Plant Communications
Correspondence

CellPress
Partner Journal

CentIER: Accurate centromere identification for plant genomes

Dear Editor,

Centromeres, the basis for cell division, offer essential insights into cell dynamics, genome stability, and evolutionary processes (McKinley and Cheeseman, 2016). Because of ultra-high complexity, high-quality sequences of centromeric regions have long been difficult to obtain, hindering studies of centromere function, evolution, and variation. In recent years, advances in sequencing technology have solved the problem of centromere assembly to a large extent, and dozens of telomere-to-telomere-level reference genomes with complete centromeric sequences have been assembled (Li et al., 2024; Liao et al., 2024). However, there have been few studies on centromere detection in telomere-to-telomere assemblies using only computational methods, limiting a larger scale and broader range of centromere analysis.

In this paper, we first evaluate and show the disadvantages of existing centromere detection methods and then introduce a novel software, CentIER, which is the first bioinformatic tool designed to detect complete centromeric regions (including both repetitive and non-repetitive regions) without additional wet experiments. Finally, we assess the accuracy of CentIER using diverse plant genomes, including *Arabidopsis* (dicot, small genome), maize (monocot, large genome), and mulberry (metapolycentric chromosomes). The results show that CentIER can perform significantly more complete and accurate detection than the existing tool.

DISADVANTAGES OF EXISTING CENTROMERE IDENTIFICATION METHODS

The existing methods for centromere identification can be divided into two categories. The wet experimental method involves chromatin immunoprecipitation sequencing with a centromere-specific protein (CenH3) and identifies centromeric regions through mapping of reads to the genome (Chen et al., 2023). Although this method is accurate, the difficulty of synthesizing CenH3 limits its application. Because of the low conservation of CenH3 sequences, a common CenH3 antibody does not work for all species, and the production of species-specific antibodies requires comparatively high expertise. To demonstrate this issue in plant genomes, we gathered 47 CenH3 protein sequences from 44 plant species and performed a phylogenetic analysis (Supplemental Figure 1). The results showed that the similarity between evolutionarily distant CenH3 sequences (OsCenH3 from *Oryza sativa* and BrCenH3 from *Brassica rapa*) can be as low as 60.2%, highlighting the potential for considerable sequence divergence among CenH3 proteins from different species (Figure 1A).

The bioinformatic method detects centromeres according to the locations of abundant tandem repeat sequences (TRs) on chromosomes (Pei et al., 2023). A key assumption of this method is that TRs occupy the vast majority of centromeric regions. However, we found that TR locations may not always precisely correspond to centromeric regions. First, TRs may appear in pericentromeric regions, leading to a high false positive rate. For instance, the centromere region on chromosome (chr) 8 of *Zea mays* is 49.96–52.22 Mb, whereas the distribution of TRs with over 300 copies appears at 53.19–53.25 Mb (Supplemental Table 1). Second, the centromeric region may contain a substantial portion of non-TRs, leading to a high false negative rate. To demonstrate this, we used the method provided by Shi et al. (2023) to identify centromeres in *A. thaliana*, rice, and maize genomes. The results showed that centromere detection using only TRs achieves a high level of precision but a comparatively low recall rate, suggesting an incomplete identification of centromeres (Figure 1C, TRs). For these reasons, quarTeT, the only existing bioinformatic detection tool, is theoretically able to detect only TRs in centromeres rather than complete centromeres (Lin et al., 2023).

NEW FEATURES FOR CENTROMERE DETECTION

To solve the problem of detecting complete centromeres using only computation, we observed and found three types of new features. First, because centromeres are composed mainly of repetitive sequences (Melters et al., 2013), it is reasonable to assume that the sequence specificity of centromeric regions is lower than that of other chromosomal regions. We defined sequence specificity by counting the number of non-repetitive *k*-mers per unit chromosome length and then used the *A. thaliana* and *O. sativa* (including two varieties, MH63RS3 and ZS97RS3) genomes to test our hypothesis. The results showed that the *k*-mer signals in normal (non-repetitive) regions were over 49 000 and were higher than those of repetitive regions (Figure 1B). Furthermore, compared with other repetitive regions that displayed shorter low-signal intervals (Figure 1B, blue arrow), centromeres were distinguishable as extra-long recessed regions with continuous low signals (Figure 1B, red dashed box). This phenomenon can be observed on most chromosomes of different species (see Supplemental Figure 2 for details). Therefore, low *k*-mer signal intensity can be used for centromere identification.

