# SelectION: Rapid Identification of Predefined Genomic Regions in Large Nanopore DNA Sequencing Datasets

Pay Gießelmann,[1] Bernhard Schuldt[2], Alexander Meissner[1] and Franz-Josef-Müller[1,2]

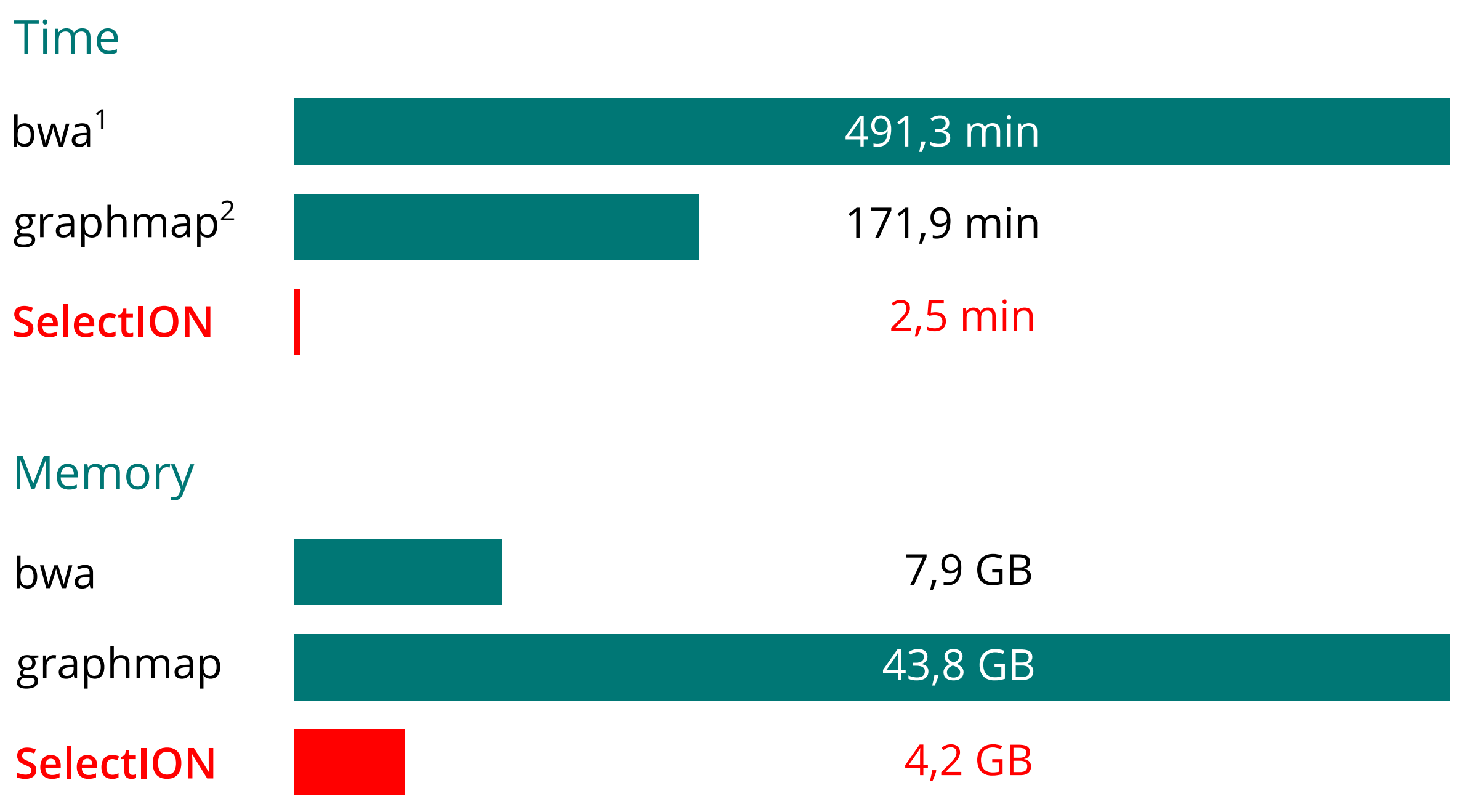
[1] Max Planck Institute for Molecular Genetics, Department of Genome Regulation, 14195 Berlin, Germany
[2] Universitätsklinikum Schleswig-Holstein Campus Kiel, Zentrum für Integrative Psychiatrie gGmbH, 24105 Kiel, Germany

