DOI 10.18699/SBB-2022-022

Less is more: filtering and trimming ONT sequencing data

Nenasheva N.V.[1, 2, 3]*, Ilyinsky V.V.[3], Makeev V.J.[1, 2]
[1] *Moscow Institute of Physics and Technology, Moscow, Russia*
[2] *Vavilov Institute of General Genetics, RAS, Moscow, Russia*
[3] *Medical Genetic Center Genotek, Moscow, Russia*
* *nenasheva.nv@phystech.edu*

Key words: Oxford Nanopore Technology (ONT), quality of ONT data, genome reference

Sequencing with Oxford Nanopore technology is rapidly gaining in its popularity as it has a number of advantages over sequencing by synthesis or other 1st and 2nd generation methods. Yet, this technology has its own limitations, particularly the unstable read quality. Since many tasks in bioinformatics, such as identification of genomic variants, structural changes, or epigenetic modifications, require high quality reads with unique genome mapping, we tested if the overall ONT data quality can be improved by filtering out low-quality reads or read fragments. To this end, we developed a set of criteria applied to the raw (fastq) base calling data, including the phred score, the level of read fragmentation and the sequence alignment coefficient normalized to the fragment length, which allowed us to remove poor quality fragments of long reads or even entire reads. To further improve the potential of ONT data we used the combined Illumina – ONT sequencing and mapped ONT reads to the "personal reference sequence" with corrected individual single nucleotide variants rather than to the standard database genome reference sequence. We assembled this reference sequence using PICARD (AddOrReplaceReadGroups, NormalizeFasta, CreateSequenceDictionary) and GATK (MarkDuplicates, BaseRecalibrator, ApplyBQSR, HaplotypeCaller, FastaAlternateReferenceMarker) tools. By so doing we visibly improved the quality of read mapping and alignment. We performed a series of comparative tests to assess contributions of the read errors and genome SNVs to mapping/alignment errors. Testing was performed on the data from six genomes belonging to the members of the same family obtained by Oxford Nanopore Technology sequencing, with the genome of one family member sequenced with Illumina. Both of these areas, improving the quality of sequencing data and improving the quality of mapping, can be applied together and used for analysis of individual DNA methylation with the help of Oxford Nanopore Technology data.