].

Several features set nanopore sequencing apart from sequencing-by-synthesis approaches which currently are the mainstay of the biomedical genome and transcriptome analysis. The digital signal generated by each ADC (raw signal) can be immediately processed at the same time it is generated. Loose *et al*., demonstrated that the signal could be mapped online to a reference sequence as proof-of-principle. The resulting information can be used to select reads for further sequencing as a read can be ejected from the pore by reverting the voltage across the two chambers, thus effectively ending the sequencing of an individual DNA strand. Predefined regions of interests in a given, larger genome could be identified and only those DNA fragments fully sequenced, which map to specific genomic regions (referred to as "read until")[6].

Notably, since the ionic current through a nanopore is determined by the three-dimensional conformation of the analyte, chemical DNA- and RNA-base modifications can also be detected[10]. This ability provides an alternative to whole genome bisulfite conversion, and subsequent sequencing of bisulfite converted DNA[11]. Real-time data analysis and selective sequencing could enable rapid gene panel re-sequencing in patients or the detection of cancer subtypes based on dynamic sampling DNA variants and methylation.

II. MOTIVATION

Advanced applications of nanopore sequencing such as "read-until" or direct determination of base modification for specific biological questions still pose complex optimization problems. For the development of these and other novel applications, increasingly sophisticated machine learning approaches have been employed[12]. Machine learning requires an abundantly sampled 'ground truth' in the form of training and test datasets. Especially, the ground truth for not well understood, currently 'unsequenceable' genomic regions[13] is often only available from a simulation of the sequencing process.

Table 1 lists several software solutions developed for simulating nanopore sequencing. Most currently available programs can only simulate nanopore reads on the already abstracted base-space level after base calling, but cannot be used to simulate raw nanopore signals as generated by the ADC. ReadSim is a software tool capable of simulating long-reads typical for Pacific Biosciences and Oxford Nanopore sequencers[14]. Another simulator for these sequencing technologies is SiLiCo[15]. Both simulators statically model the read characteristics, and no parameters may be adopted from a real-world nanopore sequencing run. A dynamically trainable model is provided by NanoSim[16] and its fork NanoSim-H[17]. Both feature a training phase where characteristics may be taken from real-world experimentally generated fasta sequence files. Deep Simulator[18], the first raw simulating software has been published in 2018. Despite outputting raw data, this software only generates idealized signal values without any noise characteristics from a given reference genome. A comparison of the resulting accuracies of all simulation tools is shown in Table 2.

Hence, it is very likely that nanopore raw signal data processing solutions optimized using existing data simulation tools would perform poorly in real-world applications. Our work attempts to fulfill the as yet unmet need for simulation software capable of generating raw nanopore signals with realistic and controllable noise characteristics.

Ideally, an optimal data simulation solution should integrate into standard software toolchains already widely deployed in the community and be adaptable to the rapidly changing parameters of the sequencing technology. The input of any reference genome and incorporation of mutations or other biological properties in the simulated result is highly relevant for supporting a wide range of use cases. Sequencing parameter input should reflect those from a real-world sequencing experiment, while also having the flexibility to allow systematic variation of individual parameters to identify optimal conditions. For optimum usability, the software should also be modular and easily modifiable.

Table 1. Comparison of nanopore read simulation software

	Input reference file	Model-based simulation	Training of characteristics	Realistic ONT sequence output	Generates raw fast5	Realistic Noise Model
ReadSim	✓	✓		✓		
SiLiCo	✓	✓		✓		
NanoSim	✓	✓	✓	✓		
NanoSimH	✓	✓	✓	✓		
Deep Simulator	✓		✓	✓	✓	
Nanopore SimulatION	✓	✓	✓	✓	✓	✓

III. SOFTWARE DEVELOPMENT

A. Design

Nanopore SimulatION was developed with current, real-world nanopore sequencing results as a template. Discrete software modules simulate experimental parameters associated with experimental design decisions in the library preparation and sequencing methodology. Each module is realized as a discrete python function, with standardized I/O interfaces between them. The overall simulation workflow can be seen as an ordered sequence of experimental steps paired with their corresponding data simulation modules. Each module alters or adds specific signal characteristics.

To provide an experimentally determined ground truth dataset, we have designed and cloned a synthetic DNA sequence (called SynthXmer6). This sequence establishes a test case, where the ground truth comprehensively reflects the parameter space and is well defined. The SynthXmer6 sequence represents all possible 6-mers on the two strands equally[19]. As a rather short sequence of the length of 2145 base pairs, it is typically sequenced as a single read. Consequently, a very high and equally distributed coverage of the synthetic sequence can be obtained with a single nanopore sequencing experiment on a MinION flow cell.

Our simulation model was further optimized based on this high coverage nanopore data set. The synthetic sequence enabled a direct comparison of currents modulated by a near ideal representation of all possible DNA 6-mers derived from a very deeply sampled nanopore experiment with the results of variable simulated noise models.

