Preliminary results from Nanopore Q20+ sequencing

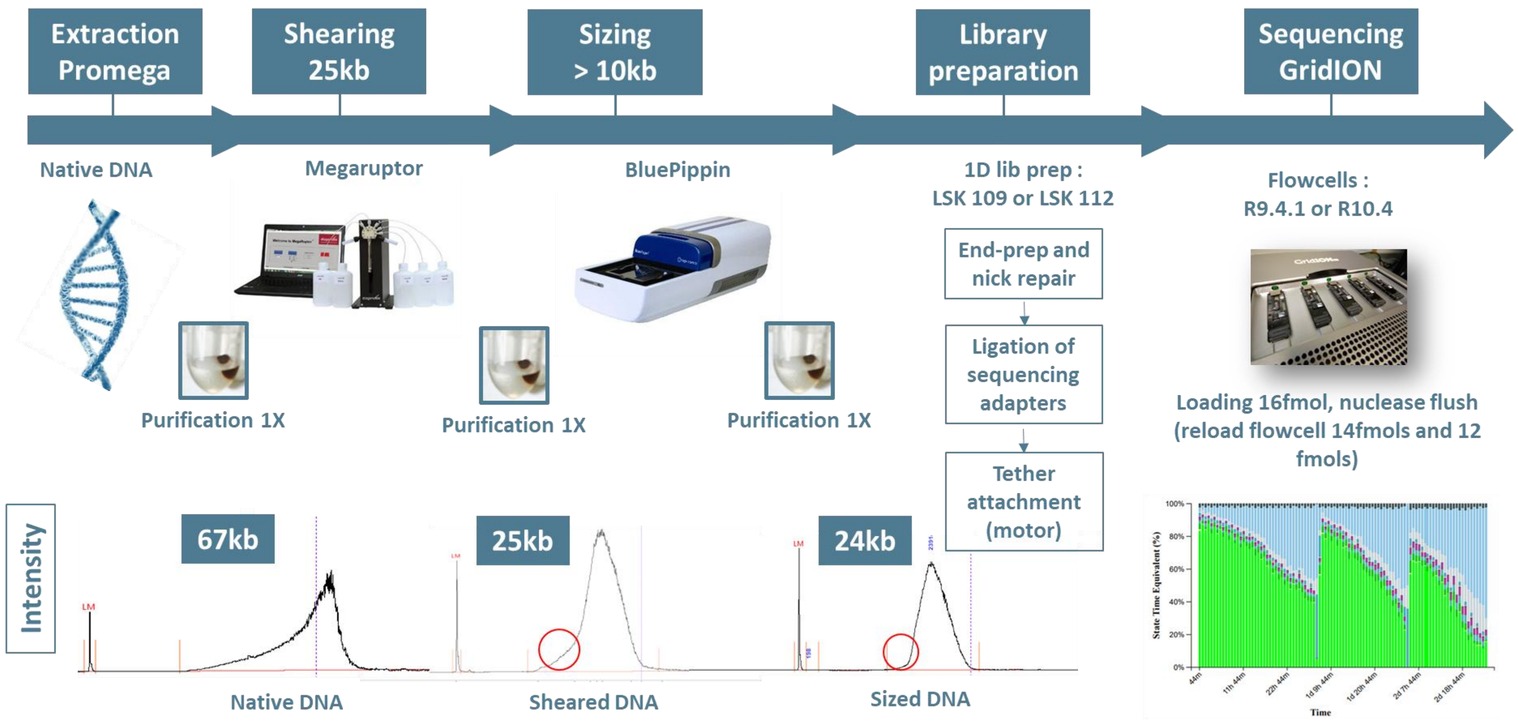
Claire Kuchly, Jules Sabban, Eden Darnige, Camille Eché, Amandine Suin, Céline Vandecasteele, Clement Birbes, Andreea Dreau, Christophe C. Klopp, Christine Gaspin, et al.

Preliminary Results from Nanopore Q20+ Sequencing

Jules SABBAN[1], Camille ECHE[1], Joanna LLEDO[1], Amandine SUIN[1], Céline VANDECASTEELE[1], Eden DARNIGE[1], Clement BIRBES[2] , Andreea DREAU[2], Christophe KLOPP[2], Christine GASPIN[2], Denis MILAN[1], Carole IAMPIETRO[1], Cécile DONNADIEU[1], Gérald SALIN[1], Céline LOPEZ-ROQUES[1] and Claire KUCHLY[1]

