Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Applications Note

Genome analysis

Improving Bionano scaffolding with BiSCoT

Benjamin Istace[1,*], Caroline Belser[1] and Jean-Marc Aury[1,*]

[1]Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, Evry, France

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Long reads sequencing and Bionano Genomics optical maps are two techniques that, when used together, make it possible to reconstruct entire chromosome or chromosome arms structure. However, the existing tools are often too conservative when two contigs are joined.
Results: We developed BiSCoT (Bionano SCaffolding COrrection Tool), a software that uses informations produced by a pre-existing assembly based on optical maps as input and improves the contiguity and the quality of the generated assembly.
Availability: BiSCoT is freely available at https://github.com/institut-de-genomique/biscot
Contact: bistace@genoscope.cns.fr;jmaury@genoscope.cns.fr

1 Introduction

Assembling large and repetitive genomes, such as plant genomes, is a challenging field in bioinformatics. The apparition of short reads technologies several years ago improved considerably the number of genomes publicly available. However, a high proportion of them are still fragmented and few represent the chromosome organization of the genome. Recently, long reads sequencing techniques, like Oxford Nanopore Technologies and Pacific Biosciences, were introduced to improve the contiguity of assemblies, by sequencing DNA molecules that can range from a few kilobases to more than a megabase in size. Nevertheless and even if the assemblies were greatly improved, the chromosome-level organization of the sequenced genome couldn't be deciphered in a majority of cases. In 2017, Bionano Genomics launched its Saphyr system which was able to generate optical maps of a genome, by using the distribution of endonuclease nicking sites. These maps were used to orient and order contigs into scaffolds but the real improvement came in 2018, when Bionano Genomics introduced their Direct Label and Stain (DLS) technology that was able to produce genome maps with an N50 several folds higher than previously (Deschamps et al, Belser et al, 2018).

However, scaffolds generated with the tool provided by Bionano Genomics do not reach optimal contiguity. Indeed, when two contigs C_1 and C_2 are found to share labels, one could expect that the tool would merge the two sequences at the shared site. Instead, the software chooses a conservative approach and outputs the sequence of C_1 followed by a 13-Ns gap and then the C_2 sequence, thus duplicating the region that is shared by the two contigs (Supplementary Figure 1) and in numerous cases, these duplicated region could reach several kilobases. As an example, on the human genome we used to evaluate BiSCoT (see Results), we could detect 168 of those regions, affecting 16 genes and corresponding to around 5Mb of duplicated sequences, the longest being 94kb in size. These duplicated regions have to be corrected as they can be problematic for downstream analyses, like copy number variation studies. In addition, contig maps can sometimes be inserted into other maps, these cases are not handled by the Bionano scaffolding tool that discards the inserted maps (Supplementary Figure 2).

We developed BiSCoT, a python script that examines data generated during a previous Bionano scaffolding and merges contigs separated by a 13-Ns gap if needed. BiSCoT also re-evaluates gap sizes and searches for an alignment between two contigs if the gap size is inferior to 100 nucleotides.

2 Methods

Briefly, BiSCoT loads all necessary files into memory, then checks if two adjacent contigs in an alignment share labels. If so, it merges those two contigs at the shared site. If no labels are shared, a gap size g is estimated. If $g \leq 100$, then a BLAT (Kent, 2002) alignment is performed between contig borders and if a match is found, contigs are merged at the corresponding alignment position. Finally, BiSCoT generates an AGP

Table 1. Metrics of the NA12878 scaffolds and contigs before or after BiSCoT treatment

	Nanopore contigs	Bionano		BiSCoT	
		Contigs	Scaffolds	Contigs	Scaffolds
Cumulative size	2,818,916,510	2,818,960,013	2,886,767,251	2,807,503,705	2,875,309,225
N50 (L50)	11,821,944 (67)	10,566,783 (71)	88,757,536 (14)	12,944,470 (62)	88,534,944 (14)
N90 (L90)	2,143,851 (280)	1,863,173 (301)	29,130,580 (36)	2,360,424 (247)	29,031,104 (36)
Number of Ns (%)	0 (0%)	0 (0%)	67,807,238 (2.35%)	0 (0%)	67,805,520 (2.36%)
NGA50	5,794,944	5,729,014	10,851,955	6,298,936	11,467,472
NGA75	1,511,206	1,495,174	2,688,891	1,573,857	2,991,730
Number of misassemblies	1,358	1,303	1,592	1,303	1,491
Complete BUSCOs	255 (84.2%)	254 (83.9%)	255 (84.2%)	255 (84.2%)	256 (84.5%)
Missing BUSCOs	29 (9.6%)	32 (10.5%)	29 (9.6%)	29 (9.6%)	29 (9.6%)

file describing changes in the contig organization and a FASTA file corresponding to scaffolds.

2.1 Mandatory files loading

During the scaffolding, the Bionano scaffolder generates one key file, several CMAP files and a XMAP file that will be used by BiSCoT. It also generates a visual representation of the hybrid scaffolds that is called an 'anchor'.