Fig. 1 shows the workflow of a real nanopore sequencing experiment on the left side. The center column introduces potential biological attributes that the DNA sample may bring in. On the right side, the modular workflow of Nanopore SimulatION is outlined. The overall process can be divided into four general steps (Fig. 1: Labelled A, B, C, and D.).

Part A covers the DNA extraction and fragmentation in real-world nanopore sequencing experiments. At this step DNA fragments of different sizes (10^3-10^6 base pairs) are available in the sample pool. For simulation purposes, the only data required is a reference genome of a biological species (e.g., *H. sapiens* or *C. elegans*). Fragment size distributions are introduced later in Nanopore SimulatION.

Part B represents the library preparation. Here, adaptor sequences attached to a motor protein are ligated to the DNA fragments. Also, several samples may be sequenced in parallel in one sequencing run. To multiplex, each sample is labeled with a short "barcode" DNA oligomer. These adaptor and barcode sequences may also be introduced into the simulation process. The library preparation may cause additional physicochemical DNA fragmentation; notably, after this step, the fragment sizes are fixed.

Part C models the sequencing process itself. Here characteristics of the nanopore sequencing technology, as well as characteristics of different library preparation methods and DNA sample properties, are combined. In a real-world sequencing experiment, all these parameters can only be observed post-hoc from a completed sequencing experiment. In our simulation process, each parameter may be altered individually to systematically study the impact of experimental variables on the outcome of an experiment. For ease of use, all parameters are introduced in specific python class methods, which facilitate writing new python methods for additional parameters that may influence the sequencing results, e.g., when modified or novel library preparation methods are introduced. The modular design will facilitate adapting to the rapid evolution of nanopore sequencing technology.

Part D as the last module encapsulates the actual data generation step. Nanopore SimulatION was designed to be fully compatible with the standard toolchain provided by Oxford Nanopore Technologies and generates files in the standard fast5 file format. It is therefore compatible with ONT's basecaller Albacore and any other tool for nanopore sequencing data working with the generic toolchain.

B. Implementation

Nanopore SimulatION requires a reference genome, a model file, a configuration file, a .ini file and a base model file as input.

The reference genome is specified by a .fasta file. The configuration file can be built with Nanopore SimulatION by extracting key parameters from an actual nanopore sequencing data set that has progressed past the basecalling stage. A .ini file containing additional parameters has to be defined. Each of the parameters may be manually modified in the configuration file to modulate the simulation behavior. Additionally, a base model file is required. This file defines a raw current model of each 6-mer and is distributed by Oxford Nanopore Technologies. For triggering the simulation process, the number of reads to be simulated as well as the output location is specified at the command line of the application.

First, a unique run-identification is generated and the start time is saved as metadata of the simulated reads. Single-Read metadata is saved to a queue so that each read can independently be processed allowing parallel, threaded pro-

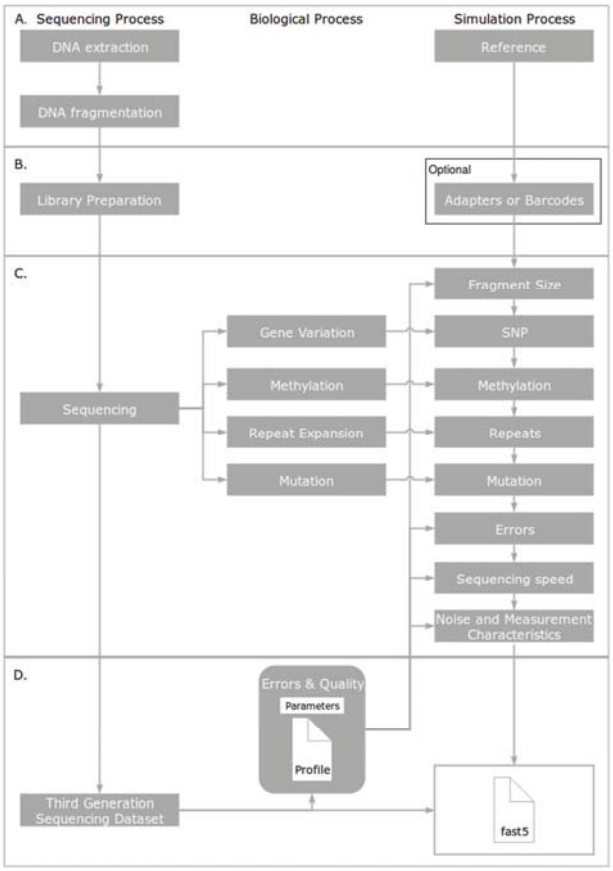

Fig. 1. Workflow of SimulatION in comparison to real nanopore sequencing. See text for details.

cessing on multi-core compute hardware. Fragment size is determined from a length distribution specified in the configuration file, and a read start position within the reference is randomly sampled for each fragment. The next step is executed via an extendable code structure enabling the simulation of biochemical modifications to the reference based DNA sequence. Several types of modifications may be introduced to the reference sequence, such as adding barcode adaptors, or single nucleotide polymorphisms (SNP). All parameters for this step are defined in the .ini file. This way the manipulation and optimization of parameters between simulation runs are most straightforward to accomplish and several parameter sets can be archived for later reproducibility. As output, the 'biochemical' API generates fragment sequences in the base space.