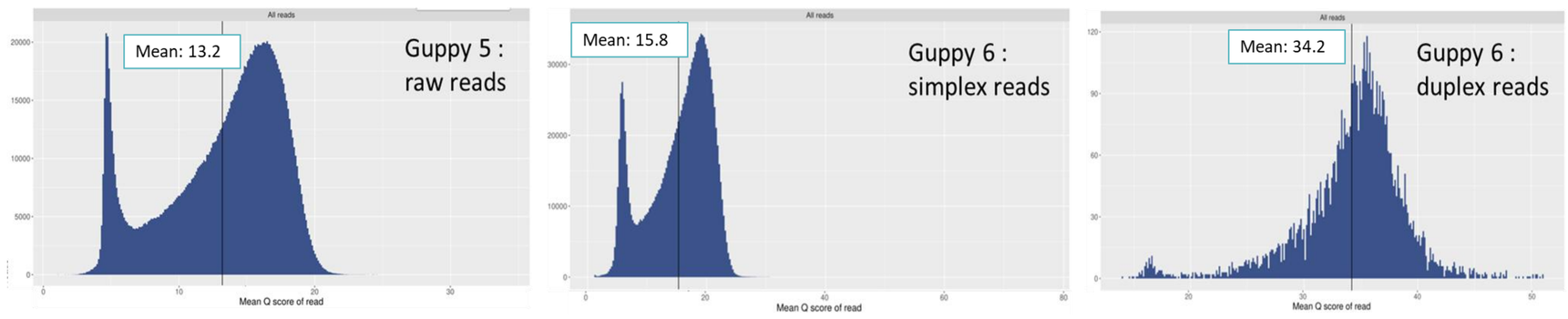
[1] *INRAE, US 1426, GeT-PlaGe, Genotoul, 31326, Castanet-Tolosan, France*
[2] *INRAE, UR 875 , MIAT, Plateforme bioinformatique, 31326, Castanet-Tolosan, France*

Library preparation



Library preparation was performed according to the manufacturer's instructions "1D gDNA selecting for long reads (SQK-LSK109 or SQK-LSK112)." At each step, DNA was quantified using the Qubit dsDNA HS Assay Kit (Life Technologies). DNA purity was tested using a Nanodrop (Thermofisher) and size distribution and degradation assessed using the Fragment analyzer (AATI) High Sensitivity DNA Fragment Analysis Kit. Purification steps were performed using AMPure XP beads (Beckman Coulter). For 1 Flowcell, 5µg of DNA was purified then sheared at 25kb using the megaruptor system (diagenode). A size selection step using the BluePippin Size Selection system (Sage Science) was performed. A one-step DNA damage repair + END-repair + dA tail of double stranded DNA fragments was performed on 1µg of sample. Then, adapters were ligated to the library. The library was loaded onto a R9.4.1 revD flowcell or onto a R10.4 flowcell and sequenced on GridION instrument at 16 pmol within 72H. Nuclease flush steps were done when necessary and possible, i.e. when 30% of the pores were still in sequencing. *There is no important difference between LSK109 and LSK112 despite the loading, which is twice less for LSK112.* Q20+ is a new chemistry combining new pores (R10.4) and a new motor enzyme (E8.1). This combination allows for the acquisition of high quality reads, especially in the case of duplex reads where one fragment is sequenced twice. The new basecaller, Guppy 6, can analyze these particular reads and produce a consensus with high confidence.

Raw data analysis



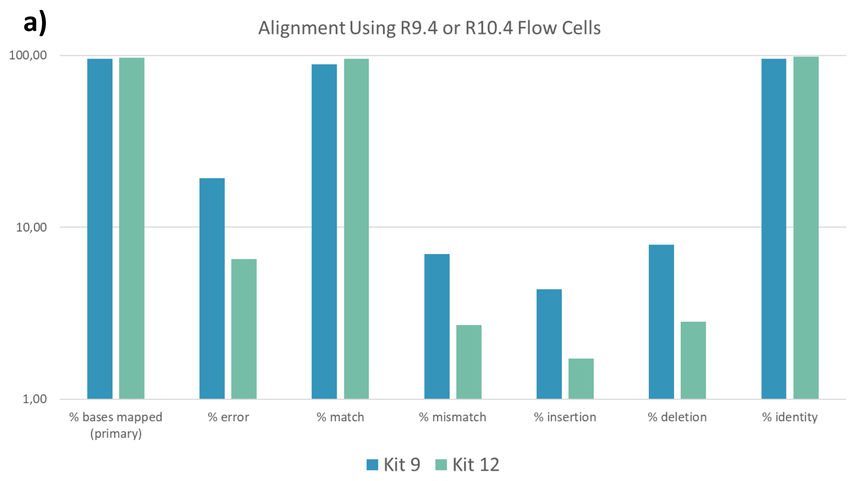
The above graphs were generated with MinIONQC [1]. Simplex reads from Guppy 6 are of better quality than Guppy 5 basecalled reads. Globally, the Qscore distribution trend is similar between raw reads with Guppy 5 and simplex reads with Guppy 6. Qscores seem to have slightly improved with the most recent version of the basecaller. A few outlying reads have very high Qscores.

