Accurate Detection of Chimeric Contigs via Bionano Optical Maps
(Supplemental Material)

Weihua Pan and Stefano Lonardi

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

Supplemental Note 1: Methods

The algorithm used by CHIMERICOGNIZER has three phases. 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 alignments either i) when they have a confidence lower than a minimum threshold or ii) when there is another alignment between the same contig and the same molecule with higher confidence. In step 3, we unify the coordinates of alignments when multiple optical maps are available. Due to imprecisions in optical mapping, the distances between restriction enzyme sites in optical maps can be inflated. To compensate for the inflation, REFALIGNER has to amplify the distances of restriction enzyme sites on the contigs by a scaling factor so that accurate alignments can be produced. Since this scaling factor is different for each optical map, in order to make the coordinates comparable across maps, we have to normalize them by the appropriate scaling factor.

After pre-processing, we identify possible conflicts between contigs and molecules. For each alignment a between an optical molecule o and a contig c, we compute the left overhang l_o and right overhang r_o from o and the left overhang l_c and right overhang r_c from c. The left-end of alignment a is declared a *conflict site* if i) both l_o and l_c are longer than some minimum length (default 50 kbp) and ii) at least one restriction enzyme sites appear in both l_o and l_c. A symmetric argument applies to the right-end of the alignment (which determines the values for r_o and r_c). The example in Supplemental Figure 2A illustrates a conflicting alignment between an optical molecules (green) and an assembled contigs (blue). Observe that a) l_o is approximately 0.37 Mb and l_c is approximately 0.27 Mb and b) the green overhang and the blue overhang contain several restriction sites. Since conditions i) and ii) are satisfied, this is an alignment conflict. Once a conflict site is recognized, the location on the optical molecule and the contig are stored as a pair of *candidate chimeric sites* (red arrows in Supplemental Figure 2A). Supplemental Figure 2B illustrates a likely chimeric optical molecule, where again the candidate locations for splits are indicated by the red arrows (here l_o is the optical molecule left overhang, l_c is the contig left overhang, r_o is the optical molecule right overhang, and r_c is the contig right overhang). Observe that the 1.5 Mb-long region between the two red arrows contains regularly-spaced restriction enzyme sites, indicating a repetitive region of the genome. It is likely the the Bionano Assembler created a mis-join in the optical map in that region.

In the second phase, high-confidence chimeric sites are selected from the list of candidate sites. The *relevance* of each candidate site is first quantified, then a maximum parsimony strategy is applied. Among all the candidate sites, we find the subset with minimum total relevance which can resolve all the conflicts. We model this problem as a weighted vertex cover problem on a *conflict graph* in which a vertex represents a

candidate site and an edge indicates that the two sites conflict with each other. Each vertex v in the *conflict graph* is weighted by its *relevance* $\text{cov}(v)/\left(1+t(v)\right)$ where $t(v) = \sum_{u \in N(v)} q_{g(u)} / \sum_i q_i$, $\text{cov}(v)$ is the number of alignments covering the candidate chimeric site corresponding to v, $N(v)$ is the set of vertices connected to v, $g(u)$ is the optical molecule or contig corresponding to u, and q_i is the quality score for contig/molecule i. The variable i ranges from 1 to the sum of the number of contigs plus the number of optical molecules. Values q_i are provided by the users. By default all optical molecules are given quality 1.5 and all contigs are given quality 1. The value of $\text{cov}(v)$ is the main factor in deciding whether to cut the contigs or the molecule in order to resolve an alignment conflict. When $\text{cov}(v)$ is a tie, the denominator in the relevance formula breaks the tie based on the "trust" users have on the optical map vs. the assemblies.

While building the *conflict graph*, candidate chimeric sites which are close to each other (i.e., when their distance is smaller than a minimum threshold) are merged into the same vertex. Then, among the set of vertices which covers all the edges, we identify the subset with the smallest total weight. To speed up the process, we find the minimum vertex cover of each connected component of the conflict graph. We run the exhaustive (optimal) algorithm on small components and Clarkson's 2-approximation algorithm on larger components [2]. In the third phase, contigs and molecules are cut at the chimeric sites determined by the solution of the minimum vertex cover.

Supplemental Note 2: Cowpea data set and evaluation criteria

In this note, we describe how we created the real and synthetic cowpea datasets used for the evaluation of CHIMERICOGNIZER and BIONANO HYBRID SCAFFOLD and the specific criteria for evaluation.

