Review of: "Characterization and simulation of metagenomic nanopore sequencing data with Meta-NanoSim"

KLOPP Christophe

Potential competing interests: The author(s) declared that no potential competing interests exist.

The authors present a tools achieving three aims, first the accurate simulation of Oxford Nanopore metagenomic reads, second the detection of chimeric reads and last the quantification of metagenomic abundances from aligned Nanopore reads. They use real reads as input to extract different metrics in order to fit the error and chimeric read profile of the simulated reads. They present assembly results on simulated reads assemblies with Flye-meta showing that Meta-NanoSim simulated reads have an interest in checking assemblies at different coverages.

Only one other specialized simulation package for nanopore reads is already available CAMISIM and the authors show that Meta-NanoSim simulated reads improve real data fitting for fragment length as well error profiles.

Even if the test on the Even dataset has an interest it is quite far from real data, therefore the author should shift their presentation to data sets having an abundance profile closer to real data.

The authors should clarify if the chimeric read abundance can be given as a simulation input parameter. This would enable to check its impact on assembly metrics.

The authors have specialized in Nanopore based data type simulation and are adding, with this new tool, metagenomic data to their repertoire.
They have published trans-NanoSim last year in which they claim part of the same results : We introduce Trans-NanoSim, a tool that simulates reads with technical and transcriptome-specific features learnt from nanopore RNA-sequncing data. The authors should explain in which length the methodologies used in both tools overlap.

Background :

"due largely to the long read lengths it generates" could be replaced by "due to its important read length"

The N50 definition lacks read ordering.

Regarding “making it possible to disambiguate between even closely-related strains”, having a few reads does not enable to disambiguate between closely related strains if the reads have 8% errors. This should be further explained.

Missing reference about Nanopore errors https://pubmed.ncbi.nlm.nih.gov/32255181/ 'Inverted duplicate DNA sequences increase translocation rates through sequencing nanopores resulting in reduced base calling accuracy'. This should be added to the references and perhaps to the simulation engine.

The authors should explicit what they call deviation in abundance models or provide a bibliographic reference.

The authors should explain why simulation should be consistent with the quantification method. Both elements are not linked.

What is a compromised abundance model?

Results :

Meta-Nanosim design sub-section feels strange in the result section. Is there another place to locate it?

Regarding "In the characterization stage, it takes ONT metagenomic reads and a reference metagenome as input to infer the ground truth through sequence alignments.", which ground truth are we speaking about here? Error and chimera rates? This should be written explicitly.

The authors should also explain how the reference metagenome should be built. Because if you build it from the same read set then the assembly is likely to contain the errors you wish to find.

Regarding "Chimeric reads are also called split reads because one read is split into two or more sub-alignments that are aligned to distinct regions of the genome", the authors should use "supplementary alignment" which is now the standard way to call "sub-alignments that are aligned to distinct regions of the genome".

Replace "different seuqncing kit" by "different sequencing kit"

Replace "another logrighmically distributed" by "another logarithmically distributed"

The Even dataset being "non realistic" for abundance estimation there is not much interest of having it in table 1, the authors should replace it by the Adp dataset.

About "Reference genome streaming is uniquely advantageous", the authors should briefly describe how this is done, which ID and reference sites are used?

Discussion :

Regarding "Existing methods generally assume uniform read lengths and therefore they only need to count the number of mapped reads or k -mers. However, it is unreasonable to treat, say, a 100 bp long read and a 1000 bp long read the same when calculating their contributions to the genome abundance." This is not evident. If you model your sequencing as random sorting in a population then you should correct for genome length not for read length. If two genome have the same abundance but one has twice the length of the other then the chance to pick a read from the longer genome is twice the one from the shorter. This is the same case as with transcript length in transcriptome data and induces the per kilo base of transcript normalization. But this is not linked with read length. Each read is random draw and the length of the drawn fragment has no impact. Correcting for the genome length is not trivial be cause you rarely know the metagenome composition. The authors should explain their position.