1 ***De-novo* chromosome level assembly of plant genomes from long read sequence data**

2

3 Priyanka Sharma¹, Ardashir Kharabian Masouleh¹, Bruce Topp¹, Agnelo Furtado¹ and Robert

4 J. Henry^{1,2}

5 ¹Queensland Alliance for Agriculture and Food Innovation, University of Queensland,

6 Brisbane 4072 Australia

7 ²ARC Centre of Excellence for Plant Success in Nature and Agriculture, University of

8 Queensland, Brisbane 4072 Australia

9 Summary

10 Recent advances in the sequencing and assembly of plant genomes have allowed the generation
 11 of genomes with increasing contiguity and sequence accuracy. The chromosome level
 12 assembly of the contigs generated from long read sequencing has involved the use of proximity
 13 analysis (Hi-C) or traditional genetic maps to guide the placement of sequence contigs within
 14 chromosomes. The development of highly accurate long reads by repeated sequencing of
 15 circularized DNA (PacBio HiFi) has greatly increased the size of contigs. We now report the
 16 use of HiFiasm to assemble the genome of *Macadamia janseni*, a genome that has been used
 17 as model to test sequencing and assembly. This achieved almost complete chromosome level
 18 assembly from the sequence data alone without the need for higher level chromosome map
 19 information. Eight of the 14 chromosomes were represented by a single large contig and the
 20 other 6 assembled into 2-4 main contigs. The small number of chromosome breaks appear to
 21 be due to highly repetitive regions of ribosomal genes that cannot be assembled by these
 22 approaches. *De novo* assembly of near complete chromosome level plant genomes now seems
 23 possible using these sequencing and assembly tools. Further targeted strategies might allow
 24 these remaining gaps to be closed.

25 **Keywords:** *de novo* genome assembly, HiFiasm, HiFi reads, genome, assembly, nuclear
 26 genome, plastid genome, mitochondrial genome, nuclear ribosomal RNA

27

28 **Significance statement (of up to two sentences):** *De novo* assembly of near complete
 29 chromosome level plant genomes is now possible using current long read sequencing and
 30 assembly tools.

Introduction

Reference genome sequences are a key resource for plant science. The challenge of producing a complete genome sequence has been greatly reduced by advances in both DNA sequencing (Levy and Myers, 2016; Hon et al., 2020) and sequence assembly tools (Phillippy, 2017; Chen et al., 2017). Final assembly of chromosome level genomes has relied upon evidence other than the sequence data alone such as genetic maps (Fierst, 2015; Yu et al., 2019).

Advancements in the field of sequencing, assembly and scaffolding technologies, along with the rapid increase in the freely available genomic data (<https://www.ncbi.nlm.nih.gov/genbank/statistics/>), has greatly facilitated the development of highly accurate *de novo* assemblers.

Short-read *de novo* assemblers are not efficient in assembling the complex and long repetitive regions of plant genomes, like centromere and telomere (Liao et al., 2019). To address this limitation, long read sequencing technologies, also known as third generation sequencers, have been developed. However, these long reads from PacBio and Oxford Nanopore have been less accurate with an average of 90% accuracy as compared to the 99.9% accuracy of Illumina reads (Amarasinghe et al., 2020; Shendure et al., 2017). Therefore, hybrid assembly pipelines were used to assemble a large number of genomes, to overcome the shortcomings of both the long reads and short reads. This has allowed assembly of larger contigs from complex genomes. However, to achieve chromosome level genome assembly, scaffolding of the contigs is usually required. Analysis of sequence proximity in the chromatin by methods such as Hi-C have made this possible (Dudchenko et al., 2017).

Recent advancements in long read sequencing technology have allowed a single molecule to be sequenced multiple times to produce long high fidelity reads (Pac Bio HiFi) with base level accuracy of 99.9 % (Wenger et al., 2019b).

