Improved Gossypium Raimondii Genome Using a Hi-C-based Proximity-Guided Assembly

Huoli Song (sglzms@163.com)
Chinese Academy of Agricultural Sciences Cotton Research Institute https://orcid.org/0000-0003-3236-9286

Qiuhong Yang
Chinese Academy of Agricultural Sciences Cotton Research Institute

Dongyun Zuo
Chinese Academy of Agricultural Sciences Cotton Research Institute

Qiaolian Wang
Chinese Academy of Agricultural Sciences Cotton Research Institute

Hailiang Cheng
Chinese Academy of Agricultural Sciences Cotton Research Institute

Xiaoxu Feng
Chinese Academy of Agricultural Sciences Cotton Research Institute

Javaria Ashraf
Chinese Academy of Agricultural Sciences Cotton Research Institute

Youping Zhang
Chinese Academy of Agricultural Sciences Cotton Research Institute

Simin Li
Chinese Academy of Agricultural Sciences Cotton Research Institute

Xiaoqin Chen
Chinese Academy of Agricultural Sciences Cotton Research Institute

Research

Keywords: Gossypium raimondii, Hi-C, genome assembly, heatmap and collinearity

Posted Date: August 24th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-63005/v1

Version of Record: A version of this preprint was published at Journal of Cotton Research on August 5th, 2021. See the published version at https://doi.org/10.1186/s42397-021-00096-2.

Abstract

Genome sequence plays an important role both in basic and applied studies. *Gossypium raimondii*, the putative contributor of the D subgenome of Upland cotton (*G. hirsutum*), highlights the need to improve the genome quality in a rapid and efficient way. Here, we performed Hi-C sequencing of *G. raimondii* and reassembled its genome based on new Hi-C data and previously published scaffolds. We identified and corrected errors of initial scaffolds before reassembled into chromosomes. In total 98.42% of sequence was clustered successfully, among which 99.72% of the clustered sequence was ordered and 99.92% of the ordered sequence was oriented with high-quality. Further evaluation of results by heat-map and collinearity analysis revealed that the current reassembled genome is significantly improved than previous one. This improvement in *G. raimondii* genome not only provides a better reference genome to increase study efficiency, but also offers a new way to assemble cotton genomes. Furthermore, Hi-C data of *G. raimondii* may be used for 3D structure research or regulating analysis.

Introduction

Over the last decade, next generation sequencing (NGS) technologies has brought immense improvements in plant genome sequencing throughput and cost, and many plant genomes have been sequenced using this technology such as *F. vesca* (Shulaev et al., 2011), *C. cajan* (Varshney et al., 2012), *G. raimondii* (Wang et al., 2012; Paterson et al., 2012, Udall et al., 2019), *G. arboreum* (Li et al., 2014) and *G. hirsutum* (Li et al., 2015; Zhang et al., 2015). However, *de novo* assembly of large eukaryotic genomes has remained a great challenge with the NGS platform due to significant amount of repeat contents (Rounsley et al., 2009). As a result, *de novo* assembly of chromosome-scale scaffolds has become a major constraint to the completion of high-quality genome sequence. Compared with the traditional method, high-throughput chromosome conformation capture (Hi-C) technology has assisted in the assembly of long scaffolds to produce chromosome-scale genome assemblies (Lightfoot et al., 2017). The Hi-C technique is based on existing scaffolds and restriction enzyme cutting sites, which are more evenly distribution and have much higher density.

The Hi-C-based proximity guided assembly was initially developed to study the three-dimensional (3-D) conformation of chromosomes of yeast gene expression (van Berkum et al. 2010). There are two regular models with Hi-C data: the first one is that the rate of Hi-C interaction is inversely proportional to the genomic distance between the pairs of loci, the second one is that the rate of Hi-C interaction of pairs of loci within a chromosome is significantly higher than that in different chromosomes (Xie et al. 2015). Based on these two model, Hi-C-based proximity-guided assembly was applied for *de novo* assembly of human, and subsequently for the assembling of mouse and *Drosophila* genomes which reported good results or improvement (Burton et al. 2013). With the success of testing and verifying this method in *Arabidopsis thaliana* (Xie et al. 2015), Hi-C based proximity-guide assembly has reported as an effective and efficient method which subsequently has been used in many other plants.

Cotton is one of the most economically important crops in the world, and its fiber is a natural and renewable source for textile industry worldwide. Besides its commercial value, cotton also serves as a perfect model system for studying cell wall biosynthesis (Zhang et al. 2018), cell elongation (Guo et al. 2017) and polyploidization (Yuan et al. 2015). The *Gossypium* genus comprises of more than 50 species including at least 5 tetraploid species and 45 diploid species. Diploid cottons are divided into 8 sub-genomes, denoted A-G and K based on chromosome pairing relationships (Wendel et al., 1992). Tetraploid cottons, such as cultivated *G. hirsutum* (AD_1) and *G. barbadense*(AD_2), had formed by an allopolyploidy event about 1–2 million years ago (Paterson et al. 2012). These tetraploid cotton species share common ancestors with the modern New World species *G. raimondii* (D_5) and the Old World A-genome species *G. herbaceum* (A_1) or *G. arboreum* (A_2). Previously, genomes of different cotton species sequenced and assembled including *G. raimondii* (Wang et al. 2012), *G. arboreum* (Li et al. 2014), and *G. hirsutum* (Li et al., 2015).

Among these cotton species, the genome of *G. raimondii* has the lowest complexity which has been sequenced and assembled using the next-generation Illumina paired-end sequencing strategy (Wang et al. 2012). Approximately 73% (281 scaffolds) of the assembled sequences were anchored to 13 chromosomes, covering 88.1% of the genome, while only 52.4% (228 scaffolds) of total sequence was both ordered and oriented. The completeness and accuracy of previous sequenced and assembled genome of *G. raimondii* (Wang et al., 2012) was relatively low due to higher numbers of repeat elements and low numbers of genetic markers. In the present study, we conducted a *de novo* Hi-C sequencing of *G. raimondii* genome, and incorporated the new Hi-C data with the existing *G. raimondii* scaffolds (Wang et al., 2012) to improve the quality of the D-genome.

