M. Eng. Christian Rohrandt
Faculty of Computer Science and Electrical Engineering
Kiel University of Applied Sciences
AMIES 12-15 September 2018

Simulating realistic raw-signal-level nanopore data for development of new methods

Overview

Introduction

- Nanopore DNA/RNA sequencing
- Oxford Nanopore Technologies MinION DNA Sequencer
- Oxford Nanopore Technologies PromethION DNA Sequencer
- Who we are and what do we do?

Simulation of realistic raw-signal-level nanopore data

- Why do we need it?
- What should it be capable of?
- How does it work?

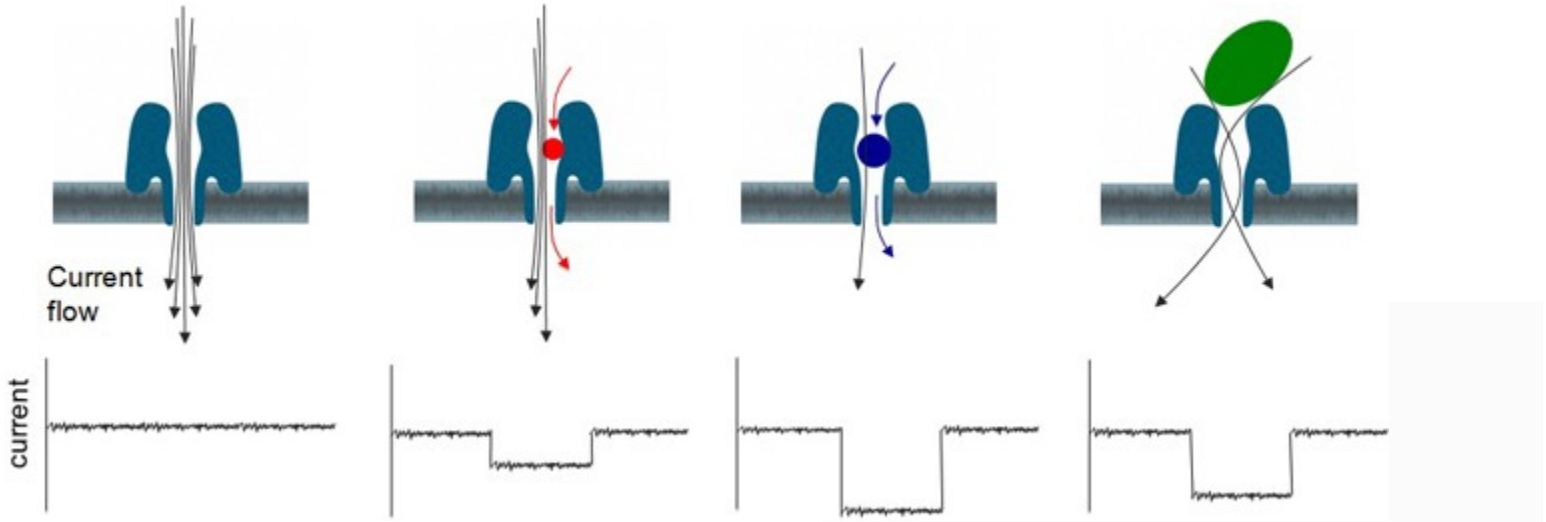
Use cases

- DNA
- Pathogenic Short Tandem Repeat expansions

Conclusion

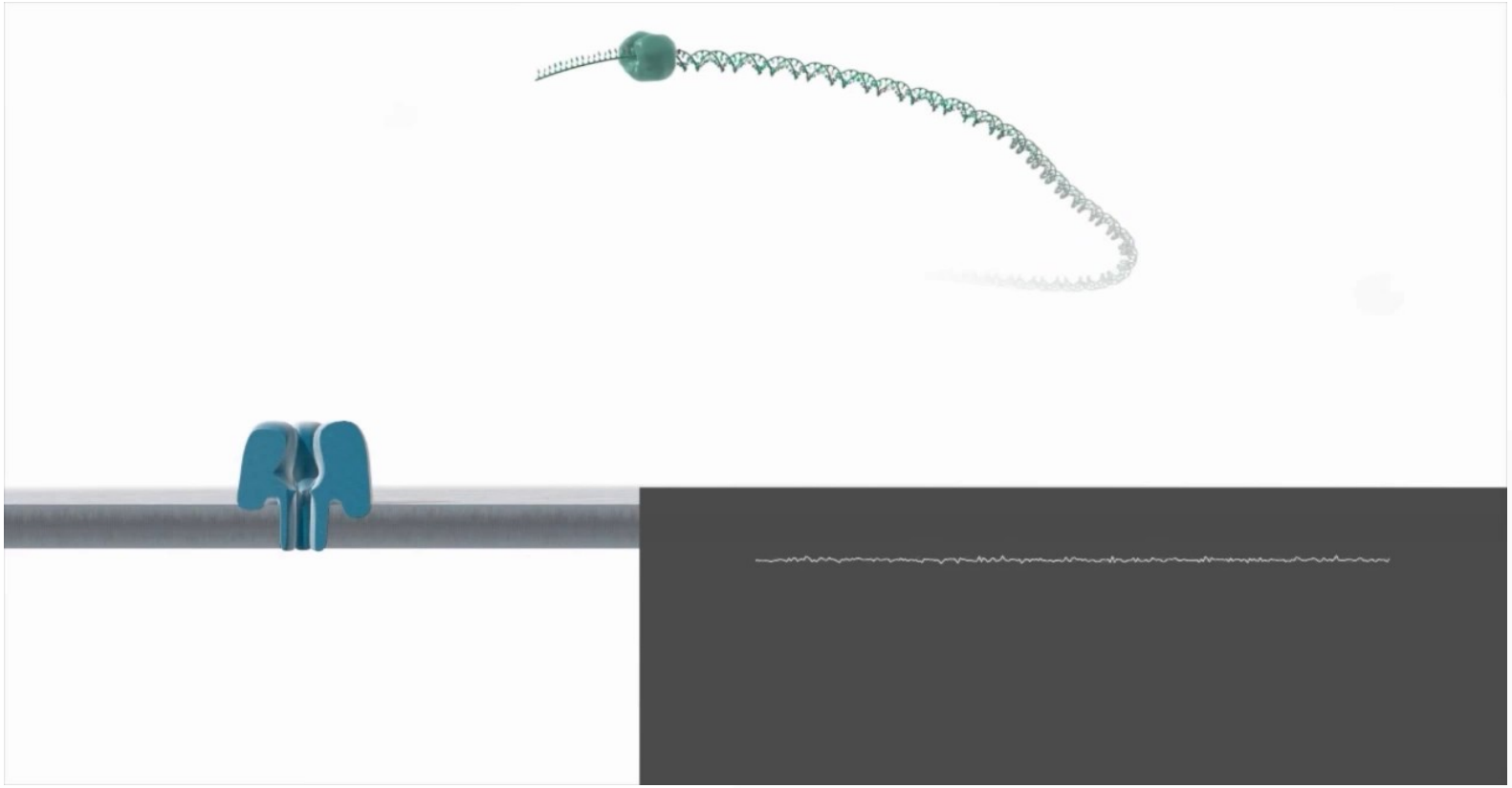
Introduction

Nanopore DNA/RNA sequencing



"Oxford Nanopore unveils portable genome sequencer – MinION." [Online].
Available: https://phys.org/news/2014-02-oxford-nanopore-unveils-portable-genome.html. [Accessed: 14-Sep-2017].

Nanopore DNA/RNA sequencing



Oxford Nanopore Technologies MinION DNA Sequencer

Image credit: Oxford Nanopore Technologies

MinION DNA Sequencer

- 512 measurement channels
  - Current measurement using an ASIC
- Sequencing speed
  - max. 70 bases/second/pore (R7)
  - max. 450 bases/second/pore (R9) (mid 2016)
  - max. 1000 bases/second/pore (R10) (not available yet)
- Basecalling accuracy
  - 80 – 85% (R7)
  - 88 – 92% (R9.4)
  - 92 – 94% (R9.5)