Sequencing data. We tested our tool on synthetic and real data of cowpea (*Vigna unguiculata*). Cowpea is a legume crop that is resilient to hot and drought-prone climates, and a primary source of protein in sub-Saharan Africa and other parts of the developing world. Cowpea is a diploid with a chromosome number 2n = 22 and an estimated genome size of 620 Mb. The genome has very low heterozygosity, in practice it can be consider haploid. We sequenced an elite African variety (IT97K-499-35) using single-molecule real-time sequencing (Pacific Biosciences RSII). A total of 87 SMRT cell yielded about 6M reads for a total of 56.84 Gbp (91.7x genome equivalent). The raw PacBio reads are available in the public domain at NCBI SRA sample SRS3721827 (study SRP159026).

Assemblies. To test CHIMERICOGNIZER we generated multiple assemblies from the PacBio data described above with a mix of parameters, polishing qualities and assembly tools. We used CANU [4], FALCON [1] and ABRUIJN [6] to generate eight assemblies. CANU was run with different parameters to generate six of the eight assemblies (parameters shown in Supplemental Table 1). CANU_4, CANU_5 and CANU_6 were polished with QUIVER.

Optical maps. We used two Bionano Genomics optical maps. The first optical map was obtained using the BspQI nicking enzyme (which recognizes "GCTCTTC"), while the second was obtained using the BssSI nicking enzyme (which recognizes "CACGAG"). The BspQI optical map had 508 assembled optical molecules with a molecule N50 of 1.62 Mb and a total length of 622.21 Mb. The BssSI optical map had 743 assembled optical molecules with a molecule N50 of 1.02 Mb and a total length of 577.76 Mb. Both optical maps were assembled at UC Davis using the Bionano IRYSSOLVE Assembler.

Synthetic chimeric contigs. To generate synthetic datasets with artificial chimeric contigs, we first used CHIMERICOGNIZER to remove and split possible chimeric contig from the eight assemblies described above. For each of the eight chimeric-free assemblies, we injected artificial chimeric contigs by pairwise joining 2% of the contigs selected at random. We create mis-joins only for contigs longer than 500 Kbp. Results of these simulations for CHIMERICOGNIZER are reported in Supplemental Table 3 (two optical maps) and Supplemental Table 4 (one optical map). Results of these simulations for BIONANO HYBRID SCAFFOLD are reported in Supplemental Table 5.

Synthetic chimeric optical molecules. To generate synthetic datasets with artificial chimeric optical molecules, we first used CHIMERICOGNIZER to remove and split possible chimeric molecules from the two optical maps described above. For each of the two chimeric-free optical maps, we created a corresponding synthetic optical map by pairwise joining 0.5% of the molecules selected at random. We created mis-joins only on molecules longer than 1 Mbp. These synthetic optical maps were given in input to CHIMERICOGNIZER along with the eight original cowpea assemblies. To produce a more realistic simulation we decided to use the original cowpea assemblies instead of chimeric-free assemblies. Results of these simulations are reported in Supplemental Table 6.

Parameters. In all the experiments, CHIMERICOGNIZER was run using default parameters (`-a 1.5 -b 1 -d 25 -e 50000 -h 50000 -r 80000`). Please refer to the README at `https://github.com/ucrbioinfo/Chimericognizer` for details about these parameters. CHIMERICOGNIZER's pipeline is illustrated in Supplemental Figure 1. BIONANO HYBRID SCAFFOLD was run using default parameters, i.e., we executed the script `hybridScaffold.pl` (v.4741) with the parameters in the XML file `hybridScaffold_config.xml`

Evaluation. To evaluate the performance of CHIMERICOGNIZER and BIONANO HYBRID SCAFFOLD on the datasets containing synthetic chimeric contigs, we measured precision and recall by comparing its results to the "ground truth". The same approach was used to measure the performance of these tools on the datasets containing synthetic chimeric optical molecules. Supplemental Figure 3 illustrates how we computed true positives, false negatives, false positives and true negatives. When a contig contains a known mis-join (TOP, condition positive), a tool may decide to cut it (true positive) or not (false negative). When a contig does not contain a mis-join (BOTTOM, condition negative), a tool may decide to cut it (false positive) or not (true negative). Precision is defined as TP/(TP+FP). Sensitivity is defined as TP/(TP+FN).

For BIONANO HYBRID SCAFFOLD the list of contigs classified as positives are those marked `cut` in the 7th and 8th column (corresponding to `ref_leftBkpt_toCut` and `ref_rightBkpt_toCut`, respectively) of output file `conflicts_cut_status.txt`. For CHIMERICOGNIZER the list of contigs classified as positives are those that are listed in the output file `qry_cuts.txt`. Among these, we determined which ones are true positive by matching them against the "ground truth".