Methods

1. Tissue collection and Hi-C sequencing

1.1 Plant materials

The seeds of *G. raimondii* $D_{5\text{-}1}$ were planted in an incubator at constant environmental condition having 27°C temperature, 60% relative humidity, 16/8-h light/dark photoperiod, and 100% fluorescent light. When sixth euphylla came out, these seedlings were transplanted into big pots. Approximately 3-gram young leaves from *G. raimondii* plants were collected and immediately treated with formaldehyde.

1.2 Hi-C pipeline

During this study, we have used the same Hi-C pipeline as in *Arabidopsis thaliana* (Xie, Zheng et al. 2015). Before start this experiment, we have tested the integrity of DNA from the formaldehyde-treated tissue, and then the DNA was isolated and digested by *Mbo*I instead of *Hin*dIII because of the shorter recognition site (only four bases of *Mbo*I). The resulting sticky ends were filled with nucleotides in which cytosine is biotinylated, and ligated the adjacent blunt ends to a chimeric circle under extremely dilute conditions. Subsequently, DNA was purified and broken into 300-500 base pairs using ultrasonic, pull-down the biotin

labeled DNA and performed the PCR reaction (10 cycles). After DNA purification, the finished Hi-C library was sequenced with an Illumina Hiseq (PE150). A total of 570,412,361 read-pairs were obtained.

2. Genome assembly based on Hi-C data

Assembling of *G. raimondii* genome involved three steps. First, valid Hi-C paired-end reads and contact matrix with a resolution of 100 kb were generated by HiC-Pro (Servant, Varoquaux et al. 2015). The raw sequence data with low quality, unmapped and invalid mapped paired reads were filtered out by HiC-Pro and contact matrix based on interaction frequency was created. At the second step, the *G. raimondii* genome was assembled with the Hi-C data by Lachesis (Burton, Adey et al. 2013),which contained clustering, ordering and orienting. Lastly, the assembled *G. raimondii* genome was assessed by Mummer and Python scripts, resulting heat-map and collinear analysis.

Results

1. Hi-C data analysis by HiC-Pro

1.1 Mapping and filtering

Initially, the low-quality and invalid paired-reads were filtered out which was caused by sequencing errors from the raw Hi-C reads of *G. raimondii* (Table 1). Results revealed that 95.6% of sequence is clean Q30 bases, showing a good quality of sequence data. Then, clean Hi-C reads were mapped to the previously sequenced genome of *G. raimondii* (Wang et al., 2012) using Bowtie2 (Langmead and Salzberg 2012), and unique mapped paired-end reads were retained (Fig.1 and Table 2). Subsequent analysis was performed to remove the invalid paired-reads. The reference genome was broken into restriction fragments by cutting them at the MobI restriction enzyme site “GTAC”. Approximately 54.7% uniquely mapped paired-read was aligned to single restriction fragments. Among uniquely mapped reads, valid paired-end reads were present in different restriction fragments, but non-ligation, self-ligation and dangling end paired-end reads were recognized by mapping orientation information in the same restriction fragment (Belton, McCord et al. 2012) (Table 3, Fig. 2). After the Hi-C data were filtered out, results showed that 81.95% of uniquely mapped sequences are valid paired-end reads. Thus, the valid paired-end reads (223,304,666) were used for further analysis.

1.2 Creating contact matrix

The genome was into non-overlapping 100-kb windows, and the number of valid paired-end reads in the 100-kb windows was referred as the contact count. The Hi-C contact matrix was built and normalized by its restriction sites because the Hi-C signal was in linear proportion to the number of restriction sites.

1.3 Identification and correction of errors within scaffolds

Errors in scaffolds of the initial draft assembly were identified and corrected following the Aedes aegypti’s *de novo* assembly procedure (Dudchenko, Batra et al. 2017). Briefly, the errors were corrected by

identifying the bins where a scaffold's long-range contact pattern changes abruptly, which is unlikely for a correct scaffold. We cut out the error bins as a new scaffold. There are 259 errors within scaffolds. The list of error bins is presented in Supplemental.1.

2 Genome assembly by Lachesis

2.1 Clustering

Lachesis is a computational method that exploits Hi-C data sets for *de novo* genome assemblies (Burton, Adey et al. 2013). Hi-C data has two classical models, Hi-C interactions within one chromosome are distinctly more than it between two chromosomes, and Hi-C interactions between two loci are inversely proportional to their distance. Based on the first model, Lachesis clustered the scaffolds into 13 groups by agglomerative hierarchical clustering (Table 4). A total of 2883 scaffolds were clustered successfully.

2.2 Ordering and orientation

In each clustered group, an acyclic spanning tree was built with vertexes corresponding to the scaffolds, while edge weights representing the normalized Hi-C interactions between pairs of scaffolds. A total of 1,328 scaffolds (744,578,885 bp, representing 99.72% of the total length) were ordered by Lachesis, among which 697 scaffolds (724,960,878 bp, representing 97.37% of the total length) are "trunk" (Table 5). For each ordered group, the acyclic spanning tree was built to represent all of the possible ways to orient the scaffolds. Lachesis has built a scoring function based on the difference between forward and backward interaction. Highest score represented the maximum likelihood through predicting orientations for each of the scaffolds. Among 1,328 ordered scaffolds, 1,129 scaffolds (743,948,690 bp, representing 99.92% of the ordered length) were of high scores (Table 6). All of the "trunk" scaffolds were of high score.

3 Assembling results

The genome of *Gossypiumraimondii* were reassembled using Hi-C data (Supplemental.2). About 98.42% of total sequence length was clustered successfully, among which 99.72% and 97.37% of the clustered sequence were ordered and high-quality ordered, respectively. And approximately 99.92% of the ordered sequence was oriented with high-quality. The statistics of pseudo-chromosome length is shown in Table 7,while the indicator statistics of the initial draft genome and reassembly results are shown in Table 8 and 9, respectively. From the parameters like scaffolds number, N50 and N60 of previously draft genome and reassemble genome, we found that the *G.raimondii* genome using a Hi-C-based proximity-guided assembly is clearly much better than the reported draft genome (Wang et al., 2012).

