High-throughput repeat expansion sequencing on PromethION: An *ABCA7* VNTR case study

Arne De Roeck[1,2], Wouter De Coster[1,2], Tim De Pooter[3], Svenn D'Hert[3], Jasper Van Dongen[1,2], Mojca Strazisar[3], Christine Van Broeckhoven[1,2], and Kristel Sleegers[1,2]

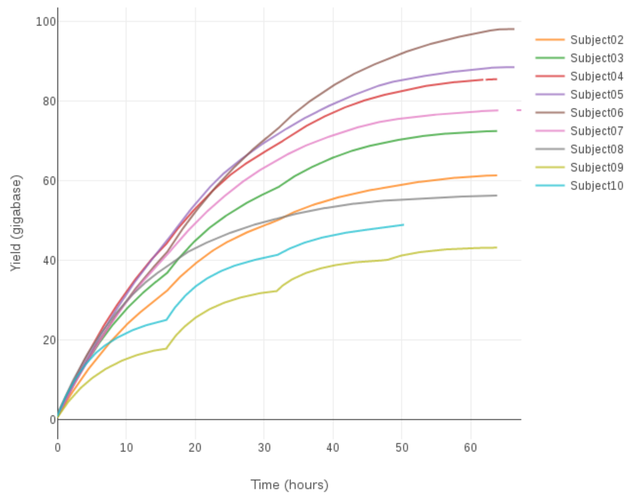
1 Neurodegenerative Brain Diseases group, Center for Molecular Neurology, VIB - University of Antwerp, Antwerp, Belgium
2 Institute Born-Bunge, University of Antwerp, Antwerp, Belgium
3 Neuromics Support Facility, Center for Molecular Neurology, VIB - University of Antwerp, Antwerp, Belgium

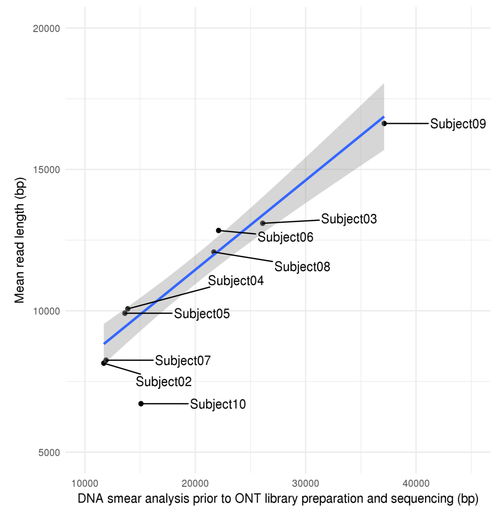
Background

Expanded *ABCA7* VNTR alleles are a strong risk factor for Alzheimer's disease. Only Southern blotting can be used to estimate the length of this tandem repeat, which precludes in-depth characterization and application in the clinic. Here, we aimed to find an alternative through long-read whole genome sequencing on the recently commercialized PromethION platform (Oxford Nanopore Technologies).