The fragment sequences are then translated into current mean values with a standard deviation specified in the model file. Furthermore, as the DNA strand does not traverse the pore with uniform velocity, the time any given 6-mer remains in the pore, and a specific current can be measured, is commonly referred to as idling or dwell time. Consequently, a sampled distribution of idling time per simulated 6-mer is employed to more realistically simulate the raw signal. For the idling distribution, two options have been implemented. The first option takes the distribution from the configuration file. The other option takes the base of this distribution from Scrappie[20]. Scrappie has an evolved base model that also takes the dwell time into account. With this option a more realistic simulation is possible.

In the next step, measurement error and an additional noise term are added to the signal. Based on the standard deviation, the mean current value and the number of samples, the stream of current measurements is generated. The error based on the standard deviation and the noise level can be adjusted independently. Afterward, the raw current values are transformed into the integer representation as generated by the ADC. Furthermore, an offset individual to each fragment and an amplitude range is taken from a distribution from the configuration and applied to the transformation.

In the final step, the raw values are written to a fast5 file with the metadata from the configuration file. This output can be analyzed with any currently available standard nanopore software toolchain. The fast5 file format is based on the HDF5 container format and can store different data in one file[2]. Therefore, not only the simulated raw data is stored in the fast5 file, but also a list of the original base sequence is stored to have the ground truth alongside the simulated data. This additional information can be used for validation or training of neural networks.

IV. Use Cases and Software Optimization

A. Simulation vs. actual measurements of nanopore sequencing data

As a most common use case, we anticipate researchers to simulate nanopore reads from already existing reference genomes. These references may encompass plasmids or virus genomes up to more complex mammalian or even larger plant genomes. Fig. 2 shows the normalized and aligned raw current values of a short stretch of a SynthXmer6 sequencing experiment (upper) and the simulated values of the same sequence (lower).

The nanopore sequencing of the synthetic DNA enabled us to investigate the range of inherent nanopore workflow signal characteristics. The high coverage of the synthetic sequence enabled to comprehensively study the effects of technical variability of the ADC measurements for each 6-mer, namely amplitude drift, the spread of current measurements per base or the noise that leads to errors in the base-called sequence[21]. As a result, we further optimized model generation and parameter selection with our large sample pool of defined sequences. Plotting raw values aligned to the corresponding base shows our underlying model produces data visually very similar to measured data. As a rigorous mean of comparing real data and simulated data, the data already shown in Fig. 2 were aligned to each other using the Dynamic Time-Warping (DTW) algorithm. A classical multidimensional scaling plot based on the DTW distances is shown in Fig. 3, demonstrating that measured and simulated raw signal data cluster together. Therefore we conclude the simulated data can be applied in algorithm development for nanopore sequencing data analysis software.

