frontiers in Genetics

DATA REPORT
published: 19 November 2021
doi: 10.3389/fgene.2021.733188

Chromosome-Scale Genome Assemblies of Two Korean Cucumber Inbred Lines

Kihwan Song[1†], Younhee Shin[2†], Myunghee Jung[2], Sathiyamoorthy Subramaniyam[2], Keun Pyo Lee[3], Eun-A Oh[3], Jin Ho Jeong[3] and Jeong-Gu Kim[3]*

[1]Department of Bioresources Engineering, Sejong University, Seoul, South Korea, [2]Research and Development Center, Insilicogen Inc., Gyeonggi-do, South Korea, [3]Genomics Division, National Institute of Agricultural Sciences, Nongsaengmyeong, Jeonju, South Korea

Keywords: Korean cucumber, genome, kimchi, slicer, Cucumis sativus

OPEN ACCESS

Edited by:
Yiqun Weng,
University of Wisconsin-Madison, United States

Reviewed by:
Jarkko Salojärvi,
Nanyang Technological University, Singapore
Yuhui Wang,
Nanjing Agricultural University, China
Shan Wu,
Boyce Thompson Institute, United States

****Correspondence:***
Jeong-Gu Kim
jkim5aug@korea.kr

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Genomics, a section of the journal Frontiers in Genetics

Received: *30 June 2021*
Accepted: *27 October 2021*
Published: *19 November 2021*

Citation:
Song K, Shin Y, Jung M, Subramaniyam S, Lee KP, Oh E-A, Jeong JH and Kim J-G (2021) Chromosome-Scale Genome Assemblies of Two Korean Cucumber Inbred Lines. Front. Genet. 12:733188. doi: 10.3389/fgene.2021.733188

INTRODUCTION

Practicing traditional food habits and using traditional ingredients are of major importance for maintaining a diet with good nutritional value. South Korea is well known for its fermented foods, particularly *banchan* (fermented side dishes such as kimchi), which are deeply rooted in Korean food culture. Moreover, Korean cuisine has unique characteristics that are widely accepted to provide various health benefits (Kim et al., 2016), with Korean food culture involving high consumption of vegetables due to the characteristics of its long agricultural history. According to the Korean Ministry of Agriculture definition of the standards and fundamentals of Korean food, one major constraint is that only food prepared with ingredients produced or cultivated in Korea can be considered Korean food. For example, kimchi prepared from imported Chinese cabbage cannot be considered Korean food; the same applies to other *banchan*. As part of the process of preserving Korean food culture, we have initiated the development of genetic resources for Korean varieties of cucumber (*Cucumis sativus* var. *sativus* L.), which is widely cultivated in Korea for both fresh and processed consumption. Cucumber originated in India and spread to other parts of the world through adaptation to various environmental factors and indigenous food habits (Sebastian et al., 2010). This process has led it to become the sixth most widely cultivated vegetable crop in the world, with 2.1 million hectares under cultivation (FAOSTAT, 2020). South Korea is the 16th largest producer of cucumber in the world, with three major cultivar groups being grown: the Baekdadagi-type, Nakhap-type, and Gasi-type cultivars (Park et al., 2021). In this study, we aimed to obtain detailed insights into the genetics of cucumber varieties by constructing chromosome-scale genome assemblies for two Korean cucumber inbred lines: JEF (semi-white Baekdadagi-type, mainly used for kimchi and other fermented foods) and KWS (Korean solid green, Nakhap-type, a slicer used fresh for salads and *gimbap* or Korean cold noodles).

