NanoSim: nanopore sequence read simulator based on statistical characterization

Chen Yang, Justin Chu, René L Warren, Inanç Birol

Background:
The MinION sequencing platform from Oxford Nanopore Technologies (ONT) is still a pre-commercial technology, yet it is generating substantial excitement in the field for its features – longer read lengths and single-molecule sequencing in particular. As groups start developing bioinformatics tools for this new platform, a method to model and simulate the properties of the sequencing data will be valuable to test alternative approaches and to establish performance metrics. Here, we introduce NanoSim, a fast and lightweight read simulator that captures the technology-specific characteristics of ONT data with robust statistical models.

Results:
The first step of NanoSim is read characterization, which provides a comprehensive alignment-based analysis, and generates a set of read profiles serving as the input to the next step, the simulation stage. The simulation tool uses the model built in the previous step to produce in silico reads for a given reference genome. NanoSim is built on our observation that patterns of correct base calls and errors (mismatches and indels) can be described by statistical mixture models. Further, the structures of these models are consistent across chemistries and organisms (*E. coli* and *S. cerevisiae*). NanoSim generates synthetic ONT reads with empirical profiles derived from reference datasets, or using runtime parameters. Empirical profiles include read lengths and alignment fractions (the ratio of alignment lengths after unaligned portions of reads are soft-clipped from their flanks to read lengths). The lengths of intervals between errors (stretches of correct bases) and error types are modeled by Markov chains, and the lengths of errors are drawn from mixed statistical models.

Conclusion:
In this work, we demonstrate the performance of NanoSim on publicly available datasets generated using R7 and R7.3 chemistries and different sequencing kits. NanoSim mimics ONT reads well, true to the major features of the emerging ONT sequencing platform, in terms of read length and error modes. The independent profiling module grants users the freedom to characterize their own ONT datasets, which is expected to perform consistently upon the improvement of nanopore sequencing technology, as the shapes of the error models hold among different datasets. NanoSim

will immediately benefit the development of scalable NGS technologies for the long nanopore reads, including genome assembly, mutation detection, and even metagenomic analysis software. The scalability of NanoSim to human-size genome will benefit the development of scalable NGS technologies for long nanopore reads. Moreover, a mixture of in silico genomes simulating a microbiome will be helpful for benchmarking algorithms with applica-tion in metagenomics, including functional gene prediction, species detection, comparative metagenomics, clinical diagnosis. As such, we expect NanoSim to have an enabling role in the field.