Further, the results were also verified by heat-map (Fig. 3) and collinear-analysis (Fig.4). The heat-map directly proves the validity of the processing methods. Based on the two regular models as we mentioned above, the heat-map of the reassembled results revealed that the diagonal interaction is much higher. The boundaries of each pseudo-chromosome are relatively clear and it is under low background noise that shows good reassembling results (Fig. 3). In addition, the collinear relationship between the draft (Wang

et al., 2012) and reassembled genomes (Fig. 4) showed that the current reassembled genome is quite different from the previous one. Further when we compared our results with another version of *G. raimondii* genome (Paterson et al., 2012), results showed that current reassembled genome is improved with respect to both quality and integrity (Fig. 5).

Discussion

With the development of sequencing techniques and bioinformatics tools, a high-quality genome sequence is the basis for cotton molecular breeding. In several previous studies, Illumina short sequencing reads are extensively used for *de novo* genome sequencing and assembling of different organisms (Ekblom and Wolf 2014) including cotton (Li et al., 2014; 2015). However, Illumina short sequencing reads are very short, which requires other types of sequencing data to assemble the genome such as BAC and fosmid libraries (Salzberg, Phillippy et al. 2012), jump libraries (Salzberg, Phillippy et al. 2012), optical mapping (Dong et al. 2013), genetic linkage maps (Fierst 2015), and single-molecule real-time sequencing (Bickhart et al. 2017). Previously, Wang et al. (2012) used the Illumina sequences and genetic linkage map to assemble the genome of *G. raimondii* (Wang et al. 2012), but this technique is insufficient due to low density of markers available in genetic linkage map. Hi-C-based proximity-guided assembly was able to more accurately reassemble the *Arabidopsis* genome from a set of scaffolds into chromosomes (Xie et al. 2015). Nowadays, this method has been successfully used in many species such as *Aedes aegypt* (Dudchenko, Batra et al. 2017), *Amaranthus hypochondriacus* (Lightfoot et al. 2017), *Rubus occidentalis* (Jibran et al. 2018). In the present study, we applied *de novo* Hi-C sequencing of the *G. raimondii* genome to improve the quality and accuracy of its previously reported draft genome from Wang et al. (2012).

Results from the comparative analysis of different parameters between the draft genome (Wang et al., 2012) and the current reassembled genome showed a significantly improved quality as compared to previous one. Such as we increased the rate of clustering from 73% of the previous draft assembly to 98.42% of current reassembly. Similarly, the rates of ordering and orienting were also from 52.4% (previous draft assembly) to 98.07% (current reassembly), confirming the better quality of current reassembled *G. raimondii* genome.

Previously, Paterson et al. (2012) also sequenced and assembled the *G. raimondii* genome with good results and abundant markers. However, these markers are not evenly distributed across the genome which might indicate some errors in its genome assembly. Thus in the current study, we also compare the reassembly genome with this version of *G. raimondii* genome (Paterson et al., 2012) by the collinear analysis. Our results showed that despite of fewer differences, the current reassembled genome is improved with respect to both quality and integrity. However, based on the differences between the current reassembled genome and the genome reported by Paterson et al. (2012), further work may be necessary for integrating the two versions of G. *raimondii* genome into a best version.

Abbreviations

NGS
next generation sequencing
Hi-C
high-throughput chromosome conformation capture
CAAS
Chinese Academy of Agricultural Science

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Not applicable.

Funding

Not applicable.

Authors' contributions

QY and GS conceived and designed the experiments; all authors performed data analysis and interpretation; QY, JA and GS wrote the manuscript.

Acknowledgements

This work was supported by CAAS.

References

1. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. "Hi-C: a comprehensive technique to capture the conformation of genomes. " Methods. 2012;58(3):268–76.
2. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST, Burton JN, Huson HJ, Nystrom JC, Kelley CM, Hutchison JL, Zhou Y, Sun J, Crisa A, Ponce FA, de Leon JC, Schwartz JA, Hammond GC, Waldbieser SG, Schroeder GE, Liu MJ, Dunham J, Shendure TS, Sonstegard AM, Phillippy CP, Van Tassell, Smith TP. "Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome.". Nat Genet. 2017;49(4):643–50.

3. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. "Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions.". Nat Biotechnol. 2013;31(12):1119–25.
4. Dong Y, Xie M, Jiang Y, Xiao N, Du X, Zhang W, Tosser-Klopp G, Wang J, Yang S, Liang J, Chen W, Chen J, Zeng P, Hou Y, Bian C, Pan S, Li Y, Liu X, Wang W, Servin B, Sayre B, Zhu B, Sweeney D, Moore R, Nie W, Shen Y, Zhao R, Zhang G, Li J, Faraut T, Womack J, Zhang Y, Kijas J, Cockett N, Xu X, Zhao S, Wang J, Wang W. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). Nat Biotechnol. 2013;31(2):135–41.
5. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, Aiden EL. "De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. " Science. 2017;356(6333):92–5.
6. Ekblom R, Wolf JB. "A field guide to whole-genome sequencing, assembly and annotation.". Evol Appl. 2014;7(9):1026–42.
7. Fierst JL. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. Front Genet. 2015;6:220.
8. Guo K, Tu L, He Y, Deng J, Wang M, Huang H, Li Z, Zhang X. "Interaction between calcium and potassium modulates elongation rate in cotton fiber cells.". J Exp Bot. 2017;68(18):5161–75.
9. Jibran R, Dzierzon H, Bassil N, Bushakra JM, Edger PP, Sullivan S, Finn CE, Dossett M, Vining KJ, VanBuren R, Mockler TC, Liachko I, Davies KM, Foster TM, Chagné D. (2018). "Chromosome-scale scaffolding of the black raspberry (Rubus occidentalis L.) genome based on chromatin interaction data." Horticulture Research 5(1).
10. Jonathan F, Wendel VAA. (1992). "Phylogenetics of the cotton genus(Gossypium L.): character-state weighted parsimony analysisof chloroplast DNA restriction sites data and its systematic and biogeographic implications " JSTOR.
11. Langmead B, Salzberg SL. "Fast gapped-read alignment with Bowtie 2. " Nat Methods. 2012;9(4):357–9.
12. Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, Li Q, Ma Z, Lu C, Zou C, Chen W, Liang X, Shang H, Liu W, Shi C, Xiao G, Gou C, Ye W, Xu X, Zhang X, Wei H, Li Z, Zhang G, Wang J, Liu K, Kohel RJ, Percy RG, Yu JZ, Zhu YX, Wang J, Yu S. "Genome sequence of the cultivated cotton Gossypium arboreum.". Nat Genet. 2014;46(6):567–72.
13. Lightfoot DJ, Jarvis DE, Ramaraj T, Lee R, Jellen EN, Maughan PJ. "Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (Amaranthus hypochondriacus) chromosomes provide insights into genome evolution.". BMC Biol. 2017;15(1):74.
14. Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, Yoo MJ, Byers R, Chen W, Doron-Faigenboim A, Duke MV, Gong L, Grimwood J, Grover C, Grupp K, Hu G, Lee TH, Li J, Lin L, Liu T, Marler BS, Page JT, Roberts AW, Romanel E, Sanders WS, Szadkowski E, Tan X, Tang H, Xu C, Wang J, Wang Z, Zhang D, Zhang L, Ashrafi H, Bedon F, Bowers JE, Brubaker CL, Chee PW, Das S, Gingle AR, Haigler CH, Harker D, Hoffmann LV, Hovav R, Jones DC,

Lemke C, Mansoor S, ur Rahman M, Rainville LN, Rambani A, Reddy UK, Rong JK, Saranga Y, Scheffler BE, Scheffler JA, Stelly DM, Triplett BA, Van Deynze A, Vaslin MF, Waghmare VN, Walford SA, Wright RJ, Zaki EA, Zhang T, Dennis ES, Mayer KF, Peterson DG, Rokhsar DS, Wang X and J. Schmutz (2012). "Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres." Nature **492**(7429): 423–427.

15. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, Marcais G, Pop M, Yorke JA. "GAGE: A critical evaluation of genome assemblies and assembly algorithms.". Genome Res. 2012;22(3):557–67.
16. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot E. "HiC-Pro: an optimized and flexible pipeline for Hi-C data processing.". Genome Biol. 2015;16:259.
17. van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES. (2010). "Hi-C: a method to study the three-dimensional architecture of genomes." J Vis Exp(39).
18. Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S, Zou C, Li Q, Yuan Y, Lu C, Wei H, Gou C, Zheng Z, Yin Y, Zhang X, Liu K, Wang B, Song C, Shi N, Kohel RJ, Percy RG, Yu JZ, Zhu Y-X, Wang J, Yu S. "The draft genome of a diploid cotton Gossypium raimondii. " nature genetics. 2012;44:7.
19. Xie T, Zheng JF, Liu S, Peng C, Zhou YM, Yang QY, Zhang HY. "De novo plant genome assembly based on chromatin interactions: a case study of Arabidopsis thaliana.". Mol Plant. 2015;8(3):489–92.
20. Yuan D, Tang Z, Wang M, Gao W, Tu L, Jin X, Chen L, He Y, Zhang L, Zhu L, Li Y, Liang Q, Lin Z, Yang X, Liu N, Jin S, Lei Y, Ding Y, Li G, Ruan X, Ruan Y, Zhang X. "The genome sequence of Sea-Island cotton (Gossypium barbadense) provides insights into the allopolyploidization and development of superior spinnable fibres. " Sci Rep. 2015;5:17662.
21. Zhang J, Huang G-Q, Zou D, Yan J-Q, Li Y, Hu S, Li X-B. "The cotton (Gossypium hirsutum) NAC transcription factor (FSN1) as a positive regulator participates in controlling secondary cell wall biosynthesis and modification of fibers.". New Phytol. 2018;217(2):625–40.

Tables

Table 1: The statistics of Hi-C data filtering.

Sample	G. *raimondii*
Read Length (bp)	100
Raw Paired-end Reads	570,412,361
Raw Bases (bp)	114,082,472,200
Clean Paired-end Reads	503,917,093
Clean Paired-end Reads Rate (%)	88.34
Low-quality Paired-end Reads	10,734,587
Low-quality Paired-end Reads Rate (%)	1.88
Ns Paired-end Reads	224,511
Ns Paired-end Reads Rate (%)	0.04
Adapter Polluted Paired-end Reads	55,536,170
Adapter Polluted Paired-end Reads Rate (%)	9.74
Raw Q30 Bases Rate (%)	94.88
Clean Q30 Bases Rate (%)	95.6

Clean paired-end reads are the high-quality reads after filtering. Ns paired-end reads having the more than 5% N's percentage. Q30 bases rate is the ratio of base's sequencing quality which is higher than 30, it means the base's sequencing error percentage is less than 0.1%.

Table 2: The statistics of mapped Hi-C data.

Sample	*G. raimondii*
Clean Paired-end Reads	503,917,093
Unmapped Paired-end Reads	20,995,471
Unmapped Paired-end Reads Rate (%)	4.17
Paired-end Reads with Singleton	127,442,481
Paired-end Reads with Singleton Rate (%)	25.29
Multi Mapped Paired-end Reads	82,998,864
Multi Mapped Ratio (%)	16.47
Unique Mapped Paired-end Reads	272,480,277
Unique Mapped Ratio (%)	54.07

Multi mapped paired-end read means Hi-C sequencing read mapped to more than one loci of reference sequence.

Table 3: The statistics of Hi-C data after filtering.