### Abstract

Going from bacterial to mammalian genomes increases the computational complexity of the Nanopore sequencing pipeline significantly. Among other issues, sequence alignment is considered a major bottleneck, commonly taking days on typical workstation computers. As the focus shifts towards structural variations and DNA modifications, the alignment step serves only as a starting point for a detailed analysis based on raw signal or event data. We present SelectION, a lightweight approximate alignment tool which allows Nanopore reads to be rapidly anchored to a reference. Using SelectION it is possible to work with basecalled Nanopore data on a laptop. As a proof of concept we apply this method to the identification of repeat expansions in the human genome and use the selected data to quantify repeat lengths. Application of SelectION in conjunction with alignment of distinct reduced datasets enables the rapid identification of repeat expansion carriers or qualitative analysis of its epigenetic modifications.

## 1 Background

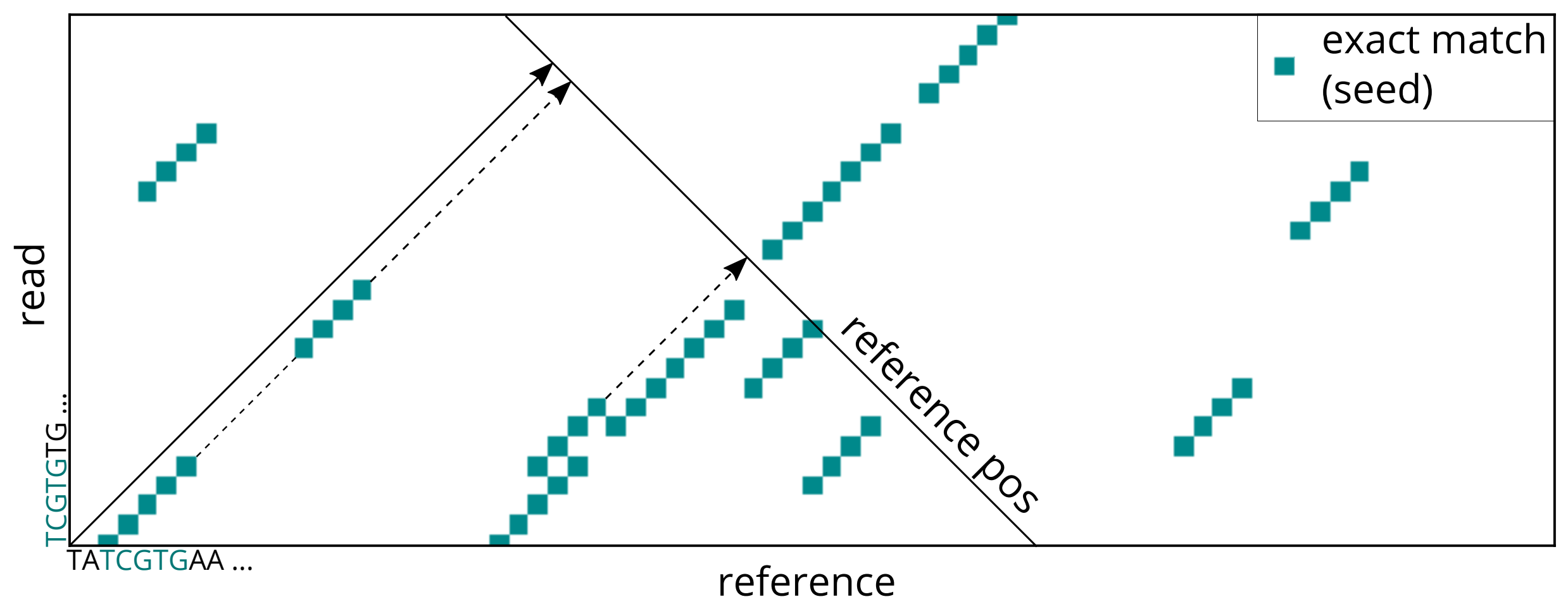
Sequence Alignment/ Anchoring of MinION DNA long reads against the human genome h38 in minutes:



Alignment/ Anchoring of 75k Nanopore reads (FAB47779 ONT-HG1[3])
8 threads Intel Xeon E5-2683 v3 @2.0 GHz

## 3 Methods

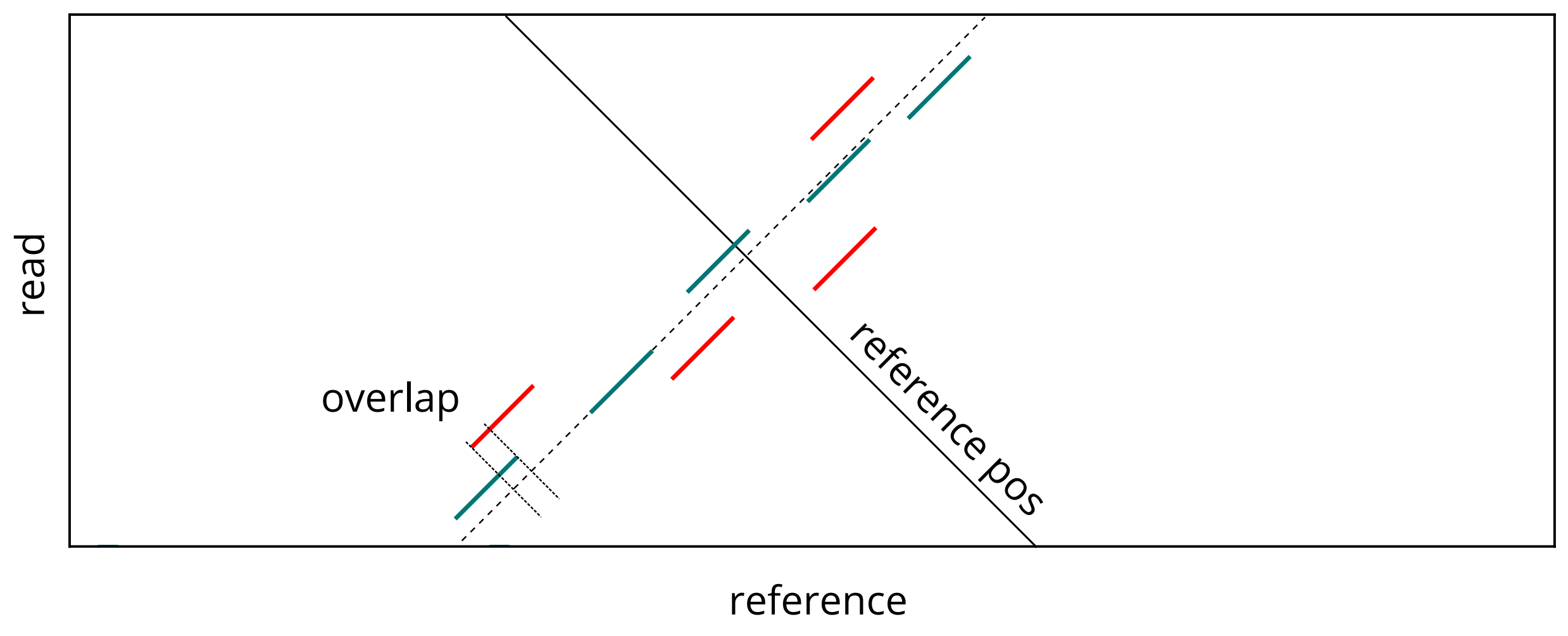
Long Common Subsequence Seeds



Seeds are computed as the longest common subsequences (lcs) of reference and a sliding window over the first N basepairs of a read. Matches are expected to cluster at the origin of the read in the reference. A precomputed FM-Index is used for fast and memory efficient exact match lookups.