- powered by USB
- portable

Oxford Nanopore Technologies PromethION DNA Sequencer

PromethION DNA Sequencer

- 24 Flowcells
  - 3.000 channels per flowcell
  - 72.000 channels in parallel

- PromethION compute module
  - Dual Intel© Xeon© 48 Cores 69 Threads
  - 377 Gigabytes RAM
  - 28 TB SSD storage
  - Bank of 4 NVIDIA TESLA V100 GPUs for basecalling
  - 20 Gigabytes per second dual fibre interface

Who are we and what do we do?

- Cooperation of University Hospital Kiel, Kiel University of Applied Sciences and Max Planck Institute for Molecular Genetics Berlin
- Started MinION sequencing with pore generation R7 in 2015
- Recently started PromethION sequencing in Kiel
- Human DNA & RNA sequencing, plasmid sequencing
- Use nanopore sequencing technology for biological question that could not be answered with previous sequencing technologies

Simulation of realistic raw-signal-level nanopore data

Simulation of realistic raw-signal-level nanopore data
Why do we need it?

- Implementation of algorithms
- Testing of algorithms
- Setting up analysis pipelines
- NO need to waste flowcells just for bioinformatic purposes
- Experiments can be planned extensively beforehands
  - Patient sample amount can be determined
  - Is the question answerable with nanopore sequencing
  - How many runs need to be done

Simulation of realistic raw-signal-level nanopore data
What should it be capable of?