Robust Oxford Nanopore PromethION sequencing on 1 flow cell per genome

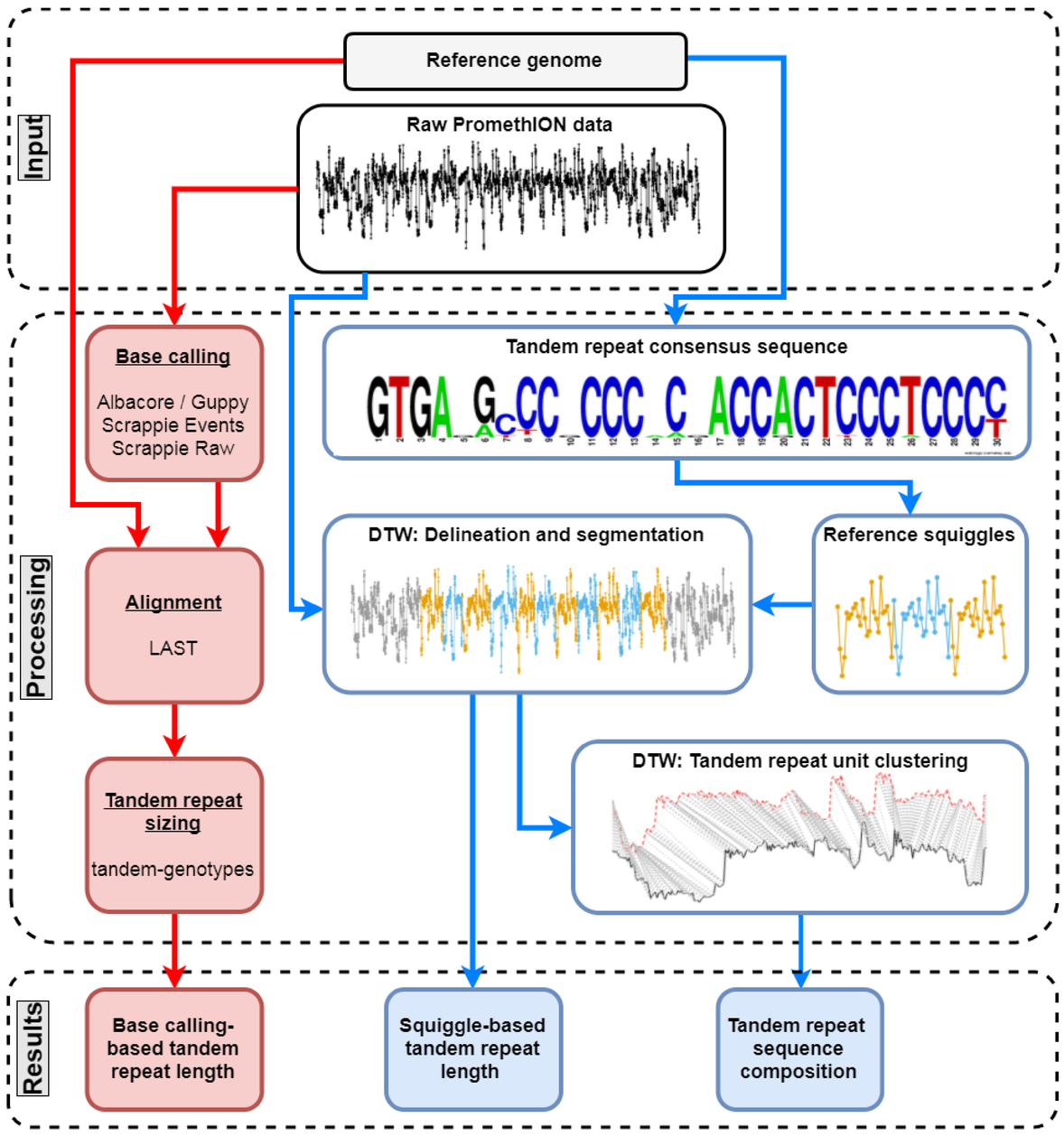
Individual	Number of flow cells	Yield (Gb)	Read length N50 (kb)
Subject01	6+7	74.4	7.0
Subject02	1	61.3	11.2
Subject03	1	72.5	16.3
Subject04	1	85.5	11.5
Subject05	1	88.5	11.9
Subject06	1	98.0	16.2
Subject07	1	77.7	10.6
Subject08	1	56.2	14.7
Subject09	1	43.2	29.3
Subject10	1	48.9	9.1
NA19240	5	220.0	15.8





- **Recommendation: fresh DNA, 20kb shearing and size selection.**
- **Average yield: 71 Gb (22x genome coverage) per flow cell, maximum: 98 Gb (30x)**
- **Mean read length N50: 14 kb**