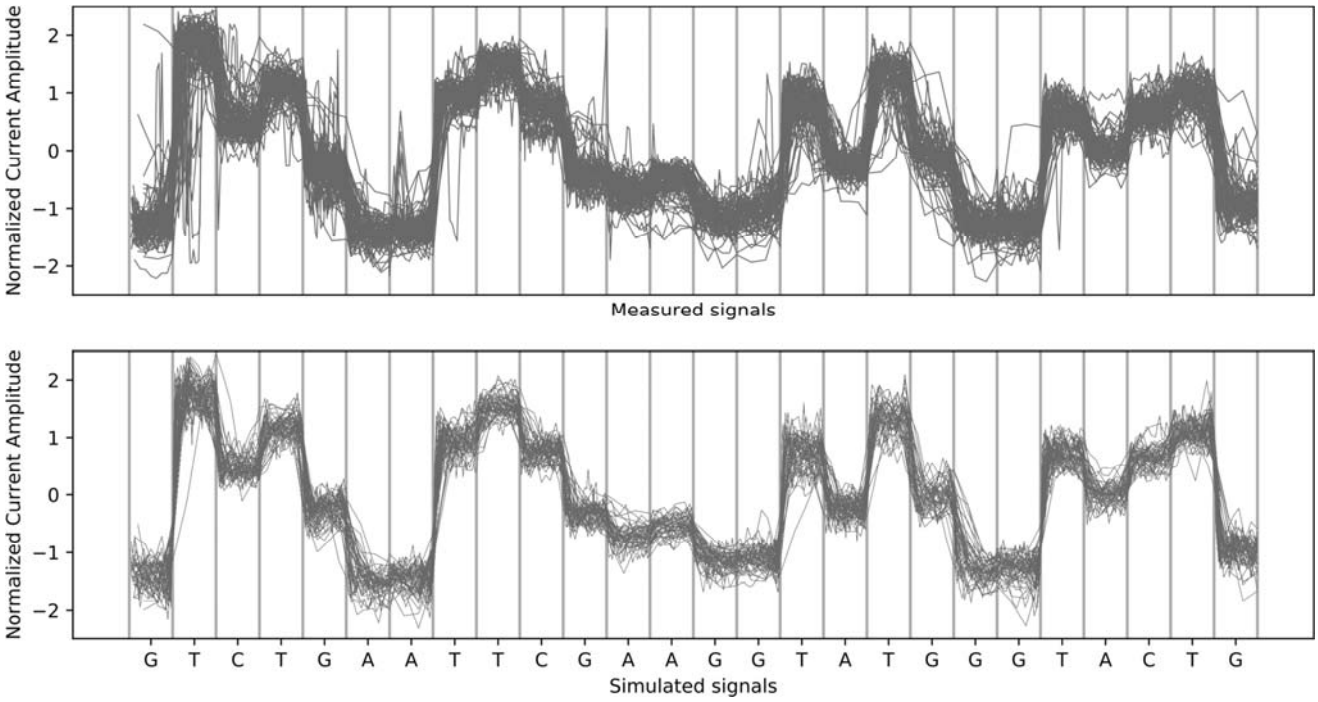

Fig. 2. Aligned raw-sample data from real nanopore sequencing experiment compared to simulated data

To further evaluate Nanopore SimulatION in later steps in the nanopore pipeline we basecalled the raw signals and aligned them to the synthetic reference sequence. For basecalling Nanopore SimulatION and DeepSimulator outputs ONT albacore[22] 2.3.1 was used, the alignments for all simulators was generated with minimap2[23] 2.11. We compared the aligned output of Nanopore SimulatION and other simulators capable of simulating nanopore sequencing long reads on the sequence level. The results are shown in Table 2. Base identity is the percentage of bases that match the reference genome after alignment.

Table 2. Statistics for a real experiment in comparison to simulated data from all simulation tools

	Base identity	Insertions	Deletions
Real data	86,75 %	1,14 %	8,53 %
ReadSim	93,98 %	2,46 %	2,60 %
SiLiCO	100,00 %	0,00 %	0,00 %
NanoSim-H	90,63 %	2,21 %	4,24 %
DeepSimulator	89,07 %	0,89 %	7,11 %
Nanopore SimulatION	86,16 %	1,52 %	7,99 %

Based on this analysis Nanopore SimulatION successfully simulates raw nanopore signals, that produce alignment statistics very close to real measurements and the sequence results are competitive with the existing sequence and raw nanopore sequencing data simulation tools.

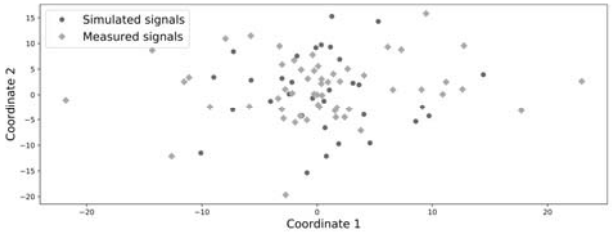

Fig. 3. Classical Multidimensional Scaling of the real and simulated data

B. Simulation of repeat expansions

As the nanopore sequencing technology generates long consecutive reads independent of DNA sequence, GC content or other features[24], nanopore reads can span very long, highly repetitive DNA sequences. Such atypical sequences cannot be reliably sequenced with conventional short-read sequencing technology[3, 12]. Pathogenic expansions of so-called Short Tandem Repeats (STRs) are an example for such unusually structured DNA sequences, which are causally linked to neuropsychiatric disorders such as intellectual disability, autism, dementia, movement disorders and epilepsy[25]. STRs consist of DNA sequences of three to six bases, which are usually consecutively repeated in healthy individual up to 20 times at a specific genomic position, but may continuously expand to several thousand repeats in case of some genetic disorders. Currently prevailing short-read sequencing technologies cannot resolve STR expansions satisfactorily[26]. Expanded repeats are usually significantly longer than the highly fragmented Illumina sequencing libraries. Because of the highly repetitive nature of the STRs, an assembly of the expanded region is typically impossibly beyond twice the read length and the maximum insert lengths in the case of paired-end sequencing libraries[26]. Nanopore sequencing might offer an obvious solution to the problem, yet we found that the current standard nanopore base calling workflows cannot reliably quantify expanded STRs above a threshold of 200 repeats. Identification of the flanking sequence on the raw signal level before and after the repeat may enable the precise repeat number quantification in nanopore reads spanning unaffected or expanded STRs with signal level analysis methods. Fig. 4 displays a single read raw current signal of a 50x “CGG” repeat in 5’ untranslated region of the FMR1 gene. The vertical lines demarcate individual CGGs. The signal is time variant, and one CGG cannot satisfactorily be distinguished.

