Genome analysis

OMSim: a simulator for optical map data

Giles Miclotte¹,², Stéphane Plaisance³, Stephane Rombauts²,⁴,⁵, Yves Van de Peer²,⁴,⁵,⁶, Pieter Audenaert¹,² and Jan Fostier¹,²,*

¹Department of Information Technology, IDLab, Ghent University–IMEC, Ghent 9052, Belgium, ²Bioinformatics Institute Ghent, Ghent University, Ghent 9052, Belgium, ³Nucleomics Core, VIB, Leuven 3000, Belgium, ⁴Center for Plant Systems Biology, VIB, Ghent 9052, Belgium, ⁵Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent 9052, Belgium and ⁶Department of Genetics, Genome Research Institute, University of Pretoria, Pretoria 0028, South Africa

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on February 21, 2017; revised on April 13, 2017; editorial decision on April 27, 2017; accepted on May 2, 2017

Abstract

Motivation: The Bionano Genomics platform allows for the optical detection of short sequence patterns in very long DNA molecules (up to 2.5 Mbp). Molecules with overlapping patterns can be assembled to generate a consensus optical map of the entire genome. In turn, these optical maps can be used to validate or improve de novo genome assembly projects or to detect large-scale structural variation in genomes. Simulated optical map data can assist in the development and benchmarking of tools that operate on those data, such as alignment and assembly software. Additionally, it can help to optimize the experimental setup for a genome of interest. Such a simulator is currently not available.

Results: We have developed a simulator, OMSim, that produces synthetic optical map data that mimics real Bionano Genomics data. These simulated data have been tested for compatibility with the Bionano Genomics Irys software system and the Irys-scaffolding scripts. OMSim is capable of handling very large genomes (over 30 Gbp) with high throughput and low memory requirements.

Availability and implementation: The Python simulation tool and a cross-platform graphical user interface are available as open source software under the GNU GPL v2 license (http://www.bioinformatics.intec.ugent.be/omsim).

Contact: jan.fostier@ugent.be

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

The Bionano Genomics platform is able to visualize occurrences of specific, short sequence motifs (e.g. 7 bp) along very long stretches of linearized DNA molecules (up to 2.5 Mbp), thus forming a unique, sequence-specific pattern per molecule, sometimes referred to as a ‘barcode’. By using those signature patterns, the molecules can be assembled in a complete consensus genome map. This view of the genome can be used to validate or improve de novo genome assembly, by providing a scaffold on which the contigs can be anchored (Shi et al., 2016), or to detect large-scale structural variation in genomes (Mak et al., 2015).
OMSim simulates the Bionano Genomics process using statistical models for which the parameters were derived from real data (see Supplementary Material data S1 for the parameter description), and generates output in BNX format. It is implemented in Python, and relies on the Scipy library to sample from the required distributions. A graphical user interface has been developed to facilitate the setup of the simulation process. OMSim requires a reference assembly as the ground truth for the simulation. Each map is simulated from a single contig, hence the contiguity of the reference assembly limits the lengths of the simulated optical maps, i.e. it is impossible to simulate an optical map that is longer than the contig from which it is simulated.

**2 Methods and results**

OMSim was designed to accurately mimic all sources of variation that occur in the Bionano Genomics data. First, false positive and false negative labels are taken into account, where labels are either erroneously placed or not placed. Second, there is the occurrence of fragile sites, where labels that occur very close to each other cause systematic breaks in the molecules. Third, each molecule has a stretch factor, which quantifies how the migration of the DNA molecule through a nanochannel causes the molecule to stretch or shrink. Fourth, there is some additional variability in the position of the labels due to local stretching. Fifth, due to the limited optical resolution, nearby labels may appear as one label in the image. Finally, also due to the optical resolution, there is the possibility of chimeric maps, which occur when distinct molecules are close together in a nanochannel such that they appear as a single molecule in the image.

The OMSim process consists of two steps. First, the locations of the sequence recognition sites in the genome are indexed using the computationally efficient Knuth-Morris-Pratt algorithm (Knuth et al., 1977). This index can be reused for future runs. Second, using this index, OMSim simulates the actual optical map data. Molecular lengths are generated from a negative binomial distribution and for each molecule a start location is uniformly chosen on the provided reference genome. Then, labels and noise are introduced in each molecule. False positive (resp. negative) labels are uniformly distributed along the molecules (resp. labels). The molecules are broken at fragile sites, based on the proximity of neighbouring labels. Stretch factor variations are normally distributed. Labels that occur close together are collapsed into a single label. After simulating the molecules, chimeras are introduced by concatenating molecules.

Optical map data was simulated from the human genome reference Hg19, and anchored using the Bionano Genomics RefAligner. The resulting alignments were compared to the alignments of real data from NA24385 (Ashkenazim Trio son, public data from http://bionanogenomics.com/science/public-datasets/). A portion of these alignments and the coverage and the size distribution of both simulated maps and real maps are shown in **Figure 1**. This figure shows that the simulated data can be aligned to the reference, that similarly as in real data missing labels are present due to false positives or collapsing labels, and that the size distributions of the simulated and real data are nearly identical. A peak memory usage of 478 MB was measured while indexing the human reference Hg19 and simulating...
optical map data from this index. Peak memory usage depends on the number of nicking sites in the reference. The indexing run time is linear in the size of the reference, while the simulation run time is linear in the size of the output. In our tests for genomes with sizes ranging from 4 Mbp up to 30 Gbp, this corresponds to a throughput of 30 Mbp per minute for indexing and 12.5 Gbp per minute for the actual simulation. Loading the index in subsequent runs took less than 30 seconds for all data sets. From these results we conclude that OMSim efficiently simulates data that resemble the real Bionano Genomics data.

**Funding**

This work was supported by The Research Foundation–Flanders (FWO) [G0C3914N].

**Conflict of Interest:** none declared.

**References**

Knuth,D. et al. (1977) Fast pattern matching in strings. *SIAM J. Comput.*, 6, 323–350.

Leung,A.K.-Y. et al. (2017) Omblast: alignment tool for optical mapping using a seed-and-extend approach. *Bioinformatics*, 33, 311.

Li,M. et al. (2016) Towards a More Accurate Error Model for BioNano Optical Maps. *Springer International Publishing*, Cham, pp. 67–79.

Mak,A.C.Y. et al. (2015) Genome-wide structural variation detection by genome mapping on nanochannel arrays. *Genetics*, 202, 351–362.

Muggli,M.D. et al. (2014) Efficient Indexed Alignment of Contigs to Optical Maps. *Springer Berlin Heidelberg*, Berlin, Heidelberg, pp. 68–81.

Muggli,M.D. et al. (2015) Misassembly detection using paired-end sequence reads and optical mapping data. *Bioinformatics*, 31, i80.

Shelton,J.M. et al. (2015) Tools and pipelines for bionano data: molecule assembly pipeline and fasta super scaffolding tool. *BMC Genom.*, 16, 1–16.

Shi,L. et al. (2016) Long-read sequencing and de novo assembly of a Chinese genome. *Nat. Commun.*, 7, 12065.