Supplemental Note 3: *D. melanogaster* data set

We also tested the performance of CHIMERICOGNIZER and BIONANO HYBRID SCAFFOLD on the *Drosophila melanogaster* (ISO) dataset from [8].

Assemblies. We downloaded three *D. melanogaster* assemblies generated in [8] (`https://github.com/danrdanny/Nanopore_ISO1`). The first assembly (295 contigs, total size = 141 Mb, N50 = 3 Mb) was generated using CANU [4] on Oxford Nanopore (ONT) reads longer than 1kb. The second assembly (208 contigs, total size = 132 Mb, N50 = 3.9 Mb) was generated using MINIMAP and MINIASM [5] using only ONT reads. The third assembly (339 contigs, total size = 134 Mb, N50 = 10 Mb) was generated by PLATANUS [3] and DBG2OLC [10] using 67.4x of Illumina paired-end reads and the longest 30x ONT reads. The first and third assemblies were polished using NANOPOLISH [7] and PILON [9].

Optical map. The Bionano Genomics optical for *D. melanogaster* map was provided by the authors of [8]. This optical map (363 molecules, total size = 246 Mb, N50 = 841 kb) was created using IRYSSOLVE 2.1 from 78,397 raw Bionano molecules (19.9 Gb of data with a mean read length 253 kb).

Reference genome. We used release 6.21 of the *D. melanogaster* genome, downloaded from FlyBase (`http://www.flybase.org`).

Parameters. CHIMERICOGNIZER was run using parameters (`-a 0.5 -b 1.0 -d 25 -e 100000 -h 100000 -r 80000`). Please refer to the README at `https://github.com/ucrbioinfo/Chimericognizer` for details about these parameters. BIONANO HYBRID SCAFFOLD was run with using default parameters, i.e., we executed the script `hybridScaffold.pl` (v.4741) with the parameters defined in the XML file `hybridScaffold_config.xml`

Evaluation. To evaluate the performance of CHIMERICOGNIZER and BIONANO HYBRID SCAFFOLD on *D. melanogaster* assemblies, we measured precision and sensitivity by comparing its results to the "ground truth" (reference genome). To determine which contigs were truly chimeric (i.e., the true positive set), we first selected all contigs from the three assemblies which (i) could be aligned to the optical map via REFALIGNER with a minimum confidence of at least 25 and (ii) had at least one BLAST alignment (v2.7.1, default parameters) to the reference genome with an e-value lower than 1e-50 and an alignment length higher than 8 kbp. A total of 73 contigs satisfied these two conditions. Among all the contigs that satisfied (i) and (ii), we defined a contig C to be a *true chimeric contig* if C had at least two alignments which satisfied any of the following three conditions: (1) C aligned to different chromosomes; (2) the orientation of C's alignments were different; or (3) the difference between the distance of alignments on the contig and the distance of alignments on the reference sequence was larger than 100 Kbp. A total of 6 contigs were identified as chimeric (out of 73). Precision and Sensitivity were defined as for cowpea (Supplemental Note 2). Experimental results are reported in Supplemental Table 9 for CHIMERICOGNIZER, and Supplemental Table 10 for BIONANO HYBRID SCAFFOLD.

References

[1] Chen-Shan Chin et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature methods*, 13(12):1050–1054, 2016.

[2] Kenneth L Clarkson. A modification of the greedy algorithm for vertex cover. *Inf. Process. Lett.*, 16(1):23–25, January 1983.

[3] Rei Kajitani, Kouta Toshimoto, Hideki Noguchi, Atsushi Toyoda, Yoshitoshi Ogura, Miki Okuno, Mitsuru Yabana, Masayuki Harada, Eiji Nagayasu, Haruhiko Maruyama, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome research*, 24(8):1384–1395, 2014.

[4] Sergey Koren et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, 27(5):722–736, 2017.

[5] Heng Li. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, 32(14):2103–2110, 2016.

[6] Yu Lin et al. Assembly of long error-prone reads using de Bruijn graphs. *Proceedings of the National Academy of Sciences*, 113(52):E8396–E8405, 2016.