Seeds are projected to a common baseline representing the position in the genome. Clustering is performed with respect to the lcs length, a dense region of long exact matches is considered to be around a likely read position.

Resolving Seed Overlaps



Overlapping of seeds is resolved by discarding matches with higher distance to the median seed reference position. The pseudo-alignment cigar is then constructed out of the remaining seeds. The alignment position is the index of the first matching base of the first lcs.

## 4 Discussion

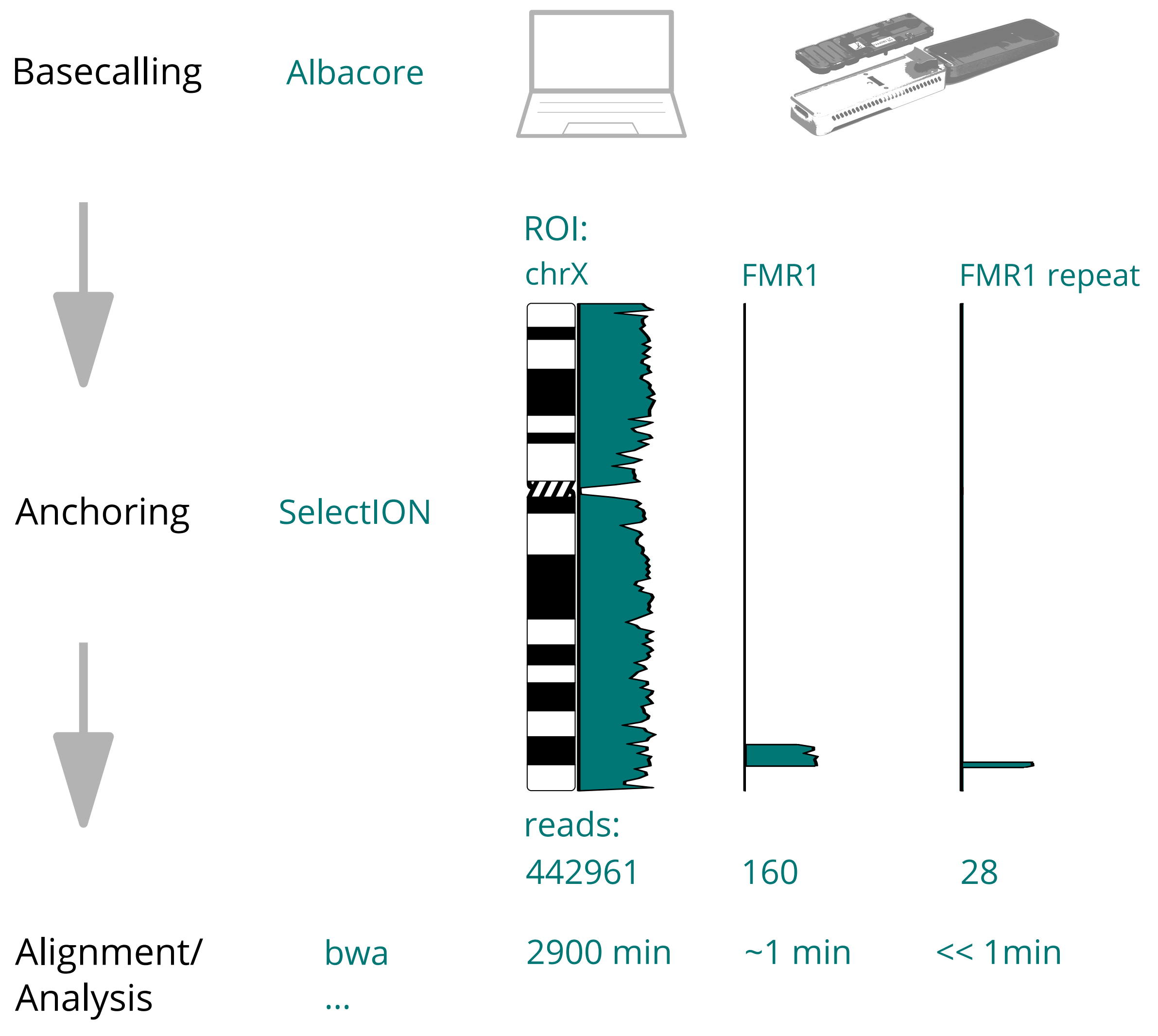
The development of sequence alignment methods for long, noisy reads from Nanopore sequencing experiments has initially been a major technical challenge. Several innovative and involved methods have been successfully implemented since[2,4,5].