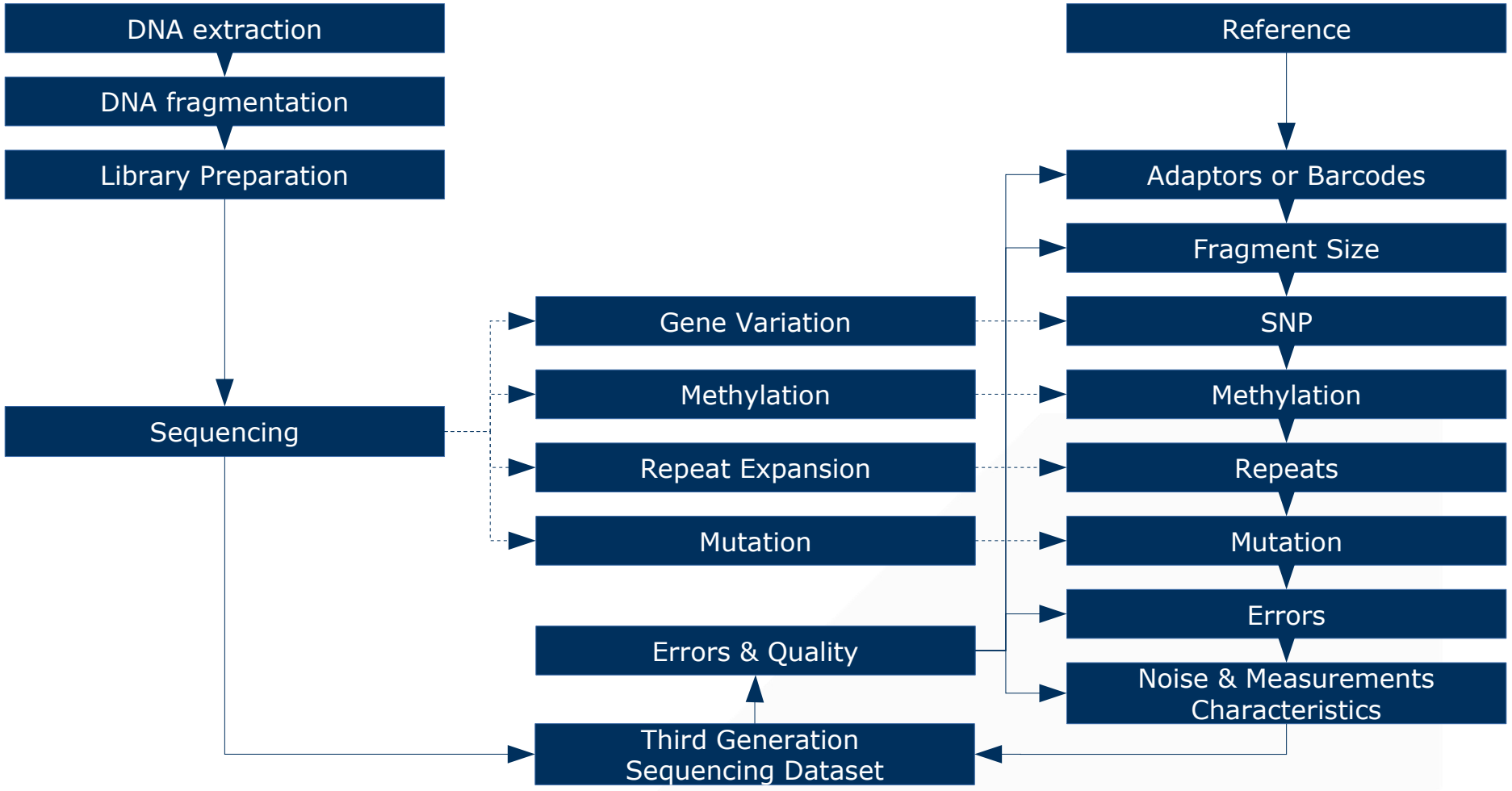
- Using parameters form real conducted experiments
  - Read length distribution
  - Derived error model
  - Metadata from MinKnow

- Generates sequencing data which is:
  - Realistic
  - Reproducable
  - Adjustable

- Should be usable without any dependency to a laboratory

- Generate output files that can be used in existing software pipelines

Simulation of realistic raw-signal-level nanopore data
How does it work?

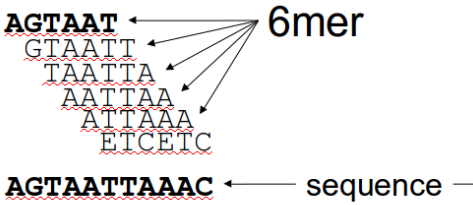


Use Cases

DNA

DNA

Synthetic DNA Sequence Ground truth



Algorithm 2. Find an optimal augmentation for a de Bruijn graph $G = (V, E)$ of odd order.