Sample	G. *raimondii*
Unique Mapped Paired-end Reads	272,480,277
Dangling End Paired-end Reads	1,151,621
Dangling End Rate (%)	0.42
Self-circle Paired-end Reads	275,606
Self-circle Rate (%)	0.10
Dumped Paired-end Reads	5,714,267
Dumped Rate (%)	2.10
Interaction Paired-end Reads	265,338,783
Interaction Rate (%)	97.38
Valid Paired-end Reads	223,304,666
Valid Rate (%)	81.95

Dangling end paired-end read means the biotin labeled base is at the end of read. Dumped paired-end reads do not contain any biotin labeled base or it's inter size is out of range. Interaction paired-end reads were mapped to different restriction fragment. Valid paired-end reads are interaction paired-end which taken out repeat paired-end reads caused by PCR.

Table 4: The statistics of clustering results.

Sample	G. *raimondii*
Number of Sequence in Draft Genome	4,974
Length of Sequence in Draft Genome (bp)	758,633,485
Number of Sequence in Clustering	2,883
Rate of Numbers in Clustering (%)	57.96
Length of Sequence in Clustering (bp)	746,659,745
Rate of Length in Clustering (%)	98.42

Table 5: The statistics of ordering results.

Sample	G. *raimondii*
Number of Sequence in Ordering	1,328
Rate of Number in Ordering(%)	46.06
Length of Sequence in Ordering	744,578,885
Rate of Length in Ordering(%)	99.72
Number of Sequence in Trunks	697
Rate of Number in Trunks(%)	52.48
Length of Sequence in Trunks	724,960,878
Rate of Length in Trunks(%)	97.37

Table 6: The statistics of orienting results.

Sample	G. *raimondii*
Number of Sequence in Orienting	1129
Rate of Numbers in Orienting (%)	85.02
Length of Sequence in Orienting (bp)	743,948,690
Rate of Length in Orienting (%)	99.92

Table 7: The statistics of pseudo-chromosome length.

Pseudo-chromosome	Scaffolds number	Length (bp)
chr1	171	73,966,406
chr2	98	65,258,067
chr3	93	63,950,047
chr4	94	62,276,720
chr5	117	60,231,642
chr6	76	60,016,944
chr7	88	59,096,273
chr8	107	58,208,991
chr9	69	56,972,541
chr10	125	53,706,612
chr11	90	51,080,163
chr12	99	49,533,500
chr13	101	30,412,479
Total anchored	1,328	744,710,385
Unanchored	3,646	14,054,600

We named the reassemble chromosome by its length order. The length of chromosome contained the 100 bp N between neighboring scaffolds.

Table 8: The statistics of the draft genome.

Parameters	Contigs length (bp)	Scaffolds length (bp)	Contigs number	Scaffolds number
Total	716,234,346	756,905,237	37,849	2,582
Max_length	333,622	10,920,000	-	-
Number>=2000bp	-	-	27,412	1,626
N50	44,885	1,600,000	4,810	139
N60	35,741	1,216,543	6,597	194
N70	27,678	947,408	8,872	264
N80	19,881	721,270	11,905	355
N90	11,331	439,650	16,558	485

All of the statistics got rid of the short scaffolds (scaffold length < 1000bp).

Table 9: The statistics of the reassembled *G. raimondii* genome.

Parameters	Contigs length (bp)	Scaffolds length (bp)	Contigs number	Scaffolds number
Total	716,283,110	757,086,669	37,921	1,329
Max_length	333,622	73,966,406	-	-
Number>=2000bp	-	-	27,412	462
N50	44,885	60,016,944	4,810	6
N60	35,740	58,208,991	6,598	8
N70	27,675	56,972,541	8,874	9
N80	19,873	53,706,612	11,907	10
N90	11,327	49,533,500	16,562	12

All of the statistics got rid of the short scaffolds (scaffold length < 1000bp).