Here, we demonstrate that the considerably improved ONT R9.x pores, sequencing chemistries and library preps with longer read lengths and overall lower error rates allow for an extremely fast approximate read localization on large reference genomes.

Selective sequencing[6] is a major promise of ONT's Nanopore platform. Our method would be compatible with the level of real-time read alignment throughput required for selective sequencing on the PromethION platform. Assuming the development of near real-time, low-latency base calling algorithms, SelectION could become an enabling step in real time selective sequencing algorithms suitable for mammalian genomes. This remaining bottleneck could be overcome by recent developments in regard to the Long-Short-Term Memory Recurrent Neural Network (LSTM-RNN) technology.

## 2 Results

### Workflow



Basecalling
Albacore

Anchoring
SelectION

Alignment/ Analysis
bwa
...

ROI: chrX
FMR1
FMR1 repeat

reads:
442961
160
28

2900 min
~1 min
<< 1min

SelectION serves as an intermediate step between base calling and alignment. Input data is specified as either a directory of basecalled reads or a FASTQ/ FASTA file of merged sequences. Output is a SAM file with the estimated read positions. Additionally reads covering regions of interest can be written to seperate output files/ folders for further analysis.

SelectION is fast and memory efficient, allowing the execution on any sequencing computer, possibly even parallel to experiment and basecalling in the near future.