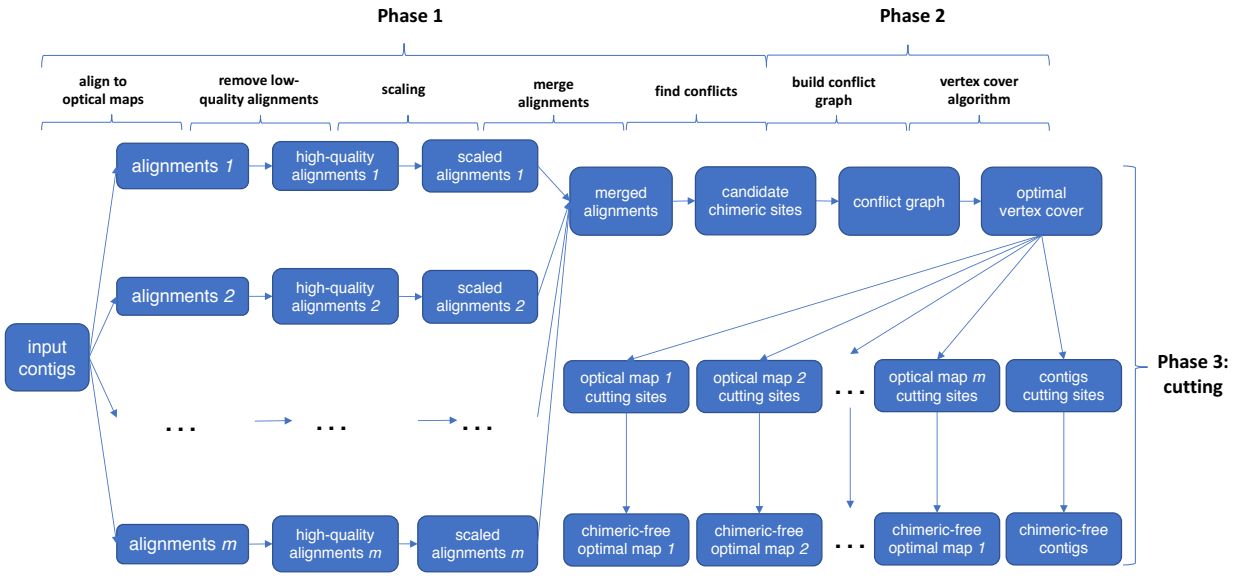
[7] Nicholas J Loman, Joshua Quick, and Jared T Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature methods*, 12(8):733, 2015.

[8] Edwin A. Solares, Mahul Chakraborty, Danny E. Miller, Shannon Kalsow, Kate Hall, Anoja G. Perera, J. J. Emerson, and R. Scott Hawley. Rapid low-cost assembly of the drosophila melanogaster reference genome using low-coverage, long-read sequencing. *G3: Genes, Genomes, Genetics*, 2018.

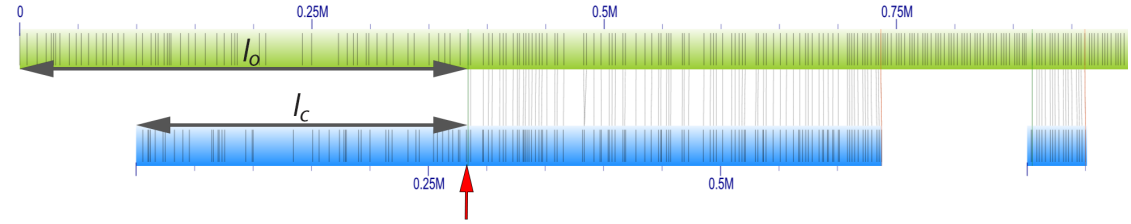
[9] Bruce J Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K Young, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one*, 9(11):e112963, 2014.

[10] Chengxi Ye, Christopher M Hill, Shigang Wu, Jue Ruan, and Zhanshan Sam Ma. Dbg2olc: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Scientific reports*, 6:31900, 2016.