Fig. 5 shows a density plot of several different repeat lengths called by both the sequence based repeat counter RepeatHMM[13] and a custom raw signal based repeat counter. In each experiment step 250 reads spanning the whole repeat sequence were simulated with a defined length of 20, 50, 75, 100, 150, 200, 250 and 300 CGG repeats. All these data sets were then analyzed with basecalling/RepeatHMM and with an in-house signal based hidden markov model (HMM) implementation. While our raw signal HMM model has a straightforward training procedure with only four states, it results in more precise repeat counting. We noted a better accuracy and that the signal level method does not degrade as much in performance with higher repeat lengths when compared to the sequence based HMM. These results highlight that in the case of STR quantification algorithms based on the raw current signal outperform sequenced-based methods due to basecalling errors introduced in the upstream toolchain. We posit that Nanopore SimulatION may enable the development of clinical grade raw signal based STR counters.

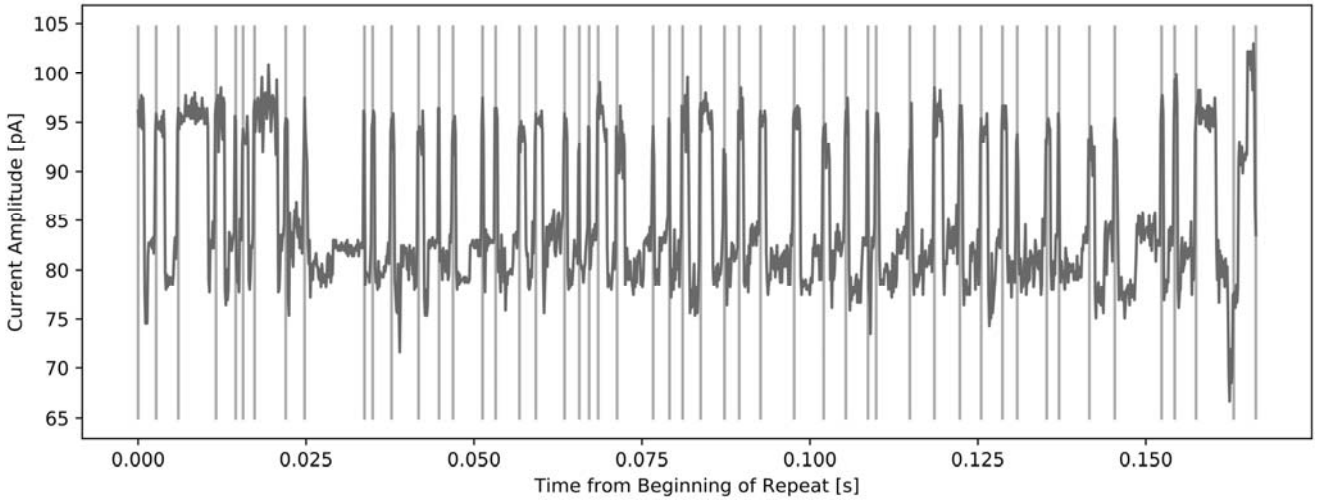

Fig. 4. Simulated raw current signal of a tandem repeat

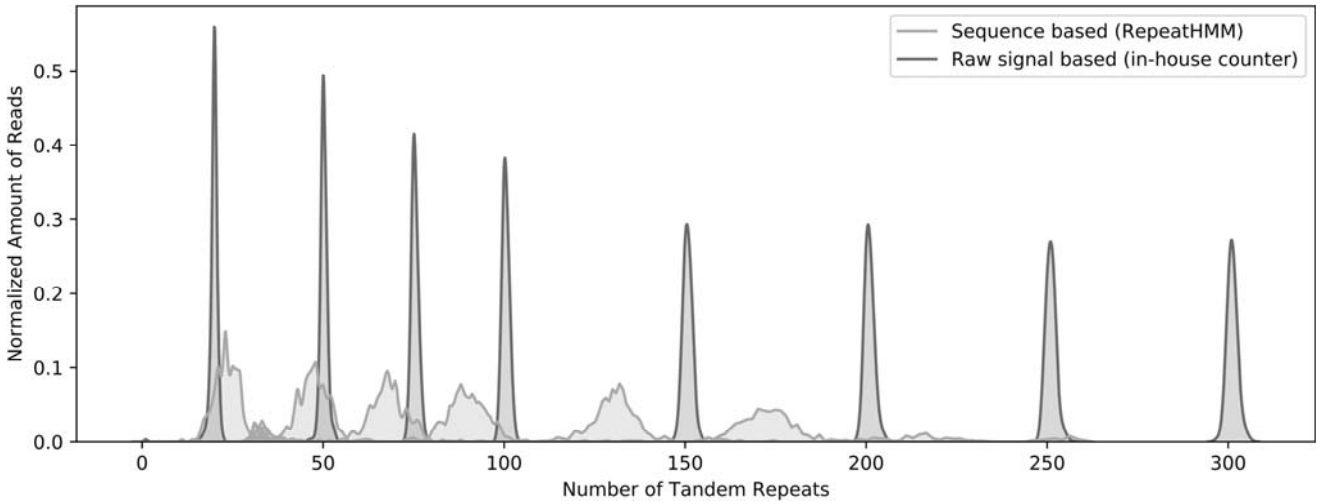

Fig. 5. Comparison of repeat counter methods in respect to their accuracy

Moreover, we anticipate that other nanopore raw signal analysis methods will benefit from our approach to simulate nanopore raw data with realistic noise characteristics. From a practical perspective, procuring experimental nanopore data from a large number of STR expansion disorders with in some cases only very few patients with a specific disorder worldwide will actively limit general and customized repeat quantification method development. Here, a realistic simulation of the anticipated data will be instrumental.