As shown by previous studies of model plants and crops, a single reference genome is inadequate to capture the variation among different genetic lineages. For example, significant structural variation among maize inbred lines has been identified through analysis of multiple genomes (Tao et al., 2019). Furthermore, the cost of assembling multiple genomes has been significantly reduced by third-generation sequencing technologies and computational methods, leading to the construction of chromosome-scale genome assemblies for various crops with the aim of obtaining detailed insights into gene–trait associations (Yang et al., 2019). The first version of the cucumber draft genome was released in 2009 for inbred line 9,930, a lineage of the 'Chinese Long' cultivar (Huang et al., 2009); the genome has since been updated to version 3 (Li et al., 2019) and the chromosomal level Northern American cucumber genome published in 2012 (Yang et al., 2012). Further insight into variations among and within varieties has recently been provided by the

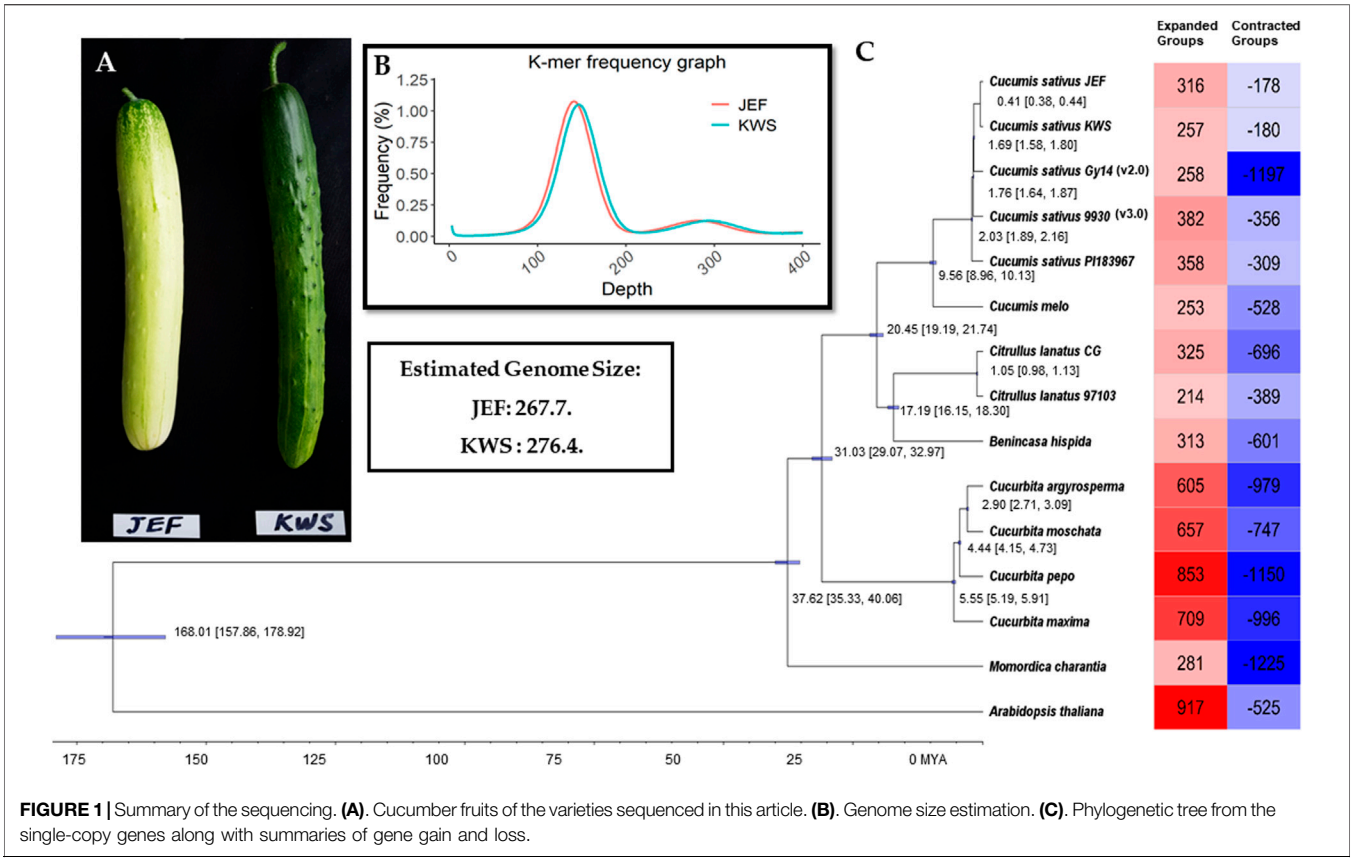

FIGURE 1 | Summary of the sequencing. **(A)**. Cucumber fruits of the varieties sequenced in this article. **(B)**. Genome size estimation. **(C)**. Phylogenetic tree from the single-copy genes along with summaries of gene gain and loss.

publication of information on the genome of the pickling cucumber "Borszczagowski" (line B10) (Osipowski et al., 2020). As the chromosome-scale haploid genome assembly of "Chinese Long" line 9,930 (2n = 2x = 14, haploid number 7) is readily available to the public, we used it as our reference for the construction of chromosome-scale assemblies for the two Korean highly inbred lines.