Second, in addition to TRs, long terminal repeat retrotransposons (LTRs) are also principal constituents of plant centromeres.

Published by the Plant Communications Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

Please cite this article in press as: Xu et al., CentIER: Accurate centromere identification for plant genomes, Plant Communications (2024), https://doi.org/10.1016/j.xplc.2024.101046

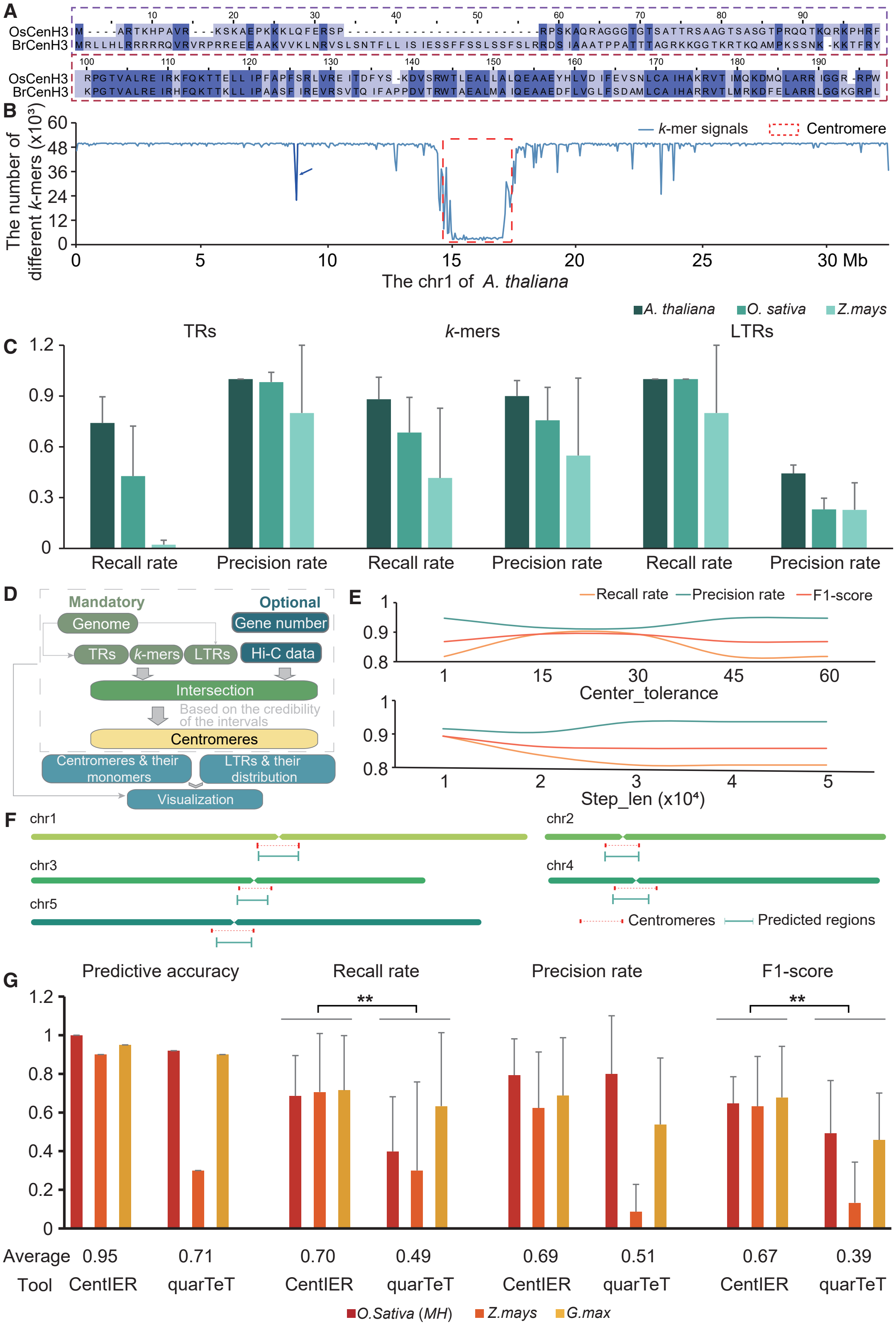

Figure 1. Evaluation of centromere prediction methods and CentIER algorithm performance with test results.

(A) Sequence alignment of OsCenH3 and BrCenH3.

(B) The distribution of *k*-mer type number on *Arabidopsis* chromosome 1.

(C) Evaluation of TR-, *k*-mer-, and LTR-based detection methods using recall and precision rates.