C. Variant calling and haplotype phasing

Single nucleotide polymorphisms (SNP) are among the most common genetic causes underlying biological variability. In the human genome, on average every 1000th base is polymorphic between two individuals. As SNP often vary between two alleles, this information may be used to determine if a specific gene variant originates from the maternal or the paternal genome. This method is referred to as phasing. Not in all cases, the originating genomes are known for analysis, and a de novo allele reconstruction needs to be done for the analysis. Nanopore sequencing is very well suited for the reconstruction of genomic haplotypes as several allelic SNPs may be covered with one long read, commonly referred to as "phasing". Consequently, we next explored this type of analysis with SimulatION.

A 10^3 base pair section of the gene MHC on chromosome 6 was modified with sets of five SNP variants, which were introduced randomly. With these two reference sequences, representing the maternal and the paternal genome, several coverage scenarios were contemplated. In ten rounds each reference was simulated with Nanopore SimulatION using a real-world read length distribution with a defined number of 10, 15, 20, 25, 30, 40, 50 and 100 reads. All read sets were base-called using Albacore and aligned to the chromosome 6 reference using minimap 2[23] as if they were sequenced together. Afterward, the widely used nanopolish tool[27] was used to detect SNPs in the dataset. A minimum number of 10 reads supporting a SNP was parameterized. The next step used the experimental function phase-reads of nanopolish to phase these SNPs, which means that an alignment is compiled with only the SNP being differentiable. This alignment is further analyzed in a custom python script, which assigns each read to a specific allele by grouping reads with several SNP in them, into clusters that form an allele profile. Fig. 6 shows the results of phasing the simulated data. The number of reads able to assign to an allele is shown as a range over all ten runs from the eight sets of data simulated, with the actual simulated number of reads per allele denoted by a black diamond. Not every set gave a result for each allele; therefore the number of sets that support the number of reads covering the specific allele is printed in bold numbers above the x-axis. As can be seen in the left half, with low coverage, SNPs cannot be distinguished well enough for enabling reliable haplotypic phasing. Coverages of 50 and more are required for interpretable results.

D. Discussion

We have developed a modular software tool that is capable of simulating the raw electrical current values of nanopore sequencing reads. Moreover, we have successfully