NanoSatellite: Tandem repeat characterization based on raw current squiggle data

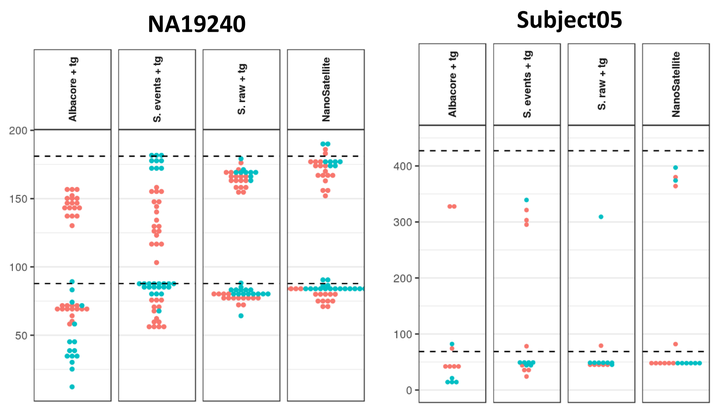


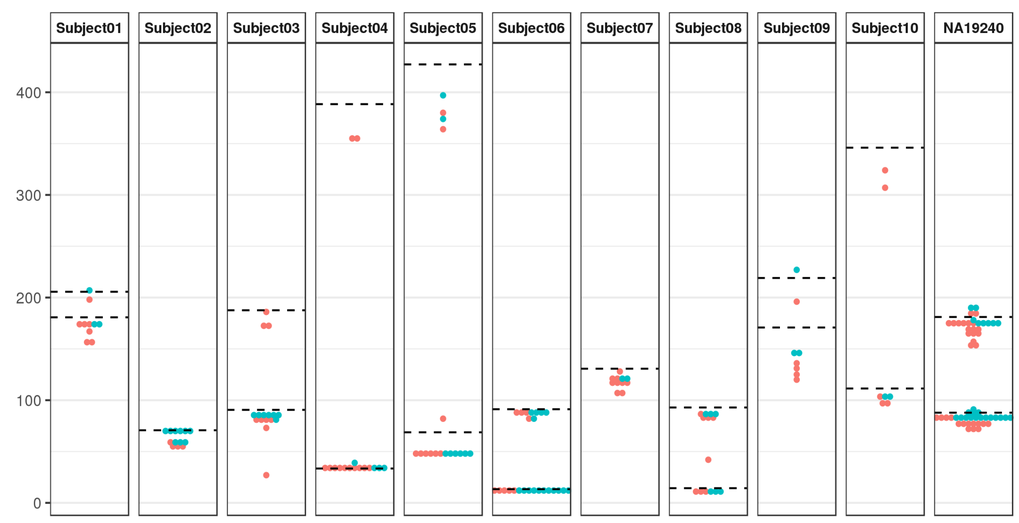
Two approaches were followed to analyze tandem repeats: existing methods (red) and our squiggle-based algorithm NanoSatellite (blue).

Accurate tandem repeat length estimation

- **NanoSatellite outperforms existing methods, which are based on base calling and alignment to a reference genome**
- **NanoSatellite works well for all *ABCA7* VNTR lengths, especially expansions**

Method	Accuracy (%)	Relative standard deviation (%)	Expanded read detection (%)
Albacore	65.6	36.0	63
Scrappie events	83.3	15.5	88
Scrappie raw	87.7	**5.5**	25
Nano-Satellite	**90.3**	**5.6**	**100**

