Title

Accurate detection of chimeric contigs via Bionano optical maps.

Journal

Bioinformatics, 35(10)

ISSN

1367-4803

Authors

Pan, Weihua
Lonardi, Stefano

Publication Date

2019-05-15

DOI

10.1093/bioinformatics/bty850

Peer reviewed

Bioinformatics, 35(10), 2019, 1760–1762
doi: 10.1093/bioinformatics/bty850
Advance Access Publication Date: 8 October 2018
Applications Note

Genome analysis

Accurate detection of chimeric contigs via Bionano optical maps

Weihua Pan and Stefano Lonardi*

Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on March 30, 2018; revised on September 5, 2018; editorial decision on September 27, 2018; accepted on October 4, 2018

Abstract

Summary: A *chimeric contig* is contig that has been incorrectly assembled, i.e. a contig that contains one or more mis-joins. The detection of chimeric contigs can be carried out either by aligning assembled contigs to genome-wide maps (e.g. genetic, physical or optical maps) or by mapping sequenced reads to the assembled contigs. Here, we introduce a software tool called Chimericognizer that takes advantage of one or more Bionano Genomics optical maps to accurately detect and correct chimeric contigs. Experimental results show that Chimericognizer is very accurate, and significantly better than the chimeric detection method offered by the Bionano Hybrid Scaffold pipeline. Chimericognizer can also detect and correct chimeric optical molecules.
Availability and implementation: https://github.com/ucrbioinfo/Chimericognizer
Contact: stelo@cs.ucr.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

De novo genome assembly is a fundamental problem in genomics and computational biology. The objective of *de novo* genome assembly is to generate the longest possible set of contigs with the smallest number of errors (mis-joins) from sequenced reads. Despite significant algorithmic progress, this problem remains challenging due to the high repetitive content of eukaryotic genomes, insufficient read length, uneven sequencing coverage, non-uniform sequencing errors and chimeric reads. Irrespective on the type of sequencing technology or the algorithmic strategies employed, mis-joins are hard to avoid. A *chimeric contig* is contig that has been incorrectly assembled from reads originating from non-adjacent regions of the genome. Failing to recognize and correct chimeric contigs can have dramatic consequences in downstream steps in the assembly pipeline, e.g. scaffolding or construction of pseudo-molecules.

2 Materials and methods

Here, we introduce Chimericognizer, a tool that can detect large-scale mis-joins in either assembled contigs or Bionano optical molecules. The presence of mis-joins induces conflicts in high-quality alignments between contigs and optical molecules (Jiao *et al.*, 2017) (Supplementary Fig. S2). The quality of an alignment depends on the consistency of shared distances between adjacent restriction enzyme sites and the total length of the alignment. Due to the requirement for high-quality alignments, Chimericognizer can detect mis-joins only on assembled contigs that are sufficiently long to be reliably aligned, e.g. 50 Kbp or longer. Contigs produced from the assembly of third-generation sequencing data (e.g. PacBio and Oxford Nanopore) generally meet this criterion. In this case, the detection of chimeric contigs appears straightforward if one assumes that optical maps are error-free and all the alignment conflicts are caused by mis-joins in the contigs. Unfortunately, since optical maps are obtained via an assembly process similar to sequence assembly, optical molecules can also be chimeric. According to (Jiao *et al.*, 2017), in about '7% *of the (alignment) conflicts, the consensus map (optical map) was wrong*'. Mis-joins in optical molecules typically occur in repetitive regions of the genome, which induce long stretches of regularly-spaced restriction enzyme sites

Chimericognizer depends on the availability of multiple assemblies and one (or more) Bionano optical map to accurately detect

chimeric contigs and reduce the possibility of incorrectly splitting non-chimeric contigs. Multiple assemblies can be obtained by either running several assembly tools or by using one assembler with multiple parameter settings on the same input data.