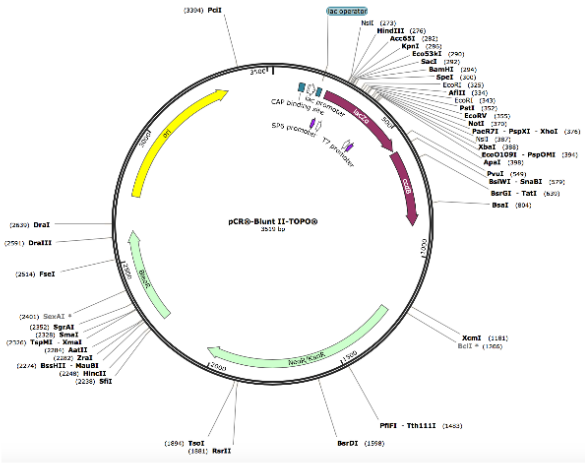
1. Add to G the set E' of palindromic edges.
   The resulting (multi-)graph is $G' = (V, E \cup E')$.
2. Define $V^+ = \{v \in V | (v, u) \in E' \wedge (u, v) \notin E' \text{ for some } u\}$

$$V^- = \{u \in V | (v, u) \in E' \wedge (u, v) \notin E' \text{ for some } v\}.$$

3. Define a complete bipartite graph $H = (V^-, V^+, F)$
   with edge weights $w(x, y) = |ov(x, y)|$.
4. Find a maximum weight-matching M in H.
5. Define $G^2 = (V, E \cup E' \cup E'')$
   where $E'' = \{(u, v) \in p(x, y) | (x, y) \in M\}$.
6. Modify M, so that each cycle in the graph $(V, E' \cup E'')$
   contains exactly two palindromic edges (Lemma 6).