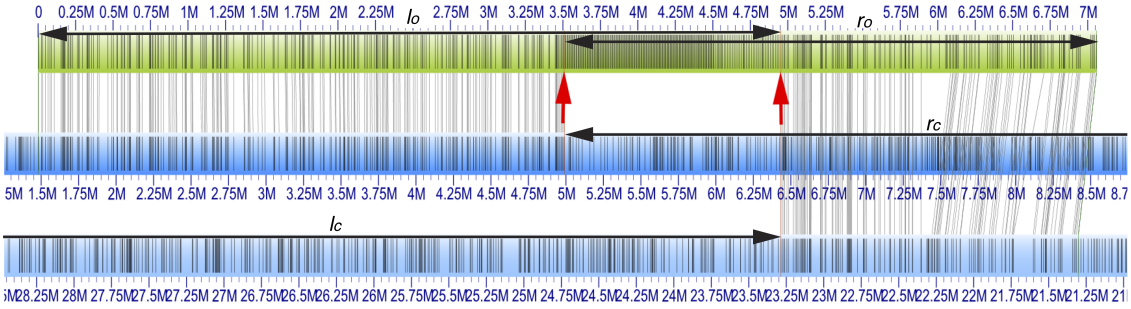
Supplementary Figures and Tables



Supplemental Figure 1: Algorithmic pipeline of CHIMERICOGNIZER

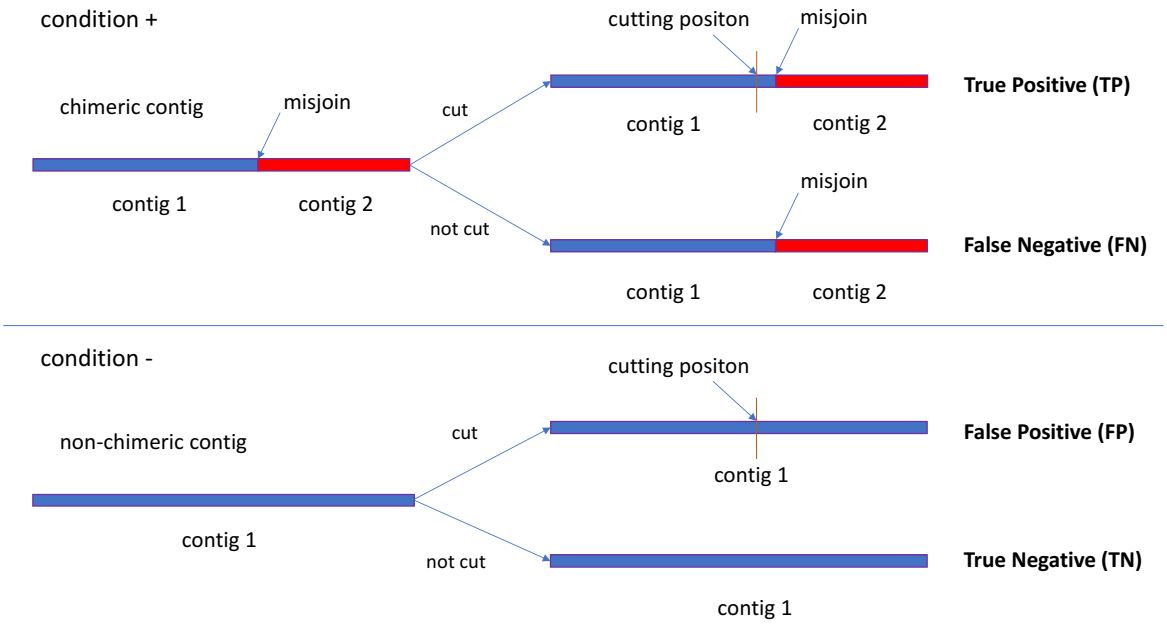


(A)



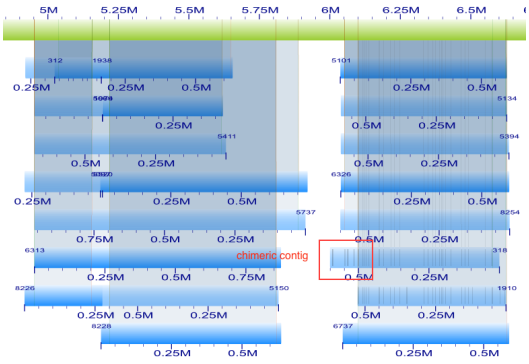
(B)

Supplemental Figure 2: Examples of a conflicting alignment between an optical molecule (green) and an assembled contig (blue); vertical lines indicate the location of restriction enzyme sites; (A) a chimeric contig (blue) and its candidate location for a split indicated by the red arrow (l_o is the optical molecule left overhang, l_c is the contig left overhang; the left end of alignment is declared a *conflict site* if i) both l_o and l_c are longer than some minimum length (default 50 kbp) and ii) at least one restriction enzyme sites appear in both l_o and l_c; both conditions are satisfied in this case); (B) a chimeric optical molecule (green) and candidate locations for splits indicated by the red arrows (l_o is the optical molecule left overhang, l_c is the contig left overhang, r_o is the optical molecule right overhang, r_c is the contig right overhang)

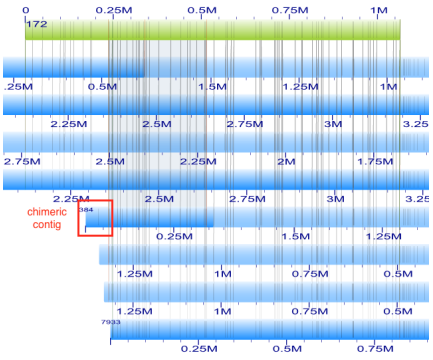


Supplemental Figure 3: Illustrating how we computed true positives, false negatives, false positives and true negatives; when a contig contains a mis-join (TOP, condition positive), CHIMERICOGNIZER may decide to cut it (true positive) or not (false negative); when a contig does not contain a mis-join (BOTTOM, condition negative), CHIMERICOGNIZER may decide to cut it (false positive) or not (true negative); precision is TP/(TP+FP), sensitivity is TP/(TP+FN)