VALUE OF THE DATA

These new genomes will serve as an additional genetic resource that can be used as a basis and reference for more detailed study into genetic variation and domestication history among Korean cucumber varieties. In addition, they may be valuable for conducting comparative analysis among and within the species in the genus *Cucumis*, which could improve the genome selection process in molecular-assisted breeding.

MATERIALS AND METHODS

Sample Collection and Genomic DNA Extraction

The inbreed lines (i.e., JEF and KWS) are obtained from the leading varieties "Joeun Baekdadagi" and "Gyeoulsal-i Cheongjang" from Fomer Heungnong Seeds Co. After selecting the individual that best characteristics represent of each group in the F_2 populations, two inbreeds were raised through self-fertilization. The resulted breed line i.e., JEF is gynoecious, which is semi-white fruit skin color with white spine and KWS is monoecious which is uniform dark green skin color with black spine (**Figure 1A**). The *Cucumis sativus* breeding line plants were directly harvested in June 2018 in a field in Wanju, Jeollabuk-do, South Korea (35°90′ N, 127°15′ E), near the National Institute of Agricultural Sciences. Sampled fruits are shown in **Figure 1A**, and the complete work flow followed in this study is given in **Supplementary Figure S1**.

DNA Sequencing and *de novo* Genome Assembly

Total DNA was isolated from the samples individually according to sequencing protocols. The isolated DNA was sequenced using two different sequencing systems, PacBio Sequel (Pacific Biosciences, Menlo Park, CA, United States) and Illumina HiSeq 2,500 (Illumina, San Diego, CA, United States), which are widely used in long- and short-read sequencing. For Illumina sequencing, DNA was prepared using the TruSeq Nano DNA Library Prep Kit (Illumina). For PacBio sequencing, DNA was prepared using the SMRTbell Express Template Prep Kit (Pacific Biosciences; catalog no. 101–357–000). The experimental procedures were fully conducted by DNA Link (Seoul, Korea), an authorized service provider in South Korea. The Illumina

paired-end sequences were initially subjected to filtering of technical artifacts (i.e., base-calling errors [Phred quality score ≤ Q20]) and adapters using Trimmomatic v. 0.32 (Bolger et al., 2014). These Illumina reads were used for error correction of PacBio reads in CLC Assembly Cell v. 5.1.1 (Qiagen, Hilden, Germany). The corrected PacBio reads were used to prepare the initial draft version of the cucumber genomes in FALCON-Unzip v. 0.30, a haplotype assembler program (Chin et al., 2016). Finally, using the RaGOO method (Alonge et al., 2019), the genome contigs were clustered and reordered according to their alignment with chromosomal units in the reference genome ('Chinese Long' 9,930). The assembled genomes were assessed for completeness using BUSCO v. 4.1.4 with the Viridiplantae_odb10 reference dataset (Seppey et al., 2019).

Reference Mapping of Bacterial and Organelle Genes

To prepare a clean reference genome, it was necessary to remove bacterial contamination and organelle genomes from the database. The complete GenBank database, which contains draft and reference genomes of bacteria and organelles (mitochondria and plastids), was used as the reference to determine which reads should be removed from the raw sequences. All reference mapping of preprocessed reads was conducted using Bowtie 2 v. 2.2.8 (Langmead and Salzberg, 2012). Details regarding reference paths and sizes are given in **Supplementary Table S1**, and mapping statistics are given in **Supplementary Table S2**.