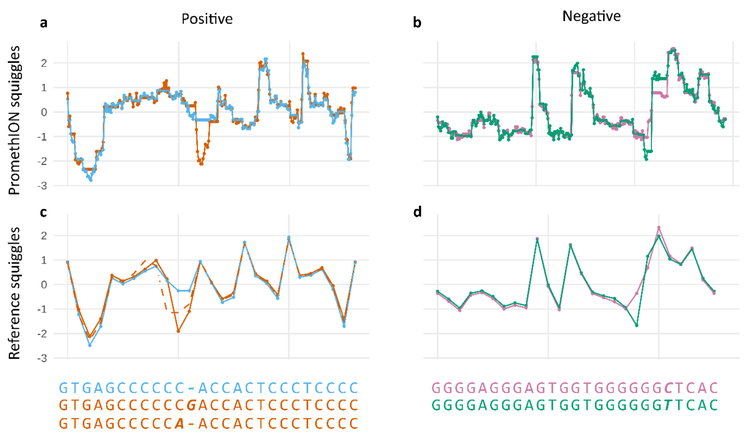




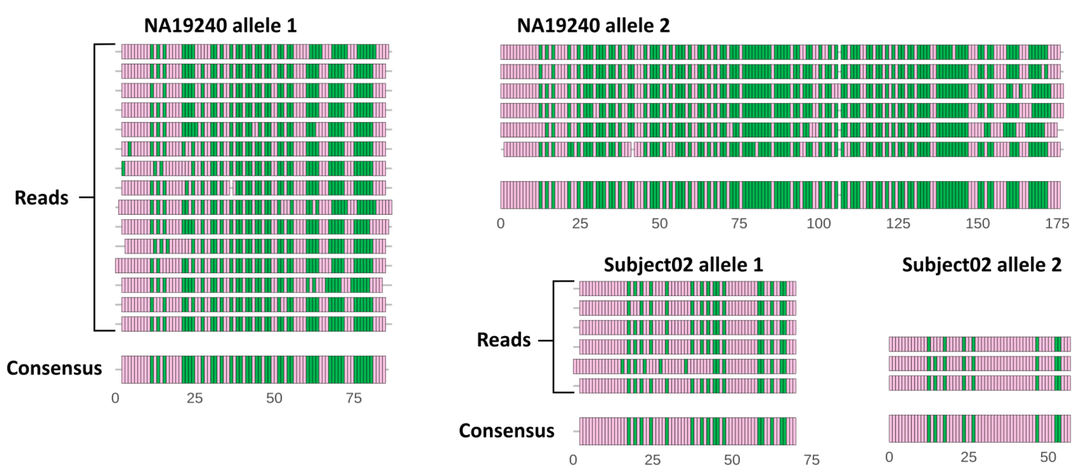
Figures legend: Tandem repeat length estimates (the number of tandem repeat units is depicted on the y-axis) per positive strand (red) or negative strand (blue) PromethION sequencing reads (dots) are shown in comparison to the Southern blotting lengths (dashed lines).

High consistent tandem repeat sequence determination

- **Clustering of tandem repeat unit squiggles for positive and negative DNA strands enables detection of underlying nucleotide changes.**
- **High consistent clustering patterns for reads of the same allele ➔ reliable nucleotide sequence determination.**
- **Separation of alleles with homozygous length.**



***ABCA7* VNTR squiggle clustering by NanoSatellite.** Centroids are shown which were extracted from hierarchical *ABCA7* VNTR squiggle unit clusters originating from positive **(a)** or negative **(b)** DNA strands. Each cluster is shown in a different color. We compared these centroids to positive **(c)** and negative **(d)** reference squiggles with corresponding sequence motifs shown below.



***ABCA7* VNTR sequence reconstruction based on squiggle clusters.** Each rectangle corresponds to a tandem repeat unit. Colors correspond to the tandem repeat unit cluster as assigned in the previous figure.

Conclusions

- **PromethION can be used to study (expanded) tandem repeats**
  - **1 flowcell provides sufficient human genome coverage**
  - **All ABCA7 VNTR alleles [1kb - 10kb] were spanned with sequencing reads**
- **Developed NanoSatellite: tandem repeat analysis on raw current data**
  - **Accurate length estimation**
  - **Consistent nucleotide sequences**

Preprint: https://www.biorxiv.org/content/early/2018/10/09/439026
GitHub: https://github.com/arnederoeck/NanoSatellite

SCIENCE MEETS LIFE