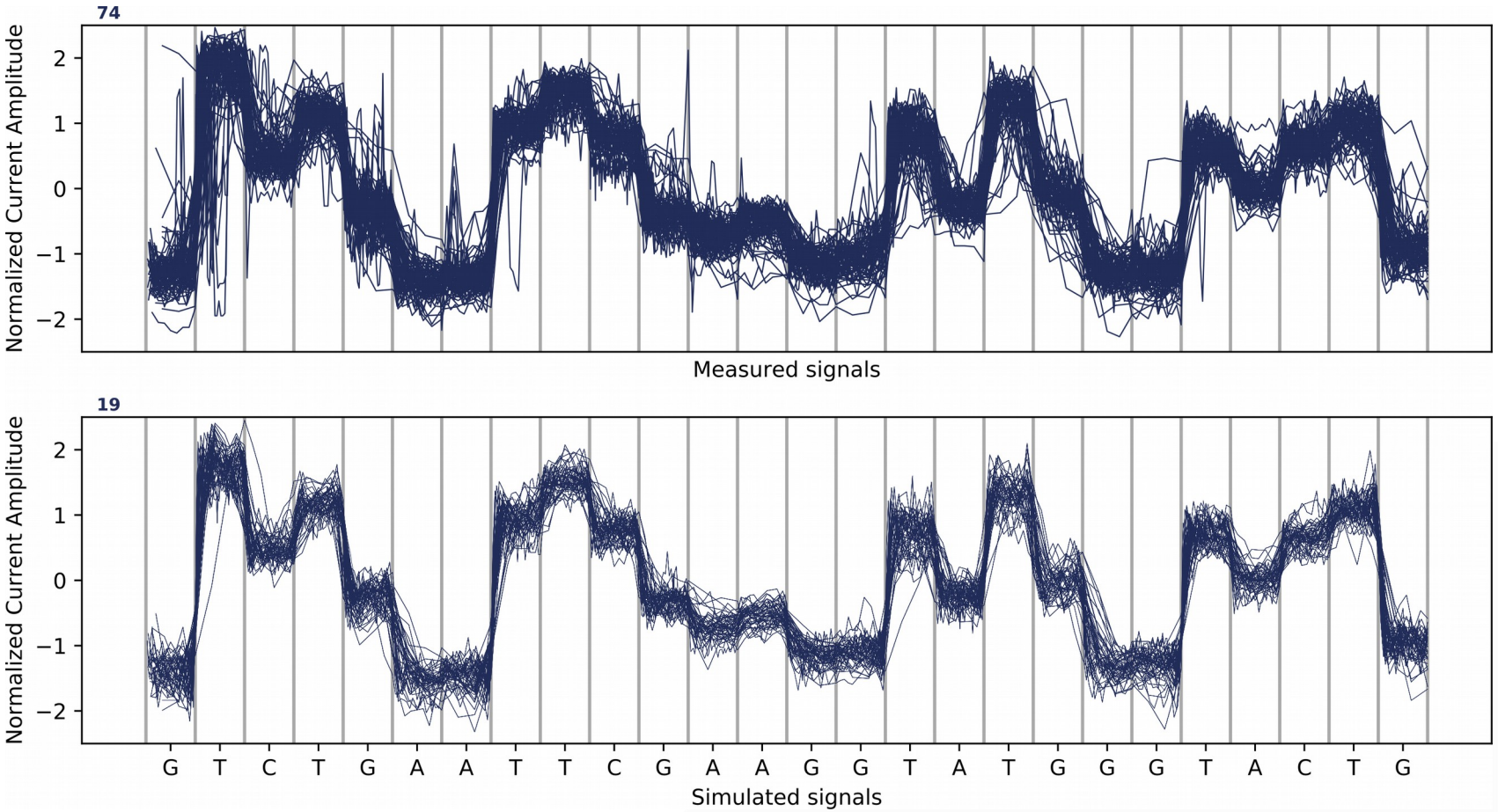
Orenstein and Shamir (Bioinformatics 2013)

```
AGTAATTAAAAATCCCCTGGACAGCTCTCACTTCACAGGGAAATCAGATCCTGCGTCCTTGGATGCATGTGCT
TTACCGCGGGCGTTTGGTTGCACGCTCGGAGGGTTAAAGTACTCAGGTCCCCCCACTGAAAATGTTGGGGTCTC
GGCTAAATTTTAGGTGCCACACATCCTTATCTTTTCGGTGGCTGTTCGATGAAACGGCAGTGATGCGCCTATC
CGGACACAGTCTGAATTCGAAGGTATGGGTACTGATCATATGTTTGCATATTCACCTTTCTCTTTAGAAGGCC
ACTTTGGGTCGGGTAAGGGGCTATTGGTGATCCCAGAAACAGTTCTGTGCACGTCCACGCGTTAACTGGTGGA
TCCGACGTCGACGCGATGGTCGCGGCCAATTGTTCCCAATGTCCTGTCGGCCGGTATCCATCCCTACAGTGCC
GAATTTAATTTGTGGAACCATGGCAGGTTGAACTATTTCACGTTGCGTATAGCTAAGGCAGCGCTCTTCAGGC
CTAGGTTCATAGTATTCTCCCGCCCCCTCATCCGGCATTAATTGCAAGCCGCTTCAAGCAAGTCTCCGGGGCAG
AAGTCGATATCGCGAAGCCCGGGGTTGGCTACATGTAACCAGATAGACTCTCTCCATAATCTCTAGATCTATG
GTATACTTCGGGGGTATTGTCGTCGCACACTGGCCATGAGACAGAGCCAGGCGAGGTTTCCGATCGATCATGA
TAGCGTTGACTGGGCCTAGAACACAATCGCTCAACTACTAAGACACGATACTGCTCTAACGACCGAGGCCGGC
TGCACCGGGAATTGAGAAGCTAGTACGTACCGGTTGTGCCTACCCTTCCATTCTATAATTACACTTGTTGTCT
AGCTCGAAAGCCATATCCTATGTACGCCCTCTGACACCATTACCAGCGAAACTCGGGACGAACTCACCCACCC
CTATAGTCCGCCTCCCTCGAGCACTAGTTATCAAAATAACCTTGCCGCAGTCGGTAGAAAAAAGCTGCAGACC
ACGAGGACCCGGCGACACTCCGCATAGCGCCAGACGGGATAAAAACAACGAGTCGCTAACTCTACACGCACGT
ATAATATAGCATACGAGCTCAGAATCCTCACAAGCGCGCGGGGAAGTTACGTCTTCGTGAGCCGTCGTGCTGC
TTAAATGGCCCGCACCCGCTGATATTTGCCTCAGTTTACAGATTCCTGGGTTTGACCGTCCGAACCCGTACGA
CTACAATGCTTCCGGTCCGTTCAGCTACGATCTCGTCTAAGCCTAAGTAGGTACCGTTGGAATGATTCGCATG
CACTTACTTGGGCTGGCATCTTGGCGGCGCGAGTATGTGACCCTAACCGCAAGAGGCTGAGATACACCTGTAT
GACGATTTGAGTGCGCATTTGGCAACTTGACAGTAGATTTACGGTTTAAAAGTTTTGTTACTCCAATCTTCCT
TTATTGCAGTTGTAAGCTTATGTCTGTTGATTGCGAGAGGTAGTACAAGACCTTAACATCAATATTAACGGGG
CCCTTGTCACGCCACGGCCCGATTAGATGATGGGCGCCGATAGTGGAGGAACTTAAGATCAAGGCTAGCGGAA
TAAGTCAGTCCTAGTAAAACGTCACCGCTCCGTCAATGACTAACACTAAACGATGCTGGGGAGCCTGCTAATA
ATGTATTTATAAGAAATTGCCCACGTGTTGCTGACCAGGTAACGTAGGAGTCACTACGGATCGCCCGTGCCCT
AGCAATGGGAGTTGGTCCATGTCGAGAACCGATGTCATGCTCCAGTAATCCGCGTAGACCGCCGTATCACGGT
GTGTACAGCAGATGGCGAACGTTATATACAAACTAATGAATCGAGTGAACATGAAGGAGGTGGTGCTACTCTT
ACGAATAGAGGAGCTTGGTAGCGCAAAAGGCGTCAGGAGATGTACCACATTCAATCCAGCTGCGCGTAATAGG
TCGAATGCAGGGTGTTTATGAGCAAAGGGACAAAGACTAGATATATGCCCCATGCCTGTTATGCGTTCTTTCC
TAATTCTTCTTGAGGTCATCAGGGCTCGCTTAATCACACGGAACGAAATGAGTCCCACAGCGTCTCTGGAAAA
CCACTCTGCGATAATACGCTTTGAAAGTCCAAAAATATGGAT
```