Figures

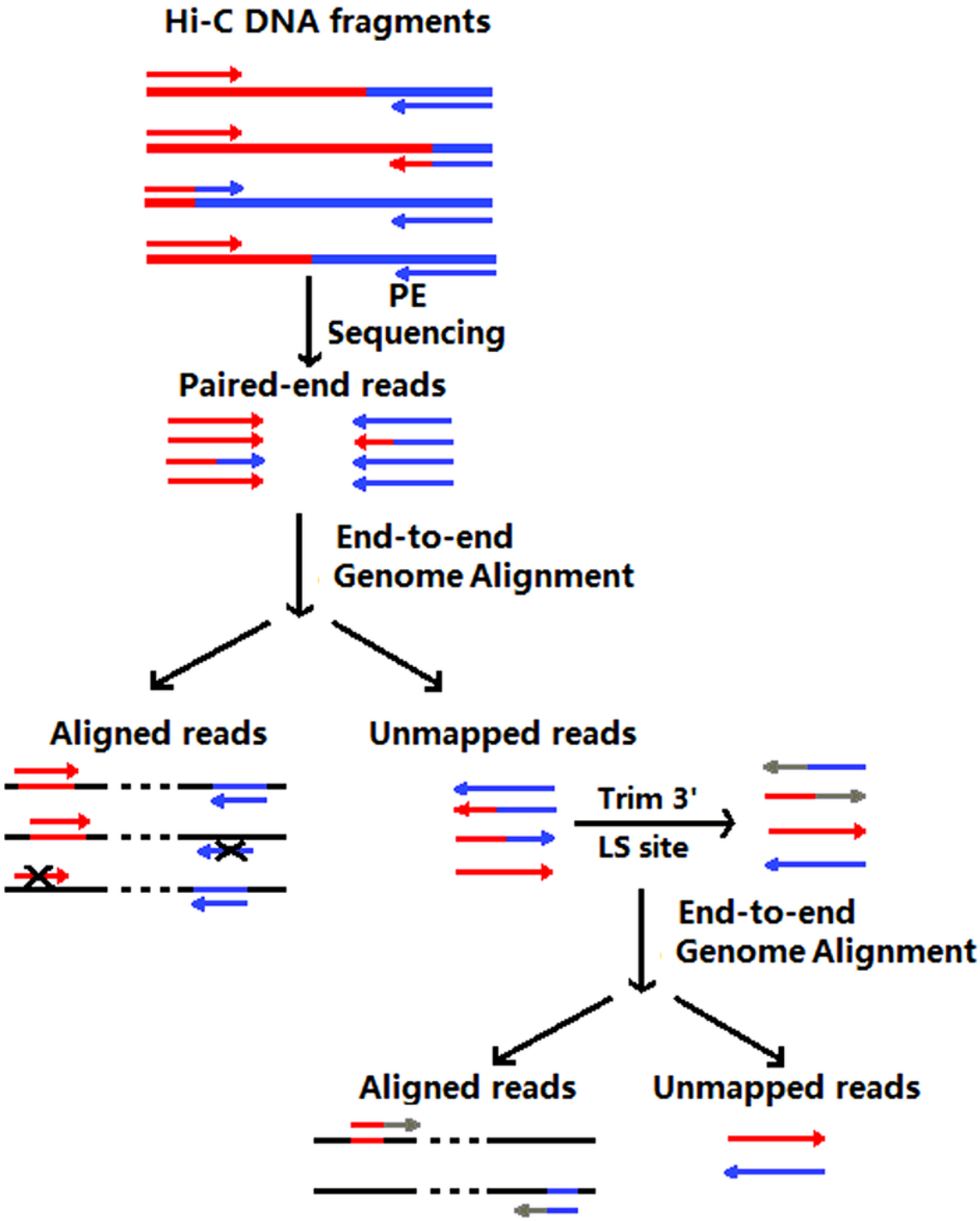

Figure 1

The pipeline of bowtie2. First, Reads1 and Reads2 are aligned to reference genome and then we cut the tails from 3' to the ligation-site and realigned it to reference genome. At last we merged the two results as the final mapped results.

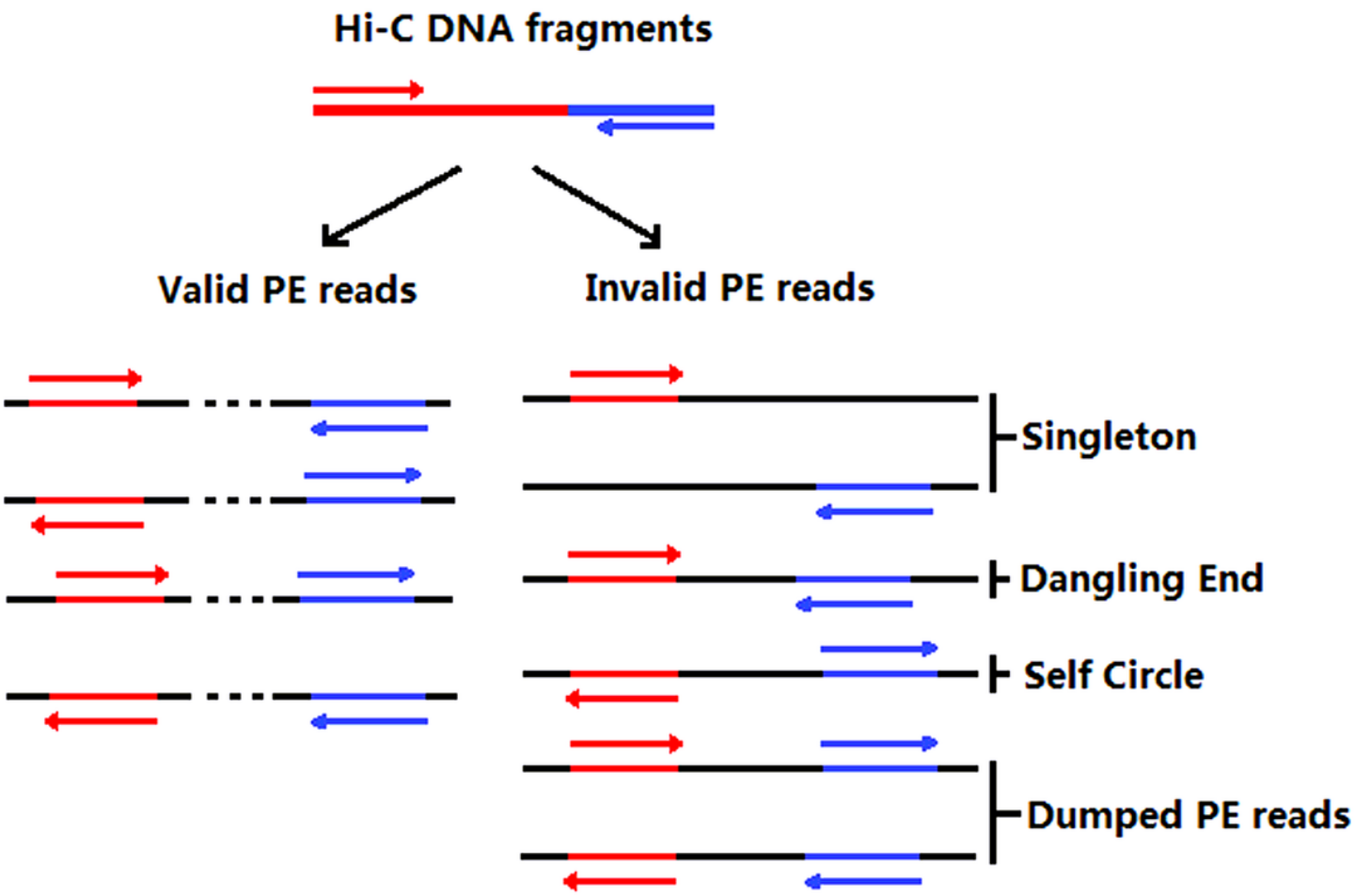

Figure 2

The corresponding situations of different alignment orientations.

