BiSCoT: Improving large eukaryotic genome assemblies with optical maps

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**

Corresponding author:
Benjamin Istace[1]

Email address: bistace@genoscope.cns.fr

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 and organization of contigs into scaffolds is not always optimal.
Results. We developed BiSCoT (Bionano SCaffolding COrrection Tool), a tool 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. BiSCoT was tested on a human genome and three plant genomes sequenced with Nanopore long reads and improved significantly the contiguity and quality of the assemblies. BiSCoT generates a fasta file of the assembly as well as an AGP file which describes the new organization of the input assembly.
Availability. BiSCoT and improved assemblies are freely available on Github at http://www.genoscope.cns.fr/biscot

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 (Schmidt et al. (2017); Kim et al. (2019); Shafin et al. (2019)). Nevertheless and even if the assemblies were greatly improved, the chromosome-level organization of the sequenced genome cannot 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 a N50 several folds higher than previously (Belser et al. (2018); Formenti et al. (2018); Hu et al. (2019)).

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 (Figure 1 case 1 and 2) and in numerous cases, these duplicated regions could reach several kilobases. As an example, on the human genome we used to evaluate BiSCoT (see Results), we could detect 515 of those regions, affecting 16 genes and corresponding to around 24.5Mb of duplicated sequences, the longest being 237kb in size. These duplicated regions affect the contiguity and have to be corrected as they can be problematic for downstream analyses, like copy number

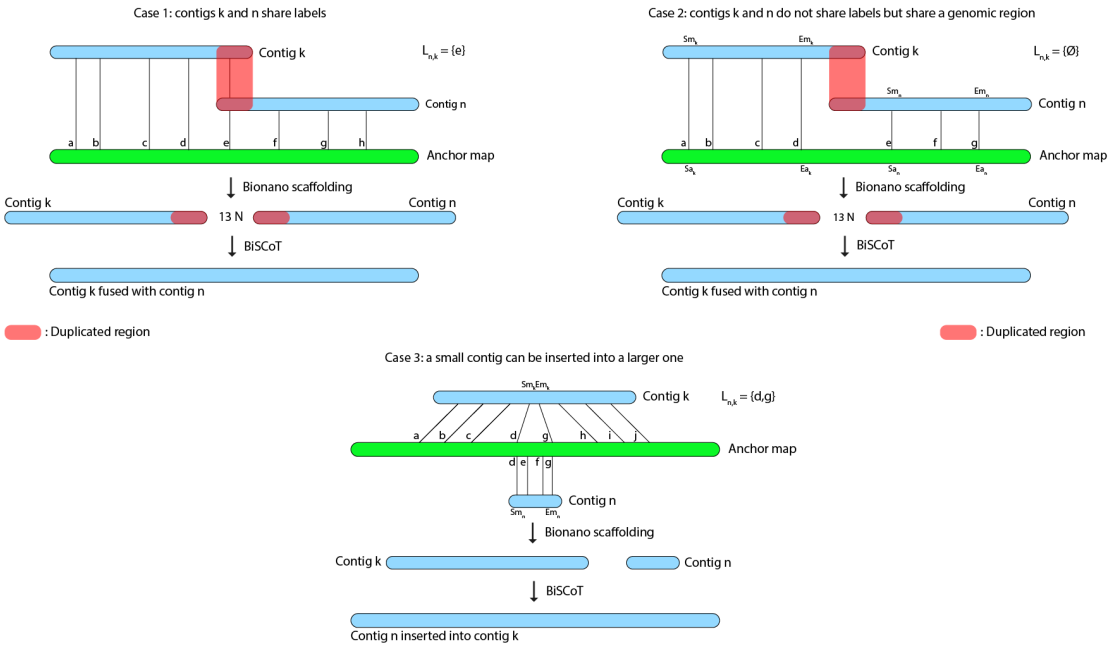

Figure 1. The Bionano scaffolding tool doesn't merge contigs even if they share labels. Instead, it inserts 13 N's gap between contigs, thus artificially duplicating the shared region. **Top left panel.** BiSCoT merges contigs that share enzymatic labelling sites. **Top right panel.** If contigs don't share labels but share a genomic region, BiSCoT attempts to merge them by aligning the borders of the contigs. **Bottom panel.** The Bionano scaffolding tool doesn't handle cases where contigs can be inserted into others. BiSCoT attempts to merge the inserted map with the one containing it if they share labels.

variation studies. They originate from overlaps that are not fused in the input assembly and usually correspond to allelic duplications. In addition, contigs can sometimes be inserted into other contigs, these cases are not handled by the Bionano scaffolding tool that discards the inserted contigs (Figure 1 case 3).