**Mapping > 90%**

**Accuracy > 90%**

**92,5%** reads mapped to reference

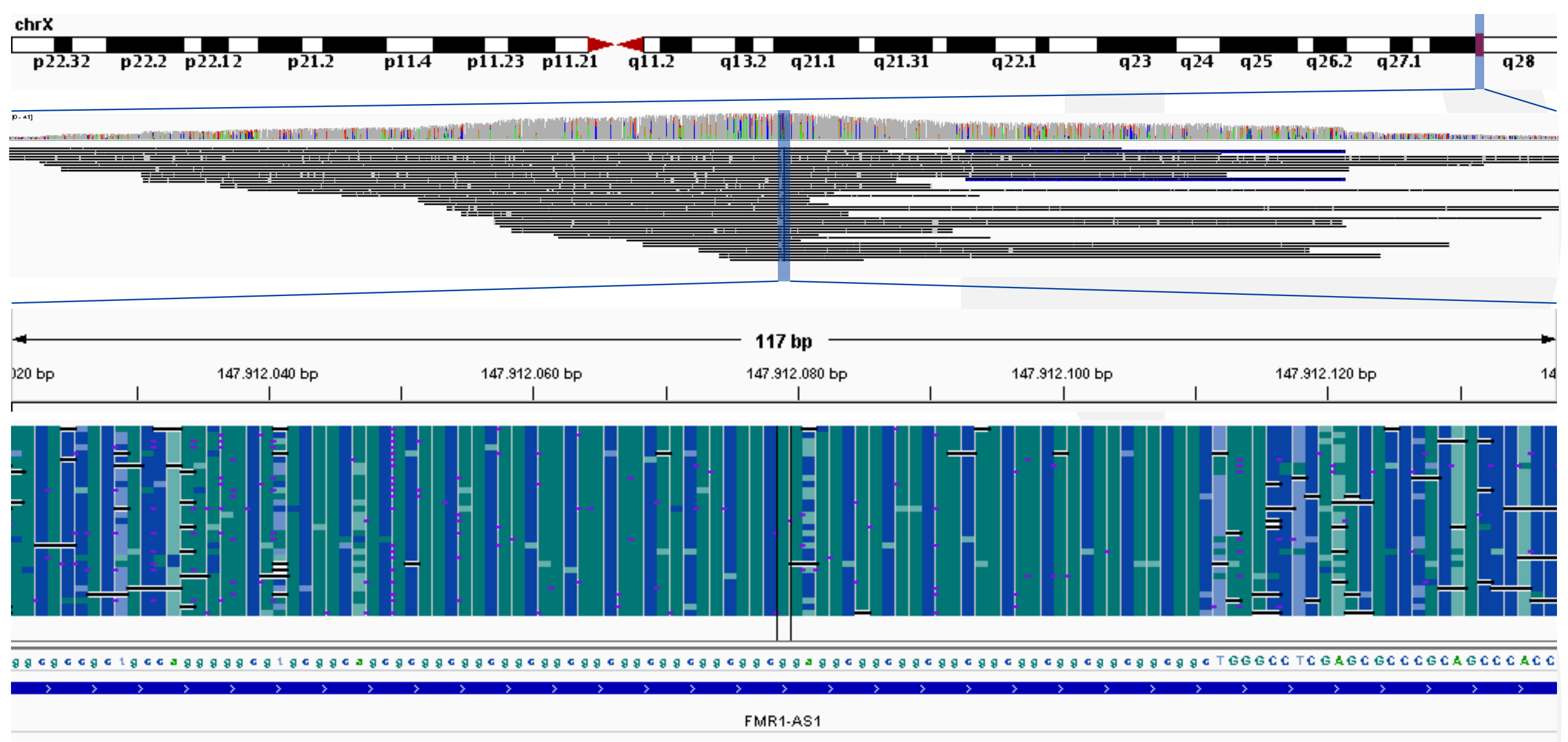
| Dist. to bwa pos. | % of mapped reads |
|---|---|
| 100 Bp | 72,9% |
| 200 Bp | 86,4% |
| 500 Bp | **93,7%** |

### Applications

**Targeted analysis** of any given spot or region in the reference from whole genome shotgun sequencing: We applied our tool to extract all reads from the publically available Cliveome (ONT-HG1) covering a repetive sequence in FMR1.