Genome Size Estimation

All the Illumina-preprocessed sequences from the paired-end library were subjected to genome size estimation based on *k*-mers. The *k*-mer frequencies (*k*-mer size = 17) were obtained using Jellyfish v. 2.0 (Marçais and Kingsford, 2011), and the genome size was calculated from the following formulas: genome coverage depth = (*k*-mer coverage depth × average read length)/(average read length–*k*-mer size +1); genome size = total base number/genome coverage depth. Here, the *k*-mer coverage depth is the major peak of the *k*-mer distribution.

Prediction and Classification of Repeat Regions

Repeat regions in the cucumber genomes were predicted using RepeatModeler (www.repeatmasker.org/RepeatModeler/) and classified into subclasses using the repbase v. 20.08 reference database (www.girinst.org/repbase/) (Bao et al., 2015). Finally, the repeats were masked in the genome using RepeatMasker v. 4.0.5 (www.repeatmasker.org) with RMBlastn v. 2.2.27+. The results are shown in **Supplementary Figure S3**.

RNA Sequencing

The mRNA library from the collected samples was prepared according to the TruSeq Stranded mRNA Prep Kit protocol (Illumina). The isolated mRNA was sequenced using the Illumina sequencer (**Supplementary Tables S4** and **Supplementary Tables S5**).

Gene Prediction and Annotation

The genes from the cucumber draft genomes were predicted using an in-house gene prediction tool that includes three modules: an evidence-based gene modeler (EVM), an *ab-initio* gene modeler, and a consensus gene modeler. The Illumina-sequenced transcriptomes were mapped to the respective repeat-masked draft genomes using TopHat, and Trinity v2.5.1 method was used to assemble the transcripts and mark gene structural boundaries (Trapnell et al., 2012). The *ab-initio* gene modeler and EVM, which included Exonerate (Slater and Birney, 2005), Geneid and AUGUSTUS (Stanke et al., 2006), were trained with several genomes. The final gene and transcript models were optimized using a consensus gene modeler and annotated using Trinotate v. 3.0.1 (Bryant et al., 2017).

Comparative Genome Analysis

Total proteins from the two cucumber genomes were subjected to ortholog analysis to provide insight into the differences between cucumber proteins and those of other plants. In total, 14 genomes from Cucurbitaceae (including the two assembled in this study) were used in the ortholog analysis, with Brassicaceae as outliers (**Figure 1D** and **Supplementary Table S3**). The complete proteins of the selected genomes were also subjected to ortholog analysis using OrthoMCL (Li et al., 2003). The single-copy genes from the given genomes were subjected to phylogenetic tree reconstruction using BEAST (Bayesian Evolutionary Analysis Sampling Trees) to assess the evolutionary time and the degree of similarity among the given genomes (Suchard et al., 2018). Furthermore, to assess the gain and loss of genes in the given genomes, the proteins were analyzed using CAFE v. 3.1 (Han et al., 2013).

Preliminary Analysis Report

Initially, the sizes of the cucumber genomes were estimated to be 267.7 (JEF) and 276.4 MB (KWS) (**Figure 1B**) based on ~50 GB of short-read sequences (**Table 1A** and **Supplementary Table S4**), but 230.8 MB (JEF) and 231.1 MB (KWS) based on the representative scaffolds assembled from ~30 GB of error-corrected long-read sequences (**Table 1A,B**). The N50s of the assembled genomes were 30.5 MB (JEF) and 31.3 MB (KWS), and 40% of the assembled contigs were covered by repeats, in which the long terminal repeat (LTR) elements dominated, accounting for 36% of contigs (**Supplementary Figure S3**). In total, 25,968 genes were predicted from the JEF genome and 26,011 from KWS, with average sizes of 4,111 and 4,114 bases respectively, and BUSCO scores of 97.88 and 98.35% completeness respectively. (**Table 1C**). Finally, 66.54% of JEF genes and 65.96% of KWS genes had homologous sequences in GenBank, while 60.25% of JEF genes and 59.82% of KWS genes had gene ontology descriptions (**Table 1D**). The two genomes were scaffolded onto the reference "Chinese Long" 9,930 genome using the RaGOO method. Overall, these genome assemblies have ~5 MB of additional bases compared with the reference and

TABLE 1 | Summary of the sequencing to annotation of the cucumber draft genomes along with the reference.