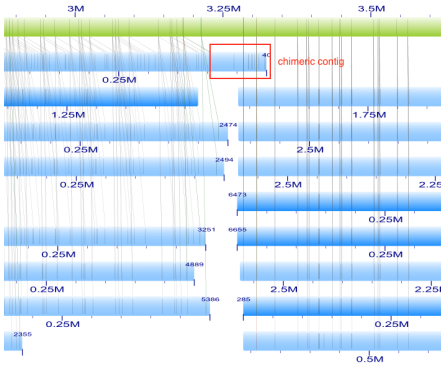
(A)



(B)



(C)



Supplemental Figure 4: A few examples of chimeric contigs missed by the human expert, but correctly identified by CHIMERICOGNIZER

CANU assembly	corMhapSensitivity	corMaxEvidenceErate	corOutCoverage	QUIVER
1	high	default	default	
2	high	0.15	100	
3	normal	0.15	100	
4	high	default	100	✓
5	low	default	default	✓
6	low	default	100	✓

Supplemental Table 1: Parameter choices for CANU v1.3: three assemblies were polished with QUIVER

	CHIMERICOGNIZER with two optical maps							
	ABRUIJN	FALCON	CANU$_1$	CANU$_2$	CANU$_3$	CANU$_4$	CANU$_5$	CANU$_6$
contig N50 (bp)	2,084,664	2,918,725	3,427,506	3,175,625	2,798,135	5,633,882	5,312,333	4,757,094
contig L50	69	47	42	48	50	28	27	31
total assembled (bp)	478,230,679	511,933,729	504,711,938	516,558,510	515,964,327	511,101,122	506,285,539	517,496,317
# contigs	516	1,826	1,061	1,099	1,125	948	879	948
# contigs ≥100kbp	410	399	287	340	316	269	201	277
# contigs ≥1Mbp	149	115	125	135	141	94	98	103
# contigs ≥10Mbp	0	1	2	4	2	10	9	10
longest contig (bp)	9,801,038	10,554,495	14,090,735	14,331,160	12,496,821	17,211,165	18,473,372	18,498,533
Illumina reads, % mapped (202M)	99.72399%	99.58149%	99.97449%	99.97389%	99.97389%	99.97743%	99.97343%	99.97763%
Illumina reads, % properly paired (202M)	92.29997%	91.94896%	92.54645%	92.63437%	92.62722%	92.64222%	92.62153%	92.64414%
Illumina reads, % mapped, MapQ≥30 (202M)	64.20883%	59.48734%	64.65541%	63.00774%	63.47912%	64.80935%	64.85658%	64.59832%
total length with 100% consistent LG (bp)	425,557,449	344,074,378	421,565,015	418,588,863	409,262,310	425,812,490	423,058,141	420,659,561
	CHIMERICOGNIZER with one optical map							
	ABRUIJN	FALCON	CANU$_1$	CANU$_2$	CANU$_3$	CANU$_4$	CANU$_5$	CANU$_6$
contig N50 (bp)	2,084,664	3,000,247	3,427,506	3,175,625	2,798,135	5,633,882	5,312,333	4,757,094
contig L50	69	46	42	48	50	28	27	31
total assembled (bp)	478,230,679	511,933,729	504,711,938	516,558,510	515,964,327	511,101,122	506,285,539	517,496,317
# contigs	510	1,814	1,059	1,098	1,125	947	879	947
# contigs ≥100kbp	407	391	286	340	316	268	201	277
# contigs ≥1Mbp	149	115	125	135	141	94	98	103
# contigs ≥10Mbp	0	1	2	4	2	10	9	10
longest contig (bp)	9,801,038	10,554,495	14,090,735	14,331,160	12,496,821	17,211,165	18,473,372	18,498,533
Illumina reads, % mapped (202M)	99.72400%	99.58149%	99.97449%	99.97389%	99.96996%	99.97743%	99.97343%	99.97763%
Illumina reads, % properly paired (202M)	92.29986%	91.94953%	92.54646%	92.63438%	92.62728%	92.64221%	92.62152%	92.64384%
Illumina reads, % mapped, MapQ≥30 (202M)	64.20894%	59.48738%	64.65538%	63.00775%	63.47915%	64.80937%	64.85659%	64.59879%
total length with 100% consistent LG (bp)	425,557,449	344,074,378	421,565,015	418,588,863	409,262,310	425,812,490	423,058,141	420,659,561
	Chimeric contigs detected/removed manually by an expert							
	ABRUIJN	FALCON	CANU$_1$	CANU$_2$	CANU$_3$	CANU$_4$	CANU$_5$	CANU$_6$
contig N50 (bp)	1,896,002	2,869,362	3,280,469	2,797,949	2,666,731	5,340,274	4,859,617	4,498,063
contig L50	74	49	42	51	55	29	30	32
contig NG50 (bp)	1,330,435	1,737,012	2,431,239	1,949,515	2,068,575	3,451,071	3,767,556	3,417,577
contig LG50	119	73	63	73	77	42	43	45
total assembled (bp)	478,230,679	511,933,729	503,187,311	516,537,734	515,949,175	507,773,747	506,154,442	516,817,613
# contigs	538	1,820	1,038	1,110	1,140	897	894	928
# contigs ≥100kbp	437	404	299	354	334	278	220	288
# contigs ≥1Mbp	151	118	128	142	145	103	104	107
# contigs ≥10Mbp	0	1	2	2	0	9	7	8
longest contig (bp)	8,846,014	10,554,495	14,090,735	14,331,160	9,775,097	17,211,165	18,473,372	18,498,533
Illumina reads, % mapped (202M)	99.72397%	99.58150%	99.94933%	99.97389%	99.94468%	99.97474%	99.96894%	99.97707%
Illumina reads, % properly paired (202M)	92.30106%	91.95107%	92.52969%	92.63057%	92.62330%	92.59763%	92.59433%	92.64181%
Illumina reads, % mapped, MapQ≥30 (202M)	64.21367%	59.49035%	64.38425%	63.00587%	63.22414%	62.84466%	64.35764%	63.50279%
total length with 100% consistent LG (bp)	379,029,914	312,593,019	356,505,616	349,534,672	347,586,448	425,812,490	331,956,528	338,556,993