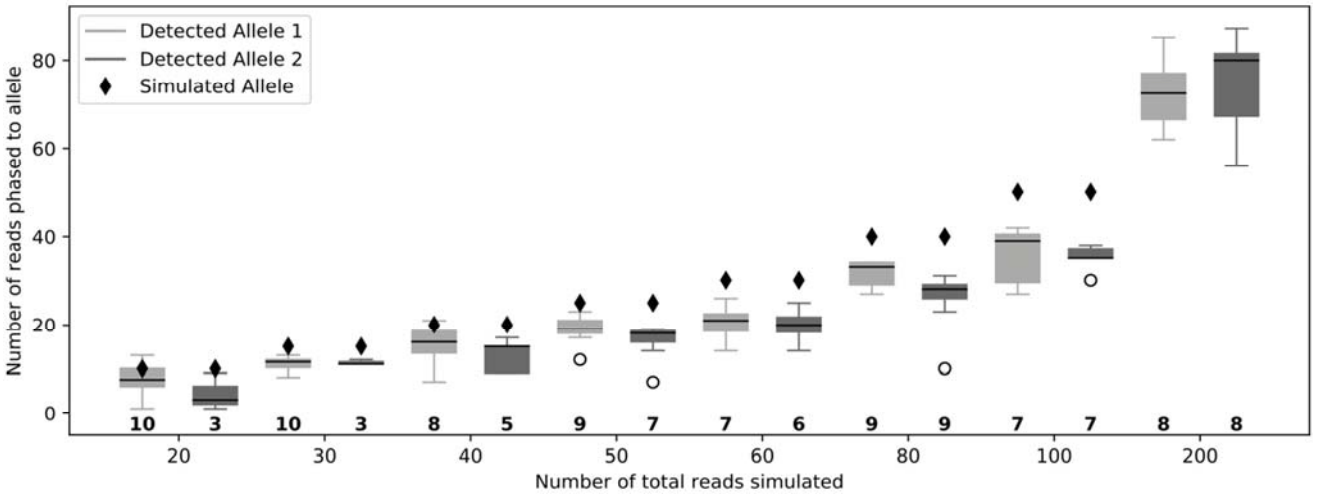

Fig. 6. Phased reads in a two allele scenario

demonstrated that this type of simulation software can be used for developing algorithms based on raw current values. Based on several use cases, we highlight the potential advantages of methods for raw signal space analysis. We conclude that ground truth datasets based on our simulation of nanopore sequencing experiments may greatly facilitate the development of new analysis methods particularly in use cases, where experimental data currently cannot be satisfactorily simulated with abstracted DNA sequences because of still relevant error rate inherent to current state-of-the-art basecalling methods. Additionally, Nanopore SimulatION enables to develop and optimize their algorithms without the need of difficult to procure DNA samples, molecular laboratory staff, and equipment. Nanopore SimulatION may enable a more useful resource usage within a research group and could shorten the question-to-answer for hypothesis-driven, biological questions.

Based on our experience, the main challenge for answering research or clinical question with nanopore sequencing is to find an optimal solution for cost/resource utilization and custom library preparation method development.

For example, with the PromethION, human genomes can be rapidly sequenced with very high coverage at a price point of several thousand US$ comparable to Illumina whole genome sequencing. Methods for highly selective DNA fragment enrichment methods could provide a specific result, e.g. for the diagnosis of a defined set of repeat expansions disorders with much less sequencing effort using consumables amounting to less than $150. For such a tailored solution involved molecular biology workflows with a large tunable parameter set (e.g., fragment size, enrichment ratios, background noise, adaptor ligation efficiencies) need to be optimized. In this example, Nanopore SimulatION can predict each quality metric or parameter in the library preparation stage that needs to be controlled for a reliable diagnosis of a repeat expansion disorder such as the Fragile X Syndrome.

We also believe, Nanopore SimulatION will demonstrate usability in the usage for the comprehensive validation and stabilization of clinical, nanopore sequencing based processes and toolchains. With its tunable library and noise parameters, our software can effortlessly generate large "challenge" datasets for the determination of the technical limits of any future diagnostic molecular and/or software workflow involving nanopore sequencing.

Further development of Nanopore SimulatION will address related biological questions like full-length direct RNA sequencing, more involved tandem repeat detection and further optimization of generic basecalling and/or sequence alignment methods. In the field of modified DNA molecules, it already gives basic support for the researchers as it enables the simulation of methylation of single bases by using an appropriate k-mer model. All these features will be relevant in the future. Especially RNA sequencing and methylation are such complex topics that they will be addressed in future work. Homopolymers, subsequences only consisting of one single character (A, C, G, T), still pose a difficulty for the basecalling softwares[28]. As the nanopore sequencing technology delivers the information of the homopolymers only in the time information of a quasi-stable raw signal, Nanopore SimulatION will be beneficial for the further basecaller development. A comparison of homopolymeric real measured data and simulated data is provided in the source code repository.

Currently, the nanopore sequencing technology is rapidly expanding its scale and use cases: The PromethION platform can generate multi-terabase scale datasets within few days at genome centers. On the other end of the spectrum, the Flongle/SmidgeION platform is marketed for kilobase-scale, focused, point-of-care sequencing applications at the bedside.

In nearly every scenario, Nanopore SimulatION can rapidly provide insights into which experimental scale and library preparation methods will result in an optimal readout and the shortest time-to-answer.

ACKNOWLEDGEMENT

FJM and BMS were supported by grants from the BMBF (13GW0128A and 01GM1513D), from the Deutsche Forschungsgemeinschaft: German Research Foundation; DFG MU 3231/3-1 and from the DFG within the framework of the Schleswig-Holstein Cluster of Excellence, EXC 306 Inflammation at Interfaces.

REFERENCES

[1] M. Jain, I. T. Fiddes, K. H. Miga, H. E. Olsen, B. Paten, and M. Akeson, "Improved data analysis for the MinION nanopore sequencer," *Nature Methods*, vol. 12, no. 4, pp. 351–356, Feb. 2015.

[2] A. Payne, N. Holmes, V. Rakyan, and M. Loose, "Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files.," *bioRxiv*, p. 312256, May 2018.

[3] M. Jain *et al.*, "Linear assembly of a human centromere on the Y chromosome," *Nature Biotechnology*, vol. 36, no. 4, pp. 321–323, Apr. 2018.

[4] J. Quick *et al.*, "Real-time, portable genome sequencing for Ebola surveillance," *Nature*, vol. 530, no. 7589, pp. 228–232, Feb. 2016.

[5] N. R. Faria *et al.*, "Establishment and cryptic transmission of Zika virus in Brazil and the Americas," *Nature*, vol. 546, no. 7658, pp. 406–410, Jun. 2017.

[6] M. Loose, S. Malla, and M. Stout, "Real-time selective sequencing using nanopore technology," *Nat Meth*, vol. 13, no. 9, pp. 751–754, Sep. 2016.

[7] H. Ishiura *et al.*, "Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy," *Nature Genetics*, vol. 50, no. 4, pp. 581–590, Apr. 2018.

