Sequence analysis

Accurate Detection of Chimeric Contigs via Bionano Optical Maps

Weihua Pan ¹ and Stefano Lonardi ¹,

¹Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: A chimeric contig is a 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: https://github.com/ucrbioinfo/Chimericognizer

Contact: stelo@cs.ucr.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

© The Author 2018. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com
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 RelAligner. 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 (see Supplemental Note 1 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 Supplemental Note 1. The algorithm pipeline is illustrated in Supplemental Figure 1.

3 Experimental results and discussion

To test 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 ≥6M PacBio reads (see Supplemental Note 2). Supplemental Table 2 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 three hours for each assembly. The process is tedious and error-prone.

First, observe in Supplemental Table 2 that there is almost no difference between Chimericognizer’s statistics using one vs. 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 ≥200M 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. Supplemental Table 2 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-Amatriain et al., 2016. To evaluate whether the assemblies contained residual chimeric contigs, we BLASTed the 121bp-long sequence surrounding the 51,128 SNPs provided in (Muñoz-Amatriain et al., 2016) 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 Supplemental Table 2 reports the total size of contigs in each assembly for which i) they contain at least one SNP and ii) all mapped SNPs belong to the same linkage group (i.e. likely to be non-chimeric). Observe in Supplemental Table 2 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/BosSI, respectively, and 40 chimeric contigs when using both maps (some examples are shown in Supplemental Figure 4). In all these cases, he later agreed that all these chimeric contigs should 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 Supplemental Note 2 and Supplemental Figure 3. Experimental results for Chimericognizer are reported in Supplemental Table 3 and 4, while the results for Bionano Hybrid Scaffold are summarized in Supplemental Table 5. 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 Supplemental Note 2 for details). Supplemental Table 6 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. Supplemental Table 7 and 8 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 Minimap2/Miniasm from Oxford Nanopore reads. The third assembly was obtained using Platimus+DBG2OLC on Illumina and Oxford Nanopore reads (see Supplemental Note 3). Using the high-quality reference genome available for the fruit fly, we identified six true chimeric contigs in the three assemblies (see Supplemental Note 3). Chimericognizer correctly identified five of them and did not report any false positives (see Supplemental Table 9). Bionano Hybrid Scaffold detected five chimeric contigs, but none of them was correct (see Supplemental Table 10).

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 overlaps 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