Supplemental Table 2: Assembly statistics of the eight cowpea assemblies after chimeric contigs were removed (top) by CHIMERICOGNIZER using two optical map, (middle) by CHIMERICOGNIZER using one optical map, and (bottom) by an expert; reads were mapped with BWA

	ABRUIJN	FALCON	CANU$_1$	CANU$_2$	CANU$_3$	CANU$_4$	CANU$_5$	CANU$_6$
# TP	10.8	35.5	21.0	20.8	22.4	17.8	17.2	18.0
# TP + FP	10.8	35.9	21.7	20.8	23.0	18.6	17.3	18.0
# P	11.0	37.0	22.0	22.0	23.0	19.0	18.0	19.0
precision	100.00%	98.92%	96.79%	100.00%	97.45%	95.70%	99.44%	100.00%
sensitivity	98.18%	95.95%	95.45%	94.55%	97.39%	93.68%	95.56%	94.74%
avg position error (bp)	16,704	26,380	32,054	18,426	19,415	38,338	17,753	18,809

Supplemental Table 3: Performance statistics for CHIMERICOGNIZER on the eight cowpea assemblies injected with synthetic chimeric contigs (i.e., 2% of the contigs longer than 500 Kbp selected at random where joined) and two optical maps; values in this table are the averages over ten experiments; TP, FP and P represent true positive, false positive and positive, respectively; avg position error is the average distance in base pairs between CHIMERICOGNIZER's cutting position and the true mis-join position

	ABRUIJN	FALCON	CANU$_1$	CANU$_2$	CANU$_3$	CANU$_4$	CANU$_5$	CANU$_6$
# TP	9.5	30.8	17.5	18.8	18.9	15.3	14.6	14.5
# TP + FP	9.5	31.7	17.5	19.2	19.7	15.3	14.6	14.5
# P	11.0	37.0	22.0	22.0	23.0	19.0	18.0	19.0
precision	100.00%	97.17%	100.00%	98.04%	96.05%	100.00%	100.00%	100.00%
sensitivity	86.36%	83.24%	79.55%	85.45%	82.17%	80.53%	81.11%	76.32%
avg position error (bp)	17,560	27,969	18,506	21,778	73,255	19,853	16,693	22,266

Supplemental Table 4: Performance statistics for CHIMERICOGNIZER on the eight cowpea assemblies injected with synthetic chimeric contigs (i.e., 2% of the contigs longer than 500 Kbp selected at random where joined) and one optical map (BspQI); values in this table are the averages over ten experiments; TP, FP and P represent true positive, false positive and positive, respectively; avg position error is the average distance in base pairs between CHIMERICOGNIZER's cutting position and the true mis-join position