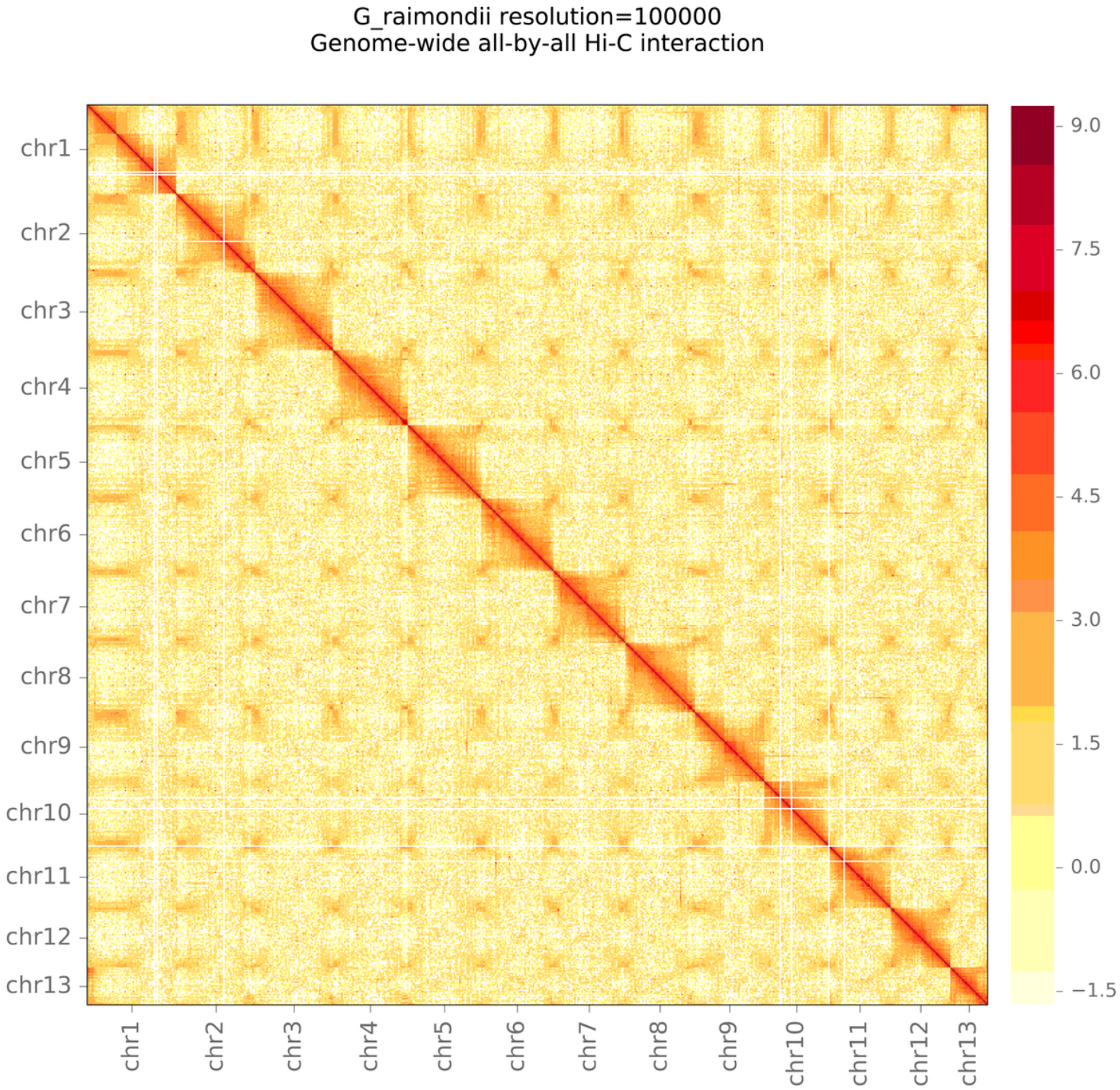

Figure 3

The heat-map of the reassembled result at a 100-kb resolution. The darker red dot is, representing the higher interaction between bins