	JEF	KWS	Cucumber_9,930_v3 (GCF_000,004,075.3)
(A) Sequencing			
Short Read	72.7 GB (315.10X)	75.9 GB (328.69X)	
Long Reads	31.7 GB (137.48X)	37.0 GB (160.38X)	
(B) Assembly			
Genome size estimation	267,736,921	276,372,239	—
Total length, bp	230,754,408	231,006,969	226,211,662
Total length/Estimation	86.19%	83.59%	—
No. of contigs	7 Chr +54 unplaced	7 Chr +57 unplaced	7 Chr +77 unplaced
Scaffold N50 (Contig N50)	30,569,742 (7,362,017)	31,270,087 (8,654,608)	31,125,843
N (%)	0.01%	0.01%	0.02%
GC (%)	33.23%	33.25%	32.82%
Repeats (MB)	93.47 (40.50%)	94.41 (40.87%)	
Repeats against references (9,930) 91.00 (39.44%)	91.00 (39.44%)	91.69 (39.69%)	84.02 (37.14%)–in our method
BUSCO	99.06%	98.82%	98.82%
(C) Structural annotations			
No. of genes	25,968	26,011	24,317
Average gene length (bp)	4,111.58	4,114.07	4,068.49
BUSCO (Viridiplantae)	97.88%	98.35%	100.00%
(D) Functional annotations			
No. hits	8,690 (33.46%)	8,853 (34.04%)	
Blast hits	17,278 (66.54%)	17,158 (65.96%)	
Gene Ontology	15,645 (60.25%)	15,561 (59.82%)	
KEGG	3,725 (66.54%)	3,718 (14.29%)	

similar BUSCO completeness scores, indicating that they are of good quality. Additionally, an average of 99% of both DNA and RNA sequences were mapped to the reference assembly as an additional measure to ensure the quality of the new assemblies (**Supplementary Figure S2**). The ortholog analysis revealed genome-specific genes, as well as gain and loss of genes, in the selected cucumber genomes (**Figure 1C** and **Supplementary Figure S4**). In addition, the RNA samples were collected from five different developmental stages, revealing that both genomes contain genes expressed differentially in different organs or at different stages (**Supplementary Figures S5** and **Supplementary Figures S6**).

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

YS, MJ, and SS: genome assembly and annotation. KS and SS: manuscript preparation. KL, E-AO, JJ, and J-GK: sampling and sequencing. KS and J-GK: funding and modeling the study.

FUNDING

This work was supported by the Cooperative Research for Agriculture Science and Technology Development (PJ01343202) of the Rural Development Administration, Republic of Korea.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.733188/full#supplementary-material

REFERENCES

Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F. J., et al. (2019). RaGOO: Fast and Accurate Reference-Guided Scaffolding of Draft Genomes. *Genome Biol.* 20, 224. doi:10.1186/s13059-019-1829-6

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase Update, a Database of Repetitive Elements in Eukaryotic Genomes. *Mobile DNA* 6, 11. doi:10.1186/s13100-015-0041-9

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a Flexible Trimmer for Illumina Sequence Data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170

Bryant, D. M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M. B., Payzin-Dogru, D., et al. (2017). A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. *Cel Rep.* 18, 762–776. doi:10.1016/j.celrep.2016.12.063

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., et al. (2016). Phased Diploid Genome Assembly with Single-Molecule Real-Time Sequencing. *Nat. Methods* 13, 1050–1054. doi:10.1038/nmeth.4035

Han, M. V., Thomas, G. W. C., Lugo-Martinez, J., and Hahn, M. W. (2013). Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3. *Mol. Biol. Evol.* 30, 1987–1997. doi:10.1093/molbev/mst100

Huang, S., et al. (2009). The Genome of the Cucumber, Cucumis Sativus L. *Nat. Genet.* 41, 1275–1281. doi:10.1038/ng.475