We developed BiSCoT, a python script that examinates 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 1,000 nucleotides.

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 $g \leq 1000$, then a BLAT (Kent, 2002) alignment is performed between contig borders and if a significant match is found, contigs are merged at the corresponding alignment position. Finally, BiSCoT generates an AGP file describing changes in the contig organization and a FASTA file corresponding to scaffolds.

Mandatory files loading

During the scaffolding, the Bionano scaffolder generates a visual representation of the hybrid scaffolds that is called an 'anchor'. It also generates one '.key' file, which describes the mapping between map identifiers and contig names, several CMAP files, which contain the position of enzymatic labelling sites on contig maps and on the anchor, and a XMAP file, that describes the alignment between a contig map and an anchor.
BiSCoT first loads the contigs into memory based on the key file. Then, the anchor CMAP file and contig CMAP files are loaded into memory. Finally, the XMAP file is parsed and loaded.

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, with C_k aligned before C_n on the anchor. 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 (Figure 1).

Case 1: contig maps share at least one anchor label

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. In this case, the gap is removed and both contigs C_k and C_n are fused.

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:

$$n = Sa_n - Ea_k \tag{1}$$

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:

$$d_k = Size_k - Em_k \tag{2}$$

$$d_n = Sm_n \tag{3}$$

Finally, we can compute the gap size g with:

$$g = n - d_k - d_n \tag{4}$$

If $g \leq 1000$, a BLAT(Kent (2002)) 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.

Case 3: insertion of small contigs

Let Sm_k and Em_k the start and end of an alignment on a contig map. If $[Sm_n, Em_n] \subset [Sm_k, Em_k]$, then the left-most shared label identifier l_l and right-most shared label identifier l_r are extracted. If C_n has more of its labels mapped in this region than C_k, the sequence of C_n will be inserted between l_l and l_r in the scaffolds. Otherwise, the sequence of C_k remains unchanged and C_n will be included as a singleton sequence in the scaffolds file.

Finally, if an Illumina polishing step was done before or after Bionano scaffolding, we recommend doing one additional round of polishing using Illumina reads after BiSCoT has been applied. Indeed, short reads tend to be aligned only against one copy of the duplicated regions, leaving the other copy unpolished.

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. 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 used the Bionano workflow to scaffold the draft assembly and launched BiSCoT using the files generated by the Bionano tools (Table 1, Supplementary Table 1, 2 and 3).

	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)
# 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
# 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%)
Duplicated BUSCOs	**9 (3.0%)**	**9 (3.0%)**	9 (3.0%)	**9 (3.0%)**	9 (3.0%)
Missing BUSCOs	**29 (9.6%)**	32 (10.5%)	29 (9.6%)	**29 (9.6%)**	29 (9.6%)

Table 1. Metrics of the NA12878 scaffolds and contigs before or after BiSCoT treatment. Bold formatting indicates the best scoring assembly among contigs.