Duplex reads have a high Qscore of around Q35, indicating an error rate of < 0.1%. However, Oxford Nanopore Technology documentation states that Qscores above Q10 are underestimated by their software. The limiting factor is the quantity of reads. Duplex read sequencing is currently in development.

Kit & software comparisons

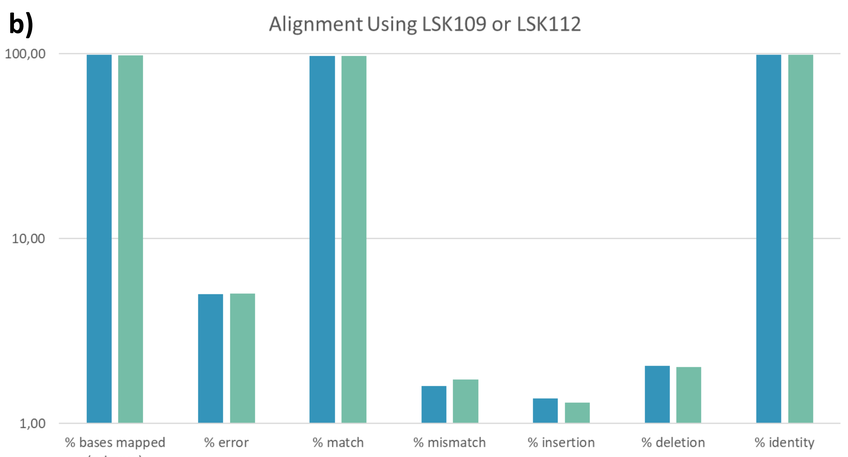
The aim of this comparison is to evaluate the improvement brought on by the new chemistry (Kit 12) as compared to the old one (Kit 9) in terms of the quality of the reads at the output of the sequencers. To make these evaluations, we mapped reads from different origins on the same reference genome (internal assembly made from 10X, Hi-C and ONT GridION data).

a)

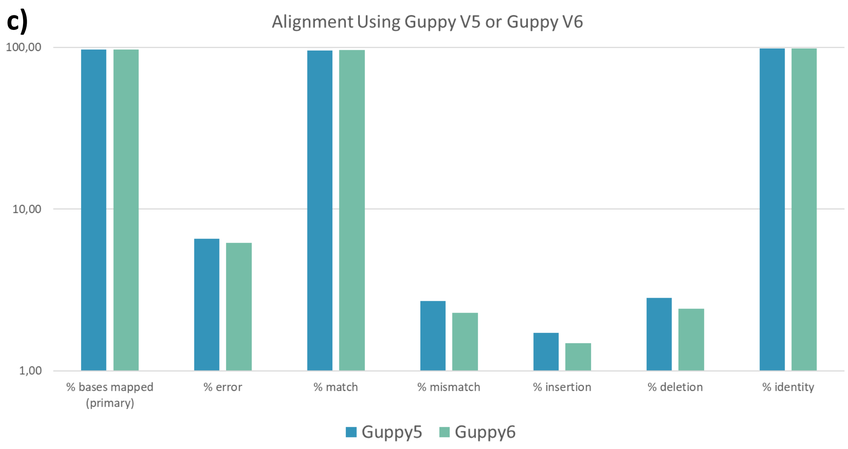


a) Comparison between R9.4 from 2019 and current R10.4 flow cells. Old data was re-basecalled with Guppy 5 to highlight the importance of the chemistry used. The high accuracy of Q20+ data is not solely due the Guppy version; the same basecaller produces a different output depending on the type of pore used.

b) Read accuracy comparison of LSK109+R9.4.1 and LSK112+R10.4 using the same software treatment (Guppy 5). No major difference in read quality; duplex reads will make the difference?



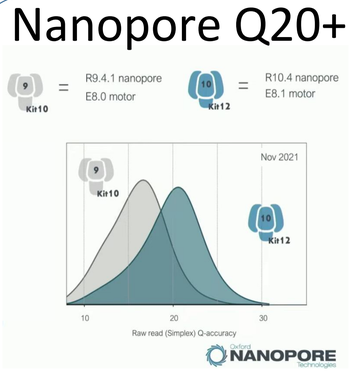
c)



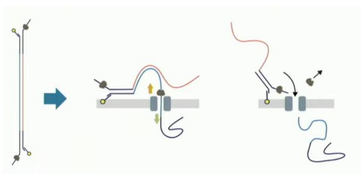
c) Here, the same data from Kit 12 were basecalled with the two versions of Guppy (5 and 6), to assess the importance of Guppy 6 in the analysis of Q20+ reads.
Guppy 6 shows small improvement at non-matching positions (mismatches, insertions or deletions). This could be even more blatant given a higher duplex reads proportion.