DNA

Simulated vs. Sequenced Data



Use Cases

Pathogenic Short Tandem Repeat expansions

Use cases
Pathogenic Short Tandem Repeat expansions

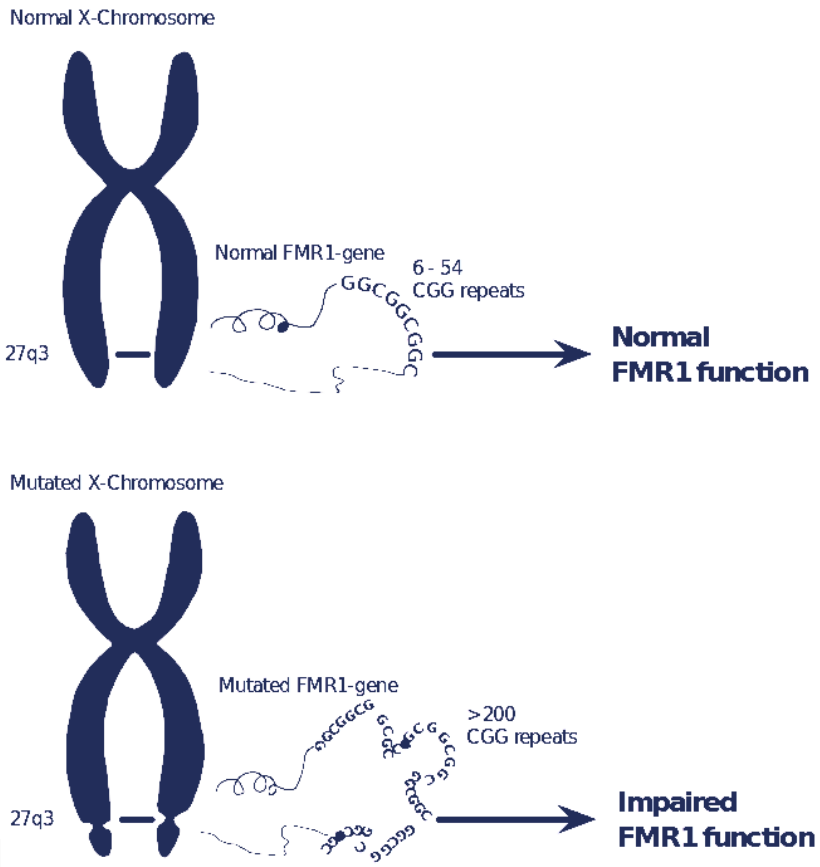
About 30 described disease entities

Causative role in

- Fragile X syndrome
- Huntington's Disease
- Friedreich's ataxia
- Spinocerebellar Ataxias
- c9FTD/ALS

Shared mechanism:

- Expansion of a short DNA repeat alter the function of a gene or may produce toxic molecules.