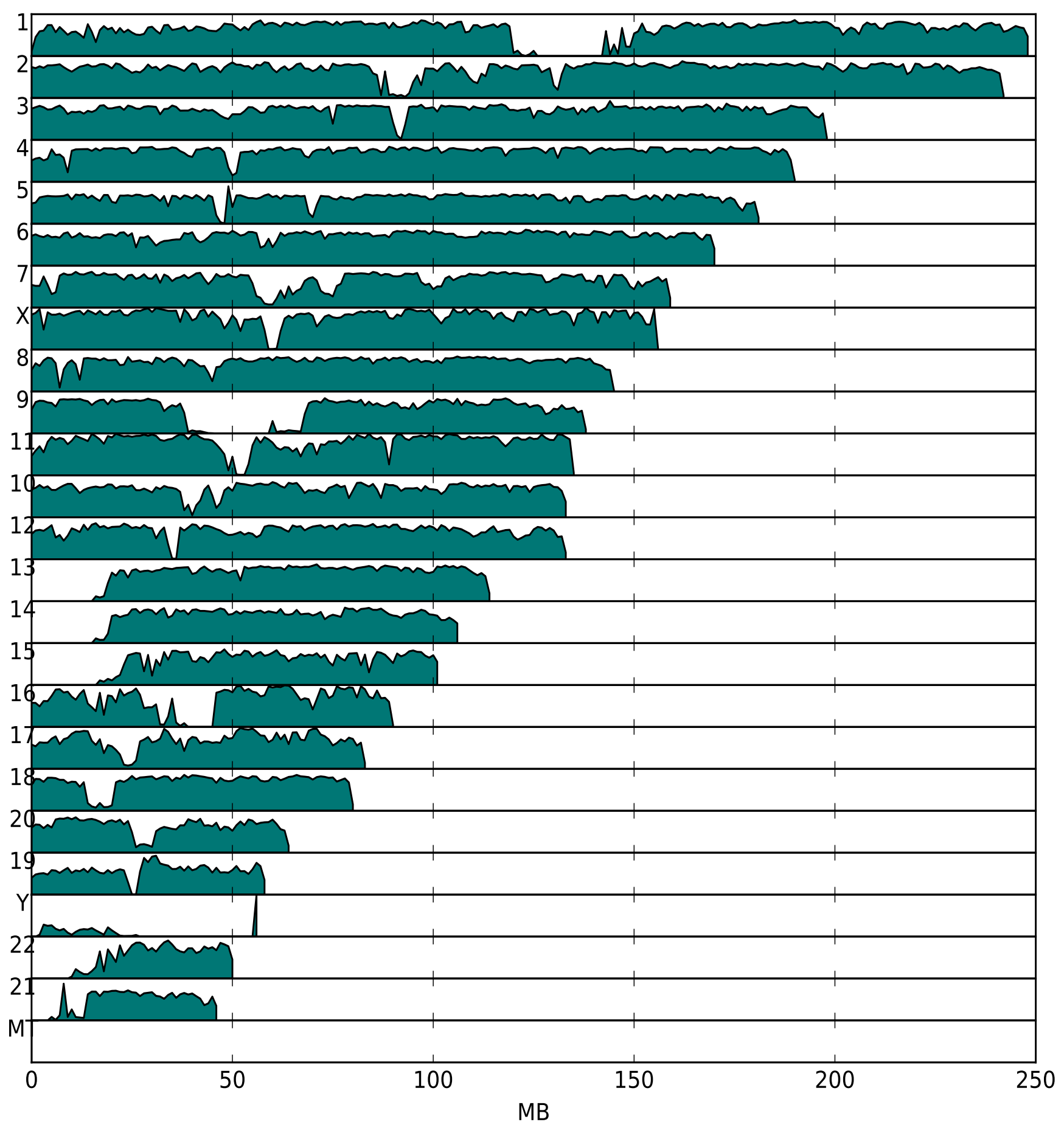
The search space of 51 Flow-Cells can be analyzed on a Laptop in less than a day.

In the given dataset the region of interest is covered by at least 41 reads, allowing for a reliable quantification of the repeat by alignment of the selected reads.



| Disease | Locus hg38 | Gene | Repeat | Length | Coverage |
|---|---|---|---|---|---|
| Huntingtons's disease | chr4: 3 074 875 | HTT | CAG | 21 | 44 |
| Spinocerebellar ataxia type 2 | chr12: 111 598 951 | ATXN2 | CAG | 23 | 45 |
| Friedreich ataxia | chr9: 69 037 284 | FXN | GAA | 7 | 49 |
| Fragile X syndrome | chrX: 147 912 042 | FMR1 | GCG | 24 | 41 |

Coverage of Cliveome (ONT-HG1)



**Rapid Coverage Evaluation**: Our approximate alignment may be used to quantify overall coverage of one or multiple Nanopore sequencing datasets. Expecting real-time basecalling to be available, this allows dynamic decisions on the amount of resources spent to reach a targeted coverage.

### Availability

Code and instructions coming soon on GitHub

https://github.com/PayGiesselmann/selectION

References:

[1] Li, H. & Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics*, 25, 1754-1760
[2] Sovic, I. *et al.* (2016) Fast and sensitive mapping of nanopore sequencing reads with GraphMap, *Naure Communications*, 7, 11307
[3] Brown, C. (2016) Cliveome ONT-HG1 https://github.com/nanoporetech/ONT-HG1
[4] Li, H. (2015) Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences, *arXiv:1512.01801*
[5] Koren, S. *et al.* (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation, *Genome Research*
[6] Loose, M. (2016) Real-time selective sequencing using nanopore technology, *Nature Methods*, 13, 751-754