These results are hopeful for the future. Oxford Nanopore Technology is currently working to increase the duplex rate, announcing that we could soon have around 40% duplex reads.

Nanopore Q20+ announcements

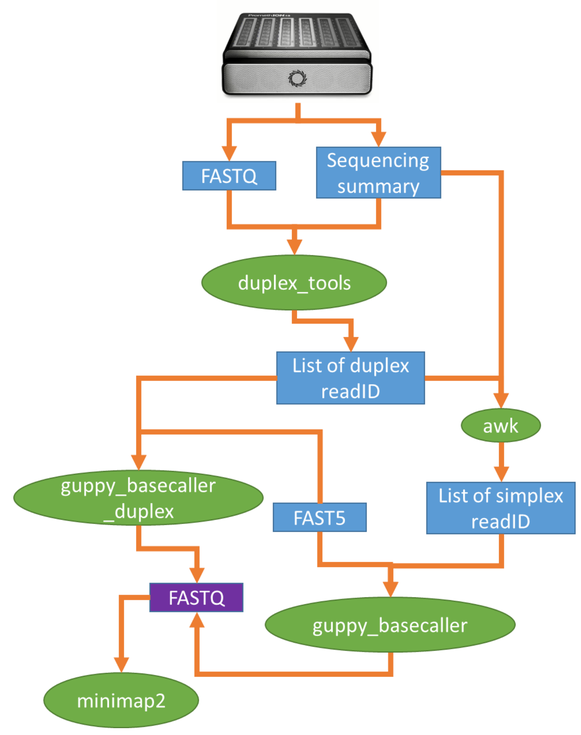


New Flowcell and motor protein slow molecule passage and increase sequencing quality.



Duplex reads:
Complementary molecules from the same duplex strand are sequenced consecutively for higher accuracy.

How the data were analyzed



Once data are generated by the GridION or PromethION using the corresponding preparation kit and machine settings, it is necessary to perform custom analyses that are not currently integrated in MinKNOW.

MinKNOW performs the first step: the live basecalling. For this step, different models of accuracy can be chosen. This first basecalling is useful to identify which reads are duplex. Oxford Nanopore Technology has developed Duplex Tools [2] for duplex read analysis. We use the "pairs_from_summary" tool, which identifies potential duplex reads based on the sequencing summary file.

Then, the "filter_pairs" tool uses sequences in the FASTQ files to align each read association by pairs and checks if they are indeed complementary to each other. A list of sure duplex reads is exported.

By comparing the list of duplex reads to the list of reads in the sequencing summary, it is possible to identify the non-duplex reads with some AWK lines.With these lists of reads and FAST5 files, separate basecalling with super accuracy mode can be performed using Guppy 6.

After that, output files are merged to align the reads to a reference genome with minimap2 [3]. Statistics from the alignments are calculated from the BAM file using a Python script provided by the CEA.

Conclusions & perspectives

Oxford Nanopore Technology is continuously improving. Library preparation with Kit 12 and basecalling of reads produced with Guppy 6 in super accuracy mode makes it possible to obtain very high quality data, especially when compared to the old chemistry. The quality of output data from Guppy 6 also heavily depends on the training data of the model used. The species that we sequenced is not part of the training dataset.

We are expecting the release of a new basecaller: Dorado. Furthermore, Oxford Nanopore Technology will soon release Kit 14, which is an upgraded version of Kit 12. The Q20+ is in its early phases and promises many new improvements in the near future.

References & Acknowledgments

[1] minionQC: R Lanfear, M Schalamun, D Kainer, W Wang, B Schwessinger, MinIONQC: fast and simple quality control for MinION sequencing data, *Bioinformatics*, Volume 35, Issue 3, 01 February 2019, Pages 523–525, https://doi.org/10.1093/bioinformatics/bty654

[2] duplex_tools: https://github.com/nanoporetech/duplex-tools

[3] minimap2: https://github.com/lh3/minimap2

We thank the GenoToul bioinformatics facility for their support in computing resources and data storage, and for helping us process the data and making available all the needed software and infrastructure.

We thank the Genoscope with whom we have always had very fruitful exchanges and who shared their test script with us.

Projet cofinancé par le Fonds Européen de Développement Régional