Pathogenic Short Tandem Repeat expansions
Nanopore data generation and analysis

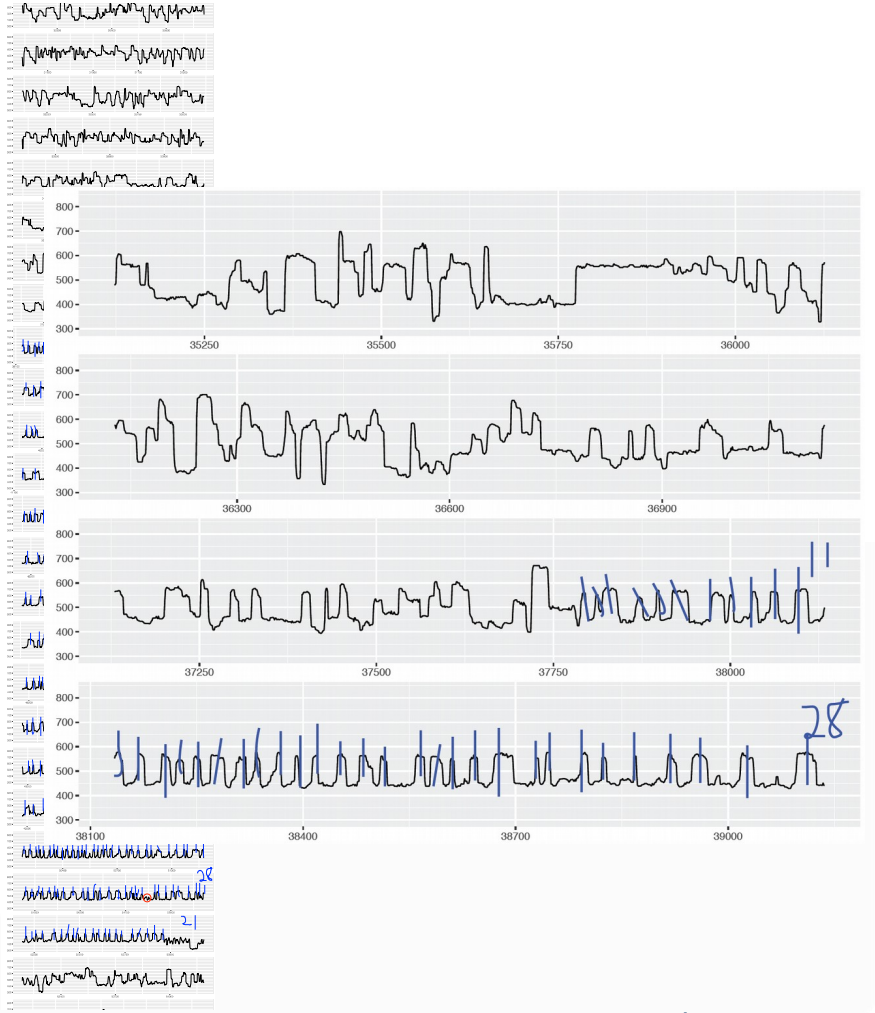
IDEA (to do's, straight forward)

- Sequence (synthetic) plasmids with known repeat lengths
- Build tool for more exact repeat length quantification
- Solve real world problem with DNA from actual patients
- Ta-da!

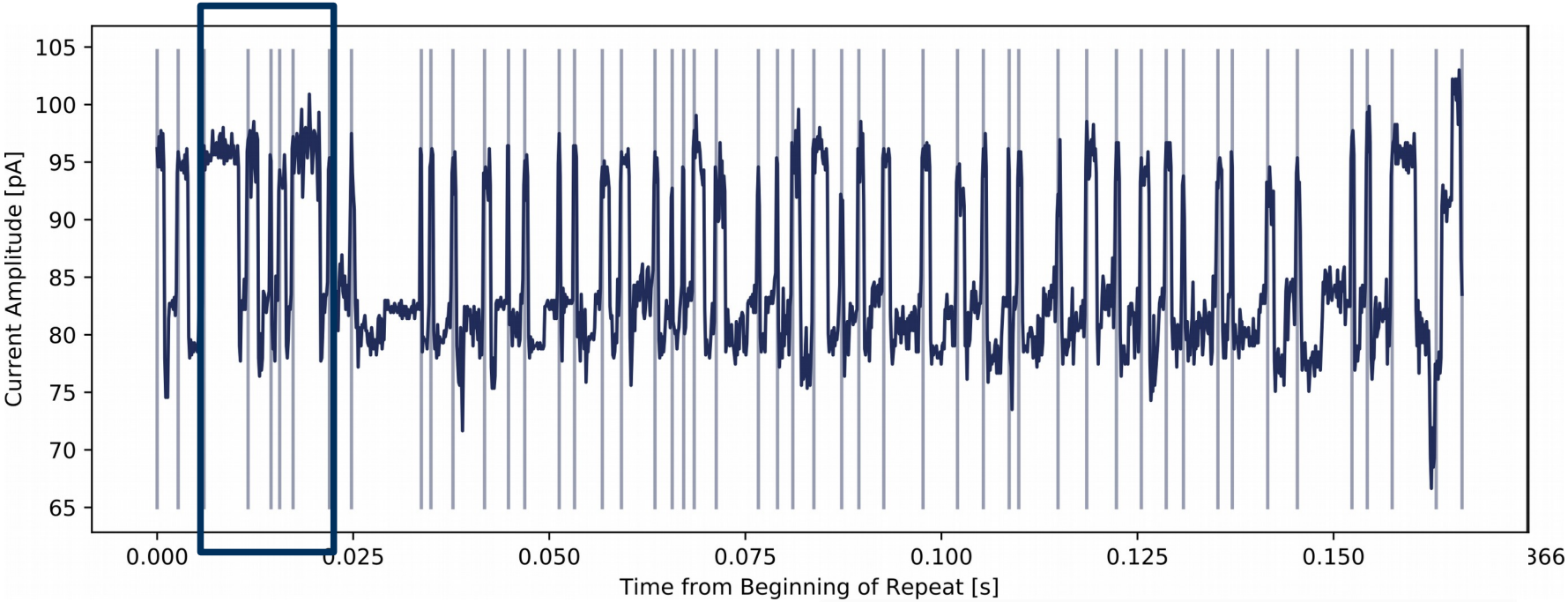
REALITY

- Repeat number can be estimated by hand in nanopore traces
- Current nanopore analysis workflows all fail to quantify expanded repeats
- No experimental DNA repeats (Plasmids, BACs, cell lines and tissues) appear to be stable