The algorithm used by Chimericognizer has three phases. In the first phase, a list of candidate chimeric sites for either assembled contigs or optical molecules is produced. The first phase has three steps. In Step 1, we concatenate all the available genome assemblies and *in silico*-digest them using the same restriction enzyme(s) used to produce the Bionano optical map(s). Then we align digested contigs to their corresponding optical map using Bionano Genomics RefAligner. In Step 2, we remove low-confidence and redundant alignments. When multiple optical maps are available, we unify the coordinates for all alignments (Step 3).

In the second phase, we select high-confidence chimeric sites from the list of candidate sites. We first compute the relevance of each candidate site (Supplementary Note S1 for the definition of *relevance*), then we find the subset with minimum total relevance which can resolve all the conflicts. In the third phase, chimeric contigs and molecules are cut at high-confidence chimeric sites. Additional details can be found in Supplementary Note S1. The algorithm pipeline is illustrated in Supplementary Figure S1.

3 Experimental results and discussion

To assess the performance of Chimericognizer, we used real and synthetic datasets for cowpea (*Vigna unguiculata*) along with two Bionano Genomics optical maps. We also tested Chimericognizer on a fruit fly (*Drosophila melanogaster*) dataset (Solares *et al.*, 2018), for which a high-quality reference genome is available. To the best of our knowledge, the Bionano Hybrid Scaffold pipeline is the only available tool that solves exactly the same problem addressed by Chimericognizer. Other chimeric detection methods are available, but they either require additional data or focus on different types of mis-joins. For example, Missequel can detect mis-joins that are much shorter than our tool, but it requires short reads in addition to an optical map (Muggli *et al.*, 2015).

For cowpea, we used Canu (Koren *et al.*, 2017), ABruijn (Lin *et al.*, 2016) and Falcon (Chin *et al.*, 2016) to generate eight assemblies from ≈6 M PacBio reads (Supplementary Note S2). Supplementary Table S2 shows the assembly statistics after the removal of chimeric contigs via Chimericognizer compared to the manually-curated assemblies (carried out by an expert several months before we developed Chimericognizer). The manual curation involves detecting chimeric contigs by visually inspecting the alignments using Bionano IrisView. For a genome of the size of cowpea, it takes about 3 h for each assembly. The process is tedious and error-prone.

First, observe in Supplementary Table S2 that there is almost no difference between Chimericognizer's statistics using one versus two optical maps. We believe that the second optical map does not help in this case because the number of input assemblies is sufficiently high (experiments below seem to support this hypothesis). Second, note that the N50 is higher for Chimericognizer's assemblies compared to the manually-curated assemblies, indicating that the expert was overly aggressive in splitting contigs. Since there is no 'ground truth' on this dataset (i.e. no high-quality reference genome), we evaluated these results using other independent metrics. First, we mapped ≈200 M paired-end Illumina reads using BWA. A comparative lower percentage of mapped reads (particularly properly-paired) would indicate an assembly that still contains chimeric contigs. Supplementary Table S2 shows there is almost no difference between Chimericognizer's and the expert's assemblies in terms of mapped reads. Second, we compared the assemblies against the high-density genetic map available from (Muñoz-Amatriaín *et al.*, 2017). To evaluate whether the assemblies contained residual chimeric contigs, we BLASTed the 121 bp-long sequence surrounding the 51 128 SNPs provided in (Muñoz-Amatriaín *et al.*, 2017) against each assembly, then we identified which contigs had SNPs mapped to them, and what linkage groups (chromosomes) of the genetic map those mapped SNPs belonged to. Chimeric contigs are revealed when their mapped SNPs belong to more than one linkage group. The last row of each panel in Supplementary Table S2 reports the total size of contigs in each assembly for which (i) they contain at least one SNPs and (ii) all mapped SNPs belong to the same linkage group (i.e. likely to be non-chimeric). Observe in Supplementary Table S2 that Chimericognizer's assemblies have higher agreement with the genetic map than the expert's assemblies. Finally, Chimericognizer determined that the expert missed 23/28 chimeric contigs in the eight assemblies using BspQI/BssSI, respectively and 40 chimeric contigs when using both maps (some examples are shown in Supplementary Fig. S4). In all these cases, he later agreed that all these chimeric contigs should be have been split.