	ABRUIJN	FALCON	CANU$_1$	CANU$_2$	CANU$_3$	CANU$_4$	CANU$_5$	CANU$_6$
# TP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
# TP + FP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
# P	11.0	37.0	22.0	22.0	23.0	19.0	18.0	19.0
precision	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
sensitivity	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
avg position error (bp)	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Supplemental Table 5: Performance statistics for BIONANO HYBRID SCAFFOLD on the eight cowpea assemblies injected with synthetic chimeric contigs (i.e., 2% of the contigs longer than 500 Kbp selected at random where joined) and one optical map (BspQI); values in this table are the averages over ten experiments; TP, FP and P represent true positive, false positive and positive, respectively; avg position error is the average distance in base pairs between BIONANO HYBRID SCAFFOLD's cutting position and the true mis-join position

	one optical map		two optical maps	
	BspQI	BssSI	BspQI	BssSI
# TP	2.3	3.4	2.8	3.7
# TP + FP	2.3	3.4	2.8	3.7
# P	3.0	4.0	3.0	4.0
precision	100.00%	100.00%	100.00%	100.00%
sensitivity	76.67%	85.00%	93.33%	92.50%

Supplemental Table 6: Performance statistics for CHIMERICOGNIZER on cowpea datasets composed by one or two synthetic optical maps and eight real assemblies; for the "one optical map" column, we injected chimeric optical molecules in either BspQI or BssSI, ran CHIMERICOGNIZER on that optical map, and measured precision/sensitivity on the molecules of that optical map; for the "two optical maps" column, we injected chimeric optical molecules in both optical maps, ran CHIMERICOGNIZER with two optical maps, and measured precision/sensitivity on molecules of each optical map separately; values in this table are the averages over ten experiments; TP, FP and P represent true positive, false positive and positive, respectively

# assemblies	1	2	3	4	5	6	7	8
# TP	20.2	39.3	56.9	78.7	107.0	121.5	142.5	163.5
# TP + FP	22.4	40.1	57.6	80.5	108.5	123.6	144.4	166.1
# P	21.6	41.6	60.2	83.0	112.5	127.7	149.1	171.0
precision	89.35%	97.86%	98.75%	97.70%	98.59%	98.33%	98.69%	98.44%
sensitivity	93.34%	94.39%	94.55%	94.81%	95.05%	95.06%	95.59%	95.61%
average position error (bp)	121,396	17,935	20,852	18,905	29,384	25,395	33,402	24,274

Supplemental Table 7: Performance statistics for CHIMERICOGNIZER on synthetic cowpea datasets composed of a variable number of assemblies and two optical maps; values in this table represent the total for all assemblies selected (averaged over ten experiments); TP, FP and P represent true positive, false positive and positive, respectively; avg position error is the average distance in base pairs between CHIMERICOGNIZER's cutting position and the true mis-join position

# assemblies	1	2	3	4	5	6	7	8
# TP	18.3	34.7	50.1	66.7	85.8	106.6	122.7	139.9
# TP + FP	25.8	38.2	51.9	68.1	87.1	108.1	124.1	142.0
# P	22.3	42.4	63.8	83.5	103.0	131.2	151.8	171.0
precision	68.43%	91.06%	96.64%	98.00%	98.56%	98.67%	98.87%	98.52%
sensitivity	81.49%	82.69%	78.76%	80.36%	83.22%	81.25%	80.85%	81.81%
average position error (bp)	270,414	102,461	19,633	41,662	21,143	25,795	25,468	29,249

Supplemental Table 8: Performance statistics for CHIMERICOGNIZER on synthetic cowpea datasets composed of a variable number of assemblies and one optical map (BspQI); values in this table represent the total for all assemblies selected (averaged over ten experiments); TP, FP and P represent true positive, false positive and positive, respectively; avg position error is the average distance in base pairs between CHIMERICOGNIZER's cutting position and the true mis-join position

# TP	5
# TP + FP	6
# P	6
precision	83.33%
sensitivity	83.33%

Supplemental Table 9: Performance statistics for CHIMERICOGNIZER on the *D. melanogaster* dataset (composed by one optical map and three assemblies); TP, FP and P represent true positive, false positive and positive, respectively

# TP	0
# TP + FP	5
# P	6
precision	0.00%
sensitivity	0.00%

Supplemental Table 10: Performance statistics for BIONANO HYBRID SCAFFOLD on the *D. melanogaster* dataset (composed by one optical map and three assemblies); TP, FP and P represent true positive, false positive and positive, respectively