[8] F. J. Rang, W. P. Kloosterman, and J. de Ridder, "From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy," *Genome Biology*, vol. 19, no. 1, p. 90, Jul. 2018.

[9] M. Jain *et al.*, "Nanopore sequencing and assembly of a human genome with ultra-long reads," *bioRxiv*, p. 128835, Apr. 2017.

[10] A. C. Rand *et al.*, "Mapping DNA methylation with high-throughput nanopore sequencing," *Nature Methods*, vol. 14, no. 4, pp. 411–413, Apr. 2017.

[11] R. Lister and J. R. Ecker, "Finding the fifth base: genome-wide sequencing of cytosine methylation," *Genome Res.*, vol. 19, no. 6, pp. 959–966, Jun. 2009.

[12] C. de Lannoy, D. de Ridder, and J. Risse, "The long reads ahead: de novo genome assembly using the MinION," *F1000Research*, vol. 6, p. 1083, Dec. 2017.

[13] Q. Liu, P. Zhang, D. Wang, W. Gu, and K. Wang, "Interrogating the 'unsequenceable' genomic trinucleotide repeat disorders by long-read sequencing," *Genome Medicine*, vol. 9, p. 65, Jul. 2017.

[14] H. Lee *et al.*, "Third-generation sequencing and the future of genomics Online methods," p. 20.

[15] E. A. G. Baker, S. Goodwin, W. R. McCombie, and O. M. Ramos, "SiLiCO: A Simulator of Long Read Sequencing in PacBio and Oxford Nanopore," *bioRxiv*, p. 076901, Sep. 2016.

[16] C. Yang, J. Chu, R. L. Warren, and I. Birol, "NanoSim: nanopore sequence read simulator based on statistical characterization," *Gigascience*, vol. 6, no. 4, pp. 1–6, Apr. 2017.

[17] K. Břinda, *NanoSim-H: Simulation of Oxford Nanopore reads. A fork of NanoSim*. 2017 [Online]. Available: https://github.com/karel-brinda/NanoSim-H. [Accessed: 07-Jun-2018]

[18] Y. Li, R. Han, C. Bi, M. Li, S. Wang, and X. Gao, "DeepSimulator: a deep simulator for Nanopore sequencing," *Bioinformatics*, Apr. 2018 [Online]. Available: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty223/4962495. [Accessed: 25-Jun-2018]

[19] Y. Orenstein and R. Shamir, "Design of shortest double-stranded DNA sequences covering all k-mers with applications to protein-

binding microarrays and synthetic enhancers," *Bioinformatics*, vol. 29, no. 13, pp. i71–i79, Jul. 2013.

[20] *scrappie: Scrappie is a technology demonstrator for the Oxford Nanopore Research Algorithms group*. Oxford Nanopore Technologies, 2018 [Online]. Available: https://github.com/nanoporetech/scrappie. [Accessed: 11-Jul-2018]

[21] A. Magi, R. Semeraro, A. Mingrino, B. Giusti, and R. D'Aurizio, "Nanopore sequencing data analysis: state of the art, applications and challenges," *Brief Bioinform* [Online]. Available: https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbx062/3869205. [Accessed: 12-Aug-2018]

[22] "nanoporetech/albacore," *GitHub*. [Online]. Available: https://github.com/nanoporetech/albacore. [Accessed: 14-Nov-2018]

[23] H. Li, "Minimap2: pairwise alignment for nucleotide sequences," *Bioinformatics* [Online]. Available: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty191/4994778. [Accessed: 22-Aug-2018]

[24] R. Krishnakumar *et al.*, "Systematic and stochastic influences on the performance of the MinION nanopore sequencer across a range of nucleotide bias," *Scientific Reports*, vol. 8, no. 1, Dec. 2018 [Online]. Available: http://www.nature.com/articles/s41598-018-21484-w. [Accessed: 20-Aug-2018]

[25] J. R. Gatchel and H. Y. Zoghbi, "Diseases of Unstable Repeat Expansion: Mechanisms and Common Principles," *Nature Reviews Genetics*, vol. 6, no. 10, pp. 743–755, Oct. 2005.

[26] M. Bahlo, M. F. Bennett, P. Degorski, R. M. Tankard, M. B. Delatycki, and P. J. Lockhart, "Recent advances in the detection of repeat expansions with short-read next-generation sequencing," *F1000Research*, vol. 7, p. 736, Jun. 2018.

[27] J. T. Simpson, R. E. Workman, P. C. Zuzarte, M. David, L. J. Dursi, and W. Timp, "Detecting DNA cytosine methylation using nanopore sequencing," *Nature Methods*, vol. 14, no. 4, pp. 407–410, Apr. 2017.

[28] P. Sarkozy, Á. Jobbágy, and P. Antal, "Calling Homopolymer Stretches from Raw Nanopore Reads by Analyzing k-mer Dwell Times," in *EMBEC & NBC 2017*, 2018, pp. 241–244.