(legend continued on next page)

Therefore, centromeric and non-centromeric regions can be distinguished on the basis of LTR density to a certain extent. Although empirical findings indicate that LTR-enriched regions are inconsistent with centromeric regions identified on the basis of other features (e.g., TR-enriched and LTR-enriched regions on maize chr10 are 51.64261–51.660022 and 41.4–48.9 Mb, respectively; Supplemental Table 1), in some situations, LTRs can be used as a type of ancillary feature for centromere detection.

Third, a Hi-C map can also be used to assist with centromere detection because of the significant difference between centromeric and non-centromeric regions on the map. Specifically, in many situations, centromeric regions appear as large, continuous missing areas in Hi-C maps owing to the high difficulty of read alignment (Supplemental Figure 3). Because they are widely used in genome assembly, Hi-C data are available for most recent sequencing projects and can therefore be used for centromere detection without additional sequencing.

To further study the effects of these features on centromere detection, we performed a comparative evaluation of the methods, each using only one type of feature (detailed data are presented in Supplemental Table 1). The results indicated that centromere identification using TRs had a higher precision rate, whereas the approach based on LTRs had a superior recall rate (Figure 1C). The *k*-mer-frequency-based method achieved a notable balance by simultaneously maintaining high rates of recall and precision. More importantly, the integration of these complementary methodologies enhanced the accuracy of centromere identification. For example, the centromeres on maize chr5 and chr8 were not successfully recognized using TR information; however, they were identified through the LTR-based method (Supplemental Table 1).

CentIER OVERVIEW AND ACCURACY EVALUATION

CentIER is a bioinformatic tool that can comprehensively take advantage of all the above-mentioned features and information such as sequence specificity (*k*-mer frequency), TRs, LTRs, and Hi-C data (optional) to accurately detect complete centromeric regions in high-quality genome assemblies using only computation.

In brief, CentIER detects candidate intervals using *k*-mer frequency, TRs, LTRs, and Hi-C data (optional) separately, then uses a voting-like strategy to decide upon a final contiguous region in which the most area is contained in as many candidate intervals as possible. An adaptive algorithm has been developed to detect candidate intervals using *k*-mer frequency, whereas long TR- and LTR-enriched regions are obtained using existing tools and pipelines, assisted by post-detection filtering steps. Candidate intervals are obtained from Hi-C data by searching for continuous bins in which the signal intensity is significantly lower than the whole-genome average, and a post-processing step has been developed to pinpoint accurate starting and ending positions. The algorithmic details are provided in the supplemental notes. In addition to the detection algorithm, CentIER integrates functions such as querying repeat sequences, annotating and statistically analyzing LTRs, and visualization, thus providing a more comprehensive platform for centromere analysis (Figure 1D).

The performance of CentIER was assessed using the genomes of multiple plants such as *Arabidopsis*, rice, maize, soybean (*Glycine max*), and mulberry, and the experimental results are shown in Figure 1F and 1G and Supplemental Tables 2 and 4. The centromeres of *Arabidopsis* were accurately detected by CentIER (Figure 1F). Furthermore, the results of CentIER were compared with those of quarTeT for the rice, maize, and soybean genomes using criteria such as predictive accuracy, recall rate, precision rate, and F1-score (definitions of these criteria are provided in the supplemental notes). CentIER accurately identified most centromeres of these three genomes, showing an average improvement of 24% in predictive accuracy, 22% in recall, 18% in precision, and 28% in F1-score relative to quarTeT (Figure 1G). In addition, we tested CentIER on a mulberry genome with metapolycentric chromosomes. CentIER accurately detected 35 out of 42 mulberry centromeres (predictive accuracy = 87%) with a recall rate of 60%, precision rate of 72%, and F1-score of 65%. To study the effect of key parameters on the performance of CentIER, we compared the results of *Arabidopsis* centromere detection using different settings of “step_len” and “center_tolerance,” which are two tunable parameters in the sequence-specificity-based detection module (Supplemental Table 3; Figure 1E). These two parameters had relatively little effect on predictive accuracy and a slight effect on recall and precision rates, demonstrating the stability of the CentIER algorithm.

DATA AND CODE AVAILABILITY

The source codes and example data can be downloaded from https://github.com/simon19891216/CentIER/releases/tag/CentIERv2.0.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant no. 32100501) and the Innovation Program of Chinese Academy of Agricultural Sciences.

ACKNOWLEDGMENTS

No conflict of interest is declared.