BiSCoT first loads the contigs into memory based on the key file, which describes the mapping between map identifiers and contig names.

Then, BiSCoT loads the anchor CMAP file and one CMAP contig file per enzyme into memory. These files contain the position of enzymatic labelling sites on contig maps and on the anchor generated during the Bionano scaffolding.

Finally, BiSCoT loads the XMAP file that describes the alignment between a contig map and an anchor.

2.2 Scaffolding

Alignments contained into the XMAP file are first sorted by their starting position on the anchor. Then, alignments on one anchor are parsed by pairs of adjacent contigs, i.e alignment of contig C_k is examined at the same time as contig C_n. Aligned anchor labels are extracted from these alignments and a list of shared labels $L_{n,k}$ is built. For the following cases, we suppose C_k and C_n to be aligned on the forward strand and C_k to be aligned before C_n on the anchor (Supplementary Figure 1 and 2).

2.2.1 Case 1: contig maps share anchor labels

First, the right-most shared label identifier l is extracted. The position of l (P_l) on both contigs C_k ($P_l(C_k)$) and C_n ($P_l(C_n)$) is searched, based on information provided in the CMAP files. In the resulting scaffold, the sequence of C_k will be included up to the $P_l(C_k)$ position and the sequence of C_n will be included from the $P_l(C_n)$ position.

2.2.2 Case 2: contig maps don't share anchor labels

Let $Size_k$ be the size of the contig C_k, Sm_k and Em_k the start and end of an alignment on a contig map and Sa_k and Ea_k the corresponding coordinates on the anchor. The number n of bases between the right-most aligned label of C_k and the left-most aligned label of C_n is then:

$$n = Sa_n - Ea_k$$

We then have to subtract the part d_k of C_k after the right-most aligned label and the part d_n of C_n before the left-most aligned label:

$$d_k = Size_k - Em_k$$

$$d_n = Sm_n$$

Finally, we can compute the gap size g with:

$$g = n - d_k - d_n$$

If $g < 100$, a BLAT alignment of the last 30kb of C_k is launched against the first 30kb of C_n. Only the best alignment is kept and if its score is higher than 5,000, C_k and C_n are merged at the starting position of the alignment. Otherwise, a number g of Ns is inserted between C_k and C_n.

3 Results

In order to evaluate BiSCoT, we downloaded the contigs fasta file of the NA12878 human genome that was sequenced and assembled by the Whole Human Genome Sequencing Project (Jain et al, 2018) and Bionano DLE and BspQI maps from the Bionano Genomics website (https://bionanogenomics.com/library/datasets/). The QUAST (Gurevich et al, 2013) and BUSCO (Simão et al, 2015) tools were used respectively to evaluate the number of misassemblies to the GRCh38.p12 human reference genome and the number of conserved genes among eukaryotes.

Furthermore, we downloaded the contigs fasta files and Bionano maps of three already published plant genomes: *Brassica oleracea* HDEM (PRJEB26621), *Brassica rapa* Z1 (PRJEB26620) and *Musa schizocarpa* (PRJEB26661).

In all cases, we first ran a classic Bionano scaffolding and used the files generated during this step to run BiSCoT (Table 1, Supplementary Table 1, 2 and 3).

Globally, using BiSCoT lowered slightly the scaffolds NX metrics but this is an expected side effect as the merging of sequences induces a lower cumulative size.

However, the contig NX and NGAX metrics increased drastically: the contigs NGA50 of NA12878 increased by around 10%, going from 5.8Mb to 6.3Mb. The scaffolds NGAX metrics also increased: the scaffolds NGA50 increased from 10.8Mb in Bionano scaffolds to 11.5Mb in BiSCoT scaffolds. Moreover, the number of Ns decreased marginally and the number of complete eukaryotic BUSCO genes increased from 255 to 256. More importantly, when aligning the assemblies against the reference genome, we could detect a decrease in the number of misassemblies going from 1,592 in Bionano scaffolds to 1,491 in BiSCoT scaffolds.

The same kind of results were observed in the three plant genomes with a slight decrease in scaffolds NX metrics and number of Ns but an increase in contigs NX metrics and complete BUSCO genes (Supplementary Tables 1, 2 and 3).

Acknowledgements

The authors are grateful to the Bionano Genomics staff for technical help and would also like to thank the Whole Human Genome Sequencing Project for providing access to the Nanopore human genome assembly.

Funding

This work was supported by the Genoscope, the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) and France Génomique (ANR-10-INBS-09-08).

References

Belser,C. et al. (2018) Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. Nat Plants, 4, 879–887.

Deschamps,S. et al. (2018) A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. Nat. Commun., 9, 4844.

Gurevich,A. et al. (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29, 1072–1075.

Jain,M. et al. (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol., 36, 338–345.

Kent,W.J. (2002) BLAT--the BLAST-like alignment tool. Genome Res., 12, 656–664.

Simão,F.A. et al. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31, 3210–3212.