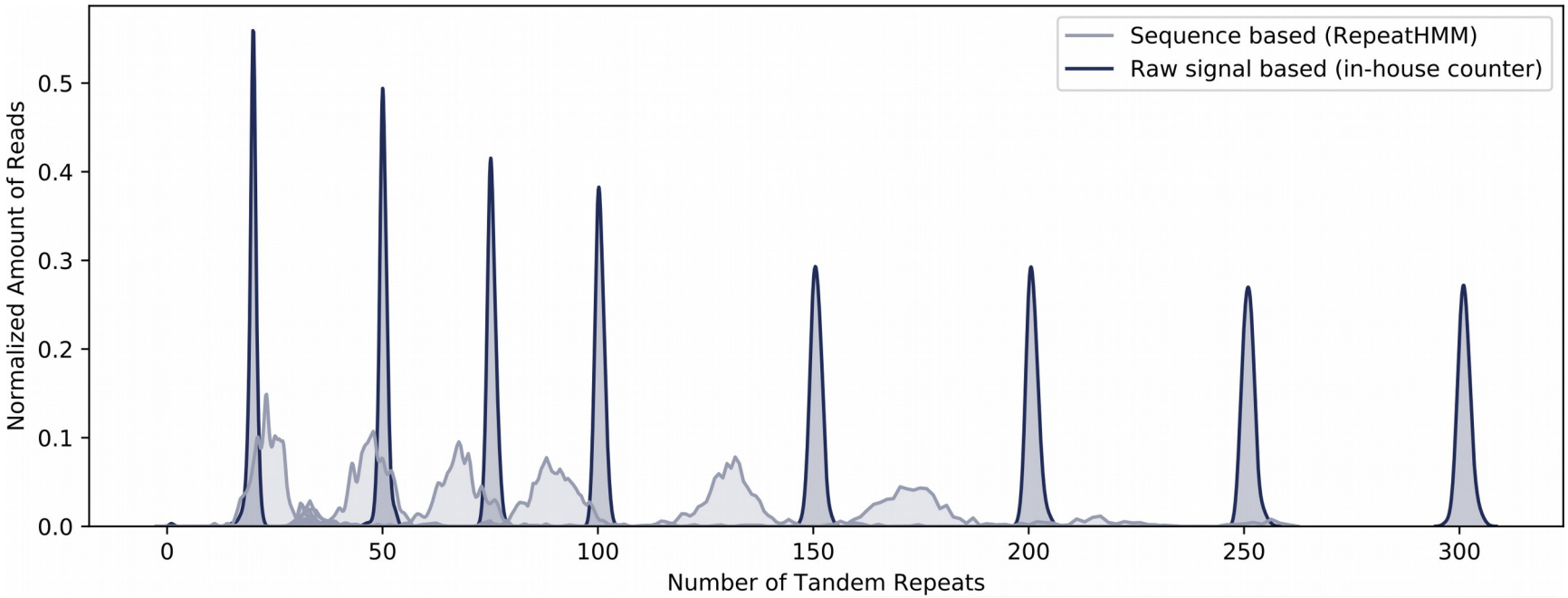
There is no "simple ground truth" to start from!



Pathogenic Short Tandem Repeat expansions
How does the simulation look like?



Pathogenic Short Tandem Repeat expansions
Benchmark with established methods



Conclusion
&
Outlook

Conclusion

We have developed a modular software tool that:

- simulates the nanopore raw signal
- adapts to different use cases
- derives relevant information form real experiments

Nanopore SimulatION

- will enable bioinformaticians to develop and optimize their algorithms in raw signal space without running flow cells
- could shorten the question-to-answer for hypothesis-driven, biological questions by in silico optimized molecular workflows

Outlook

- We’d like to test the best methods for RNA sequencing
  - Not only in space
- We plan to adapt the simulator to PromethION output data
- We will explore more clinical use cases that are on the horizon

Acknowledgment

- Zentrum für Integrative Psychiatrie Kiel
  - Dr. Franz-Josef Müller
  - Bernhard Schuldt
  - Björn Brändl
- Max Planck Institute for Molecular Genetics Berlin
  - Pay Gießelmann
- Kiel University of Applied Sciences
  - Prof. Dr. Ulrich Jetzek
  - Nadine Kraft
  - Veronika Polke

Funding

This work was partially 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.

Questions?

???