AUTHOR CONTRIBUTIONS

W.P., F.C., Y.X., and D.X. conceived the study and designed the experiments. D.X., J. Yang, and H.W. developed the CentIER algorithm. W.F. completed the data collection and tested the functionality of CentIER. X.Z., X.H., and J. Yue tested and evaluated the functionality of

(D) The workflow of CentIER.
(E) Effect of different parameter settings on CentIER accuracy.
(F) Comparison of CentIER predictions with real *Arabidopsis* centromeric regions.
(G) Comparison of the predictive accuracy, recall rate, precision rate, and F1-score of CentIER with those of quarTeT using the genomes of *O. sativa*, *Z. mays,* and *G. max*.
TRs, tandem repeat sequences; LTRs, long terminal repeat retrotransposons; chr, chromosome. $^{**}P \leq 0.01$.

Please cite this article in press as: Xu et al., CentIER: Accurate centromere identification for plant genomes, Plant Communications (2024), https://doi.org/10.1016/j.xplc.2024.101046

CentIER. D.X. wrote the manuscript. W.P., F.C., and Y.X. revised the manuscript.

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Plant Communications Online*.

Received: January 23, 2024
Revised: July 28, 2024
Accepted: August 5, 2024

Dong Xu[1,2,7], *Jinbao Yang*[1,7], *Huaming Wen*[3,7], *Wenle Feng*[4,7], *Xiaohui Zhang*[1], *Xingqi Hui*[1], *Junyang Yue*[5], *Yun Xu*[3,*], *Fei Chen*[6,*] *and Weihua Pan*[1,*]

[1]Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China
[2]Rubber Research Institute, Chinese Academy of Tropical Agricultural Science, Haikou, Hainan 571101, China
[3]School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui 230026, China
[4]College of Informatics, Huazhong Agricultural University, Wuhan, Hubei 430070, China
[5]School of Horticulture, Anhui Agricultural University, Hefei, Anhui 230026, China
[6]Nan Fan College, Hainan University, Sanya, Hainan 572000, China
[7]These authors contributed equally to this work.
***Correspondence: Yun Xu (xuyun@ustc.edu.cn), Fei Chen (feichen@hainanu.edu.cn), Weihua Pan (panweihua@caas.cn)**
https://doi.org/10.1016/j.xplc.2024.101046

REFERENCES

Chen, J., Wang, Z., Tan, K., Huang, W., Shi, J., Li, T., Hu, J., Wang, K., Wang, C., Xin, B., et al. (2023). A complete telomere-to-telomere assembly of the maize genome. Nat. Genet. **55**:1221–1231.

Li, Q., Qiao, X., Li, L., Gu, C., Yin, H., Qi, K., Xie, Z., Yang, S., Zhao, Q., Wang, Z., et al. (2024). Haplotype-resolved T2T genome assemblies and pangenome graph of pear reveal diverse patterns of allele-specific expression and genomic basis of fruit quality traits. Plant Commun. 101000.

Liao, Z., Zhang, T., Lei, W., Wang, Y., Yu, J., Wang, Y., Chai, K., Wang, G., Zhang, H., and Zhang, X. (2024). A telomere-to-telomere reference genome of ficus (*Ficus hispida*) provides new insights into sex determination. Hortic. Res. **11**:uhad257.

Lin, Y., Ye, C., Li, X., Chen, Q., Wu, Y., Zhang, F., Pan, R., Zhang, S., Chen, S., Wang, X., et al. (2023). quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. Hortic. Res. **10**:uhad127.

McKinley, K.L., and Cheeseman, I.M. (2016). The molecular basis for centromere identity and function. Nat. Rev. Mol. Cell Biol. **17**:16–29.

Melters, D.P., Bradnam, K.R., Young, H.A., Telis, N., May, M.R., Ruby, J.G., Sebra, R., Peluso, P., Eid, J., Rank, D., et al. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. **14**:R10.

Pei, T., Zhu, S., Liao, W., Fang, Y., Liu, J., Kong, Y., Yan, M., Cui, M., and Zhao, Q. (2023). Gap-free genome assembly and CYP450 gene family analysis reveal the biosynthesis of anthocyanins in *Scutellaria baicalensis*. Hortic. Res. **10**:uhad235.

Shi, X., Cao, S., Wang, X., Huang, S., Wang, Y., Liu, Z., Liu, W., Leng, X., Peng, Y., Wang, N., et al. (2023). The complete reference genome for grapevine (*Vitis vinifera* L.) genetics and breeding. Hortic. Res. **10**:uhad061.