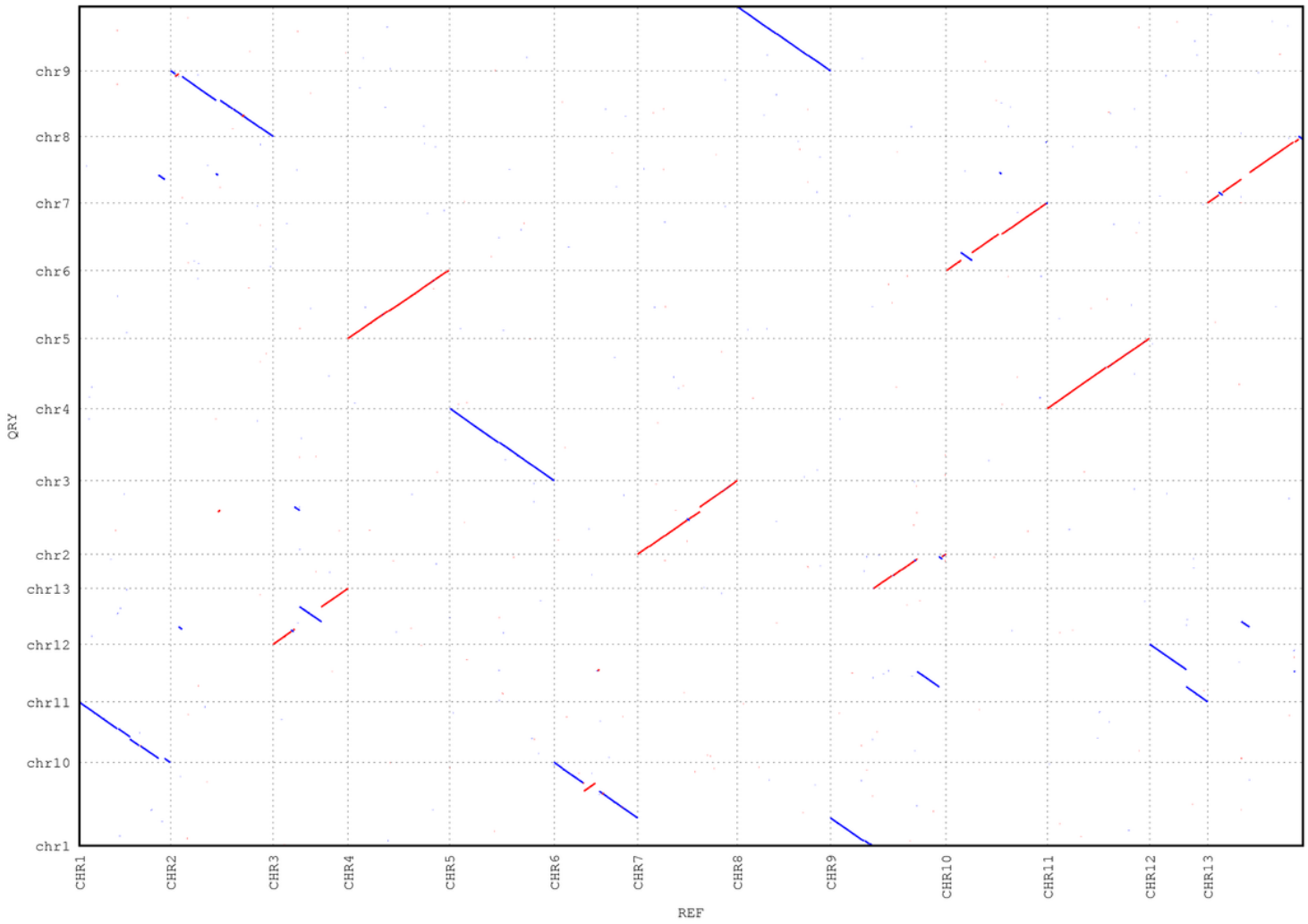

Figure 4

The collinearity analysis between reassemble (ORY) and draft genome (REF).

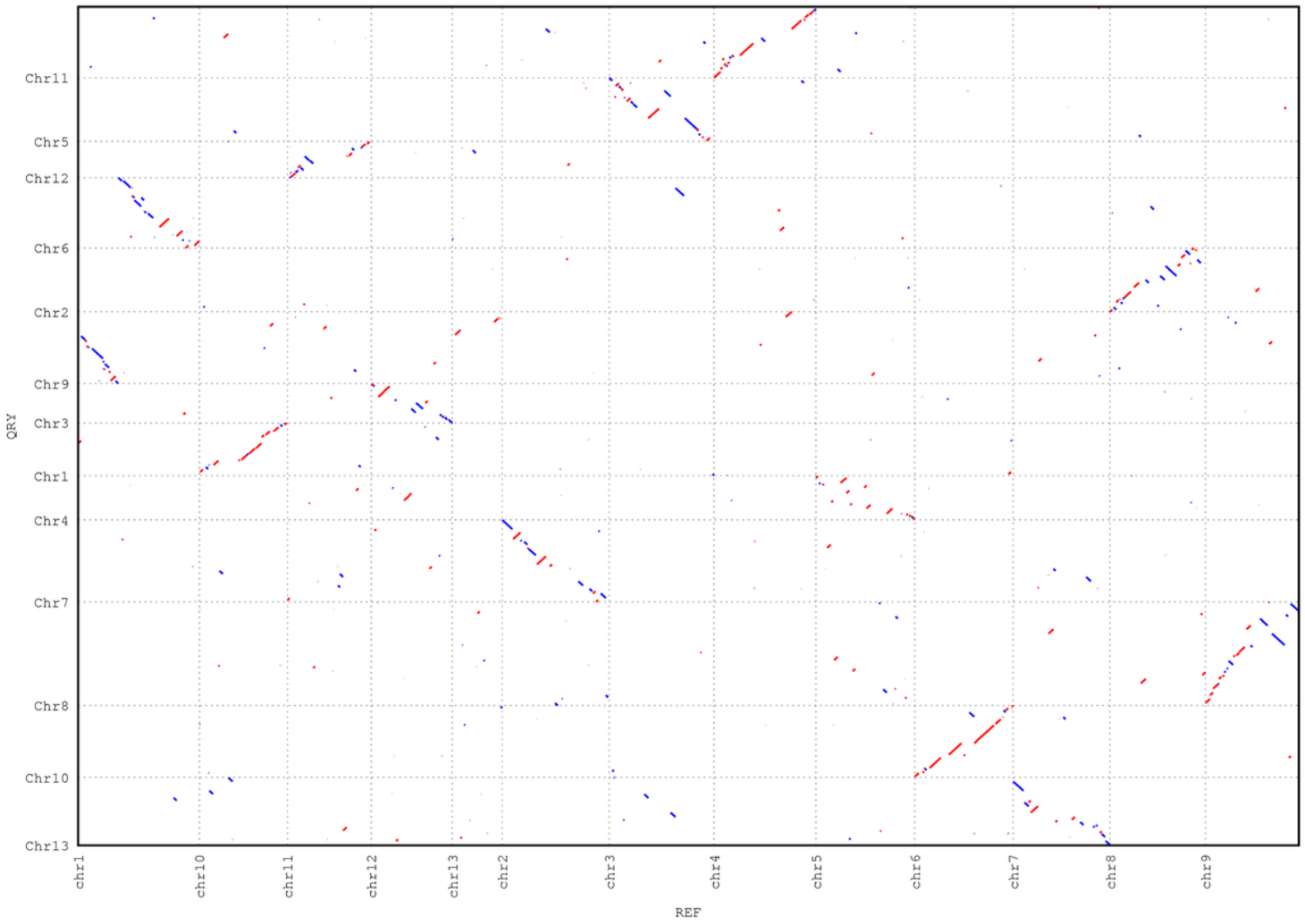

Figure 5

The collinearity analysis between the reassemble genome (ORY) and draft genome from Paterson et al. (2012) (REF).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- errorbin.bed.xls