Kim, S. H., Kim, M. S., Lee, M. S., Park, Y. S., Lee, H. J., Kang, S. A., et al. (2016). Korean Diet: Characteristics and Historical Background. *J. Ethnic Foods* 3, 26–31. doi:10.1016/j.jef.2016.03.002

Langmead, B., and Salzberg, S. L. (2012). Fast Gapped-Read Alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi:10.1038/nmeth.1923

Li, L., Stoeckert, C. J., Jr., and Roos, D. S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* 13, 2178–2189. doi:10.1101/gr.1224503

Li, Q., Li, H., Huang, W., Xu, Y., Zhou, Q., Wang, S., et al. (2019). A Chromosome-Scale Genome Assembly of Cucumber (Cucumis Sativus L.). *GigaScience* 8. doi:10.1093/gigascience/giz072

Marçais, G., and Kingsford, C. (2011). A Fast, Lock-free Approach for Efficient Parallel Counting of Occurrences of K-Mers. *Bioinformatics* 27, 764–770. doi:10.1093/bioinformatics/btr011

Osipowski, P., Pawełkowicz, M., Wojcieszek, M., Skarzyńska, A., Przybecki, Z., and Pląder, W. (2020). A High-Quality Cucumber Genome Assembly Enhances Computational Comparative Genomics. *Mol. Genet. Genomics* 295, 177–193. doi:10.1007/s00438-019-01614-3

Park, G., Choi, Y., Jung, J. K., Shim, E. J., Kang, M. Y., Sim, S. C., et al. (2021). Genetic Diversity Assessment and Cultivar Identification of Cucumber (Cucumis Sativus L.) Using the Fluidigm Single Nucleotide Polymorphism Assay. *Plants (Basel)* 10. doi:10.3390/plants10020395

Sebastian, P., Schaefer, H., Telford, I. R. H., and Renner, S. S. (2010). Cucumber (*Cucumis Sativus*) and Melon (*C. Melo*) Have Numerous Wild Relatives in Asia and Australia, and the Sister Species of Melon Is from Australia. *Proc. Natl. Acad. Sci.* 107, 14269–14273. doi:10.1073/pnas.1005338107

Seppey, M., Manni, M., and Zdobnov, E. M. (2019). "BUSCO: Assessing Genome Assembly and Annotation Completeness," in *Gene Prediction: Methods and Protocols*. Editor M. Kollmar (New York, NY: Springer New York), 227–245. doi:10.1007/978-1-4939-9173-0_14

Slater, G. S. C., and Birney, E. (2005). Automated Generation of Heuristics for Biological Sequence Comparison. *BMC Bioinformatics* 6, 31. doi:10.1186/1471-2105-6-31

Stanke, M., Schöffmann, O., Morgenstern, B., and Waack, S. (2006). Gene Prediction in Eukaryotes with a Generalized Hidden Markov Model that Uses Hints from External Sources. *BMC Bioinformatics* 7, 62. doi:10.1186/1471-2105-7-62

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian Phylogenetic and Phylodynamic Data Integration Using BEAST 1.10. *Virus. Evol.* 4. doi:10.1093/ve/vey016

Tao, Y., Jordan, D. R., and Mace, E. S. (2019). Crop Genomics Goes beyond a Single Reference Genome. *Trends Plant Sci.* 24, 1072–1074. doi:10.1016/j.tplants.2019.10.001

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562. doi:10.1038/nprot.2012.016

Yang, L., Koo, D. H., Li, Y., Zhang, X., Luan, F., Havey, M. J., et al. (2012). Chromosome Rearrangements during Domestication of Cucumber as Revealed by High-Density Genetic Mapping and Draft Genome Assembly. *Plant J.* 71, 895–906. doi:10.1111/j.1365-313x.2012.05017.x

Yang, N., Liu, J., Gao, Q., Gui, S., Chen, L., Yang, L., et al. (2019). Genome Assembly of a Tropical maize Inbred Line Provides Insights into Structural Variation and Crop Improvement. *Nat. Genet.* 51, 1052–1059. doi:10.1038/s41588-019-0427-6

Conflict of Interest: YS, MJ and SS were employed by Insilicogen Inc.

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.