To generate a dataset containing synthetic chimeric contigs, we started from the eight cowpea assemblies described above and used Chimericognizer to clean them from chimeric contigs. In each of the eight chimeric-free assemblies, we injected chimeric contigs by pairwise joining 2% of the contigs longer than 500 Kbp (selected at random). Then we used Chimericognizer and Bionano Hybrid Scaffold to detect these synthetic chimeric contigs. We measured precision and sensitivity as described in Supplementary Note S2 and Supplementary Figure S3. Experimental results for Chimericognizer are reported in Supplementary Table S3 and S4, while the results for Bionano Hybrid Scaffold are summarized in Supplementary Table S5. These are average values over ten synthetic datasets generated as described above. First, observe that Bionano Hybrid Scaffold missed all the chimeric contigs. In the case of Chimericognizer, using two optical maps the precision is very close to 100% while the sensitivity is always higher than 94%. The precision with one optical map is as good as two optical maps, but the sensitivity is worse (around 80%). We also generated a synthetic dataset in which we injected chimeric molecules in the optical map (see Supplementary Note S2 for details). Supplementary Table S6 shows that the Chimericognizer's precision is 100% and the sensitivity varies between 77% and 93%.

As said, the accuracy of Chimericognizer depends on the availability in multiple assemblies. To study Chimericognizer's performance as a function of the number of available assemblies, we randomly selected a subset of the assemblies then generated datasets containing synthetic chimeric contigs as described above. Supplementary Tables S7 and S8 report average values over ten synthetic datasets for each choice of the subset size. With one optical map and one assembly, Chimericognizer recognizes chimeric contigs and sites with relatively low precision (about 68%). The precision improves significantly (97–99%) when either two optical maps or two assemblies are used. Note that the precision increases with the number of assemblies, while the sensitivity increases with the number of optical maps. Also observe that having more than one assembly is critical when Chimericognizer can only rely on one optical map.

The fruit fly dataset contained three assemblies and one Bionano Genomics optical map. Two of the assemblies were generated by Canu and MiniMap + MiniAsm from Oxford Nanopore reads. The third assembly was obtained using Platanus + DBG2OLC on Illumina and Oxford Nanopore reads (Supplementary Note S3). Using the high-quality reference genome available for the fruit fly, we identified six true chimeric contigs in the three assemblies (Supplementary Note S3). Chimericognizer correctly identified five of them and did not report any false positives (Supplementary Table S9). Bionano Hybrid Scaffold detected five chimeric contigs, but none of them was correct (Supplementary Table S10).

As said, due to the limited resolution of optical maps Chimericognizer can detect mis-joins on assembled contigs only when they are sufficiently long to be reliably aligned. Smaller mis-joins or leftover overhangs could be removed by mapping the original long read to the contigs. Another possible complication could derive from processing highly heterozygous or polyploid genomes. Additional testing is needed to determine whether Chimericognizer would be accurate in detecting chimeric contigs in these cases.

Acknowledgements

The authors want to thank Prof. J.J. Emerson for providing the optical map of drosophila.

Funding

This work was supported in part by the US National Science Foundation [IOS-1543963, IIS-1526742, IIS-1814359]

Conflict of Interest: none declared.

References

Chin,C.-S. *et al*. (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods*, **13**, 1050–1054.

Jiao,W.-B. *et al*. (2017) Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res.*, **27**, 778–786.

Koren,S. *et al*. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, **27**, 722–736.

Lin,Y. *et al*. (2016) Assembly of long error-prone reads using de Bruijn graphs. *Proc. Natl. Acad. Sci. USA*, **113**, E8396–E8405.

Muggli,M.D. *et al*. (2015) Misassembly detection using paired-end sequence reads and optical mapping data. *Bioinformatics*, **31**, i80–i88.

Muñoz-Amatriaín,M. *et al*. (2017) Genome resources for climate-resilient cowpea, an essential crop for food security. *Plant J.*, **89**, 1042–1054.

Solares,E.A. *et al*. (2018) Rapid low-cost assembly of the *Drosophila melanogaster* reference genome using low-coverage, long-read sequencing. *G3: Genes, Genomes, Genetics*, **8**, 3143–3154.