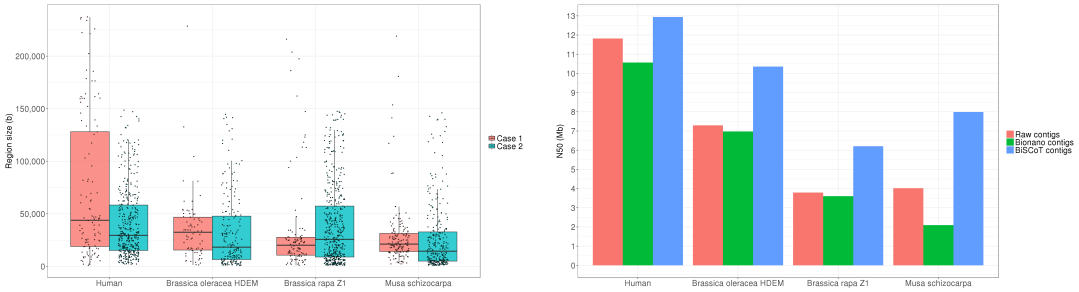

Figure 2. Left panel. Distribution of the sizes of overlapping regions in the raw assemblies. Detection was done using either Bionano labels (Case 1) or a BLAT alignment (Case 2). **Right panel.** N50 contigs of raw assemblies and assemblies before or after BiSCoT treatment.

Concerning the NA12878 genome, we could detect 515 overlapping regions with a mean size of 47kb and representing in total 24.5Mb of duplicated sequences. Among these 515 regions, 499 were corrected by BiSCoT using either shared labels (113 regions) or a BLAT alignment (386 regions) when no shared labels were found.

Globally, 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 mis-assemblies 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 (Figure 2 and Supplementary Tables 1, 2 and 3).

ACKNOWLEDGMENTS

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.

REFERENCES

Belser, C., Istace, B., Denis, E., Dubarry, M., Baurens, F.-C., Falentin, C., Genete, M., Berrabah, W., Chèvre, A.-M., Delourme, R., Deniot, G., Denoeud, F., Duffé, P., Engelen, S., Lemainque, A., Manzanares-Dauleux, M., Martin, G., Morice, J., Noel, B., Vekemans, X., D'Hont, A., Rousseau-Gueutin, M., Barbe, V., Cruaud, C., Wincker, P., and Aury, J.-M. (2018). Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature Plants*, 4(11):879–887.

Formenti, G., Chiara, M., Poveda, L., Francoijs, K.-J., Bonisoli-Alquati, A., Canova, L., Gianfranceschi, L., Horner, D. S., and Saino, N. (2018). SMRT long reads and Direct Label and Stain optical maps allow the generation of a high-quality genome assembly for the European barn swallow (Hirundo rustica rustica). *GigaScience*, 8(1). giy142.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075.

Hu, L., Xu, Z., Wang, M., Fan, R., Yuan, D., Wu, B., Wu, H., Qin, X., Yan, L., Tan, L., Sim, S., Li, W., Saski, C. A., Daniell, H., Wendel, J. F., Lindsey, K., Zhang, X., Hao, C., and Jin, S. (2019). The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. *Nature Communications*, 10(1):4702.

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., Snutch, T. P., Tee, L., Paten, B., Phillippy, A. M., Simpson, J. T., Loman, N. J., and Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36:338 EP –.

Kent, W. (2002). Blat–the blast-like alignment tool. *Genome Research*, 12:656–664.

Kim, H.-S., Jeon, S., Kim, C., Kim, Y. K., Cho, Y. S., Kim, J., Blazyte, A., Manica, A., Lee, S., and Bhak, J. (2019). Chromosome-scale assembly comparison of the Korean Reference Genome KOREF from PromethION and PacBio with Hi-C mapping information. *GigaScience*, 8(12). giz125.

Schmidt, M. H.-W., Vogel, A., Denton, A. K., Istace, B., Wormit, A., van de Geest, H., Bolger, M. E., Alseekh, S., Maß, J., Pfaff, C., Schurr, U., Chetelat, R., Maumus, F., Aury, J.-M., Koren, S., Fernie, A. R., Zamir, D., Bolger, A. M., and Usadel, B. (2017). De novo assembly of a new solanum pennellii accession using nanopore sequencing. *The Plant Cell*, 29(10):2336–2348.

Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., Sedlazeck, F. J., Marschall, T., Mayes, S., Costa, V., Zook, J. M., Liu, K. J., Kilburn, D., Sorensen, M., Munson, K. M., Vollger, M. R., Eichler, E. E., Salama, S., Haussler, D., Green, R. E., Akeson, M., Phillippy, A., Miga, K. H., Carnevali, P., Jain, M., and Paten, B. (2019). Efficient de novo assembly of eleven human genomes using promethion sequencing and a novel nanopore toolkit. *bioRxiv*.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212.