55 *Macadamia jansanii* has been used to compare methods for the sequencing and assembly of
 56 plant genomes (Murigneux et al., 2020; Sharma et al., 2021a). This genome has size (~ 800Mb)
 57 typical of many plant genomes but with a relatively low heterozygosity (Sharma et al., 2020b).
 58 Assembly of this genome using highly accurate circular consensus reads (CCS) reads (Pac Bio
 59 HiFi) was found to give a more contiguous genome than that obtained with earlier longer
 60 continuous long reads (Pac Bio CLR) (Sharma et al., 2021a). HiFiasm has been used to
 61 successfully assemble *Fragaria* × *ananassa* (garden strawberry), *Rana muscosa* (mountain
 62 yellow-legged frog) and *Sequoia sempervirens* (California redwood) (Cheng et al., 2021).
 63 Recently HiFiasm was reported to allow highly contiguous assembly of plant genomes
 64 (Druiguez et al., 2021). We now report the near complete chromosome level assembly of the
 65 *M. jansanii* genome from HiFi reads with the HiFiasm assembly tool and an analysis of the
 66 assembled genome against a HI-C chromosome level assembly.

Results

HiFiasm assembly

The estimated genome size of *M. jansanii* is 780 Mb (Murigneux et al., 2020). The size of the primary Hifiasm assembly was 826 Mb including 779 contigs (**Table S1**), with the longest contig of 71.9 Mb and an average contig length of 1 Mb. BUSCO analysis showed that the assembly covered 99.6% of universal single copy genes (**Table 1**). The contigs generated in this assembly were characterized in three groups based upon their size, large contigs > 1 Mb; medium size contigs, between 1 Mb and 100 Kb and small contigs < 100 Kb.

Larger size contigs > 1 Mb: There were 30 contigs greater than 1Mb in length. These contigs alone provided a good assembly with an N50 of 46 Mb and a BUSCO score of 99.1% (**Table 1**). Dotplot analysis against the Hi-C assembly showed that of the 9 contigs more than 46 Mb in length, 8 correspond to complete Hi-C pseudomolecules (i.e., each contig corresponds to a single chromosome (chromosome # 1, 4, 5, 6, 10, 11, 13 & 14) (**Figure 2 (A)**). One contig (Ptg000010), corresponded to a large part of the second largest chromosome (chromosome 2). Two other contigs of about 25 Mb and 2.7 Mb covered the other parts of this chromosome (**Table 2 & 3**), (**Figure 2 (B)**). The 14 contigs between 4 Mb and 46 Mb in size, covered the remaining six chromosomes, in combinations of two to four contigs. Five of the contigs between 1 Mb and 4Mb in size, corresponded to nuclear ribosomal RNA sequences, and the other two contigs matched parts of chromosome 2 and 7 (**Table 2 & 3**, **Figure 2 (B)**).

Medium size contigs: There were 64 contigs between 1Mb and 100 Kb in size. These contigs had 0% BUSCO genes (**Table 2**). Only eight contigs in the range between 100 Kb and 824 Kb corresponded to seven Hi-C pseudo-molecules (with an alignment block length of more than 100Kb) (**Figure 1B**, **Table S 2**, & **Figure S2, S3**, **Table S2 and S3**). Out of these eight contigs, five contigs corresponds to the terminal part of the Hi-C pseudo-molecules and three contigs

corresponds to the non-terminal regions of Hi-C chromosomes 3 and 7, marked as red starts in **Figure S2 (A) & (B)**. The majority of the medium size contigs corresponds to ribosomal RNA genes (**Figure 5 B**) and one contig of 183 Kb corresponded to a chloroplast assembly (**Figure 3 (B)**). None of the contigs showed similarity with mitochondrial sequences (**Figure 4 B**).

Smaller contigs: There were 685 contigs between 10 Kb and 100 Kb in size. The majority of these small size contigs from the HiFiasm assembly corresponded to small portions of the chloroplast and mitochondrial genomes. These contigs aligned together covered the complete organelle genomes (**Figure 3 (C), 4 (C)**). However, a few of contigs corresponded to nuclear ribosomal RNA sequences (**Figure 5 (C)**). This contig set also showed 0% BUSCO genes (**Table 1**).

Influence of data volume

HiFiasm assembly from CCS reads from two individual SMRT cells and the combined data is given in (**Table S4**). A HiFiasm assembly generated from the 10 X CCS data produce 4511 contigs with 909 Mb of assembly length and N50 of 0.38 Mb, whereas a larger CCS file with 18 X coverage generated an assembly with less contigs (1058), a shorter assembly length (833 Mb) with an improved N50 of 4.4 Mb (**Table 1**). The 18 X assembly closer to the combined CCS assembly (and Hi-C assembly), than the 10X assembly.

Haploid assembly details are given in **Table 1**. The haploid 1 assembly has a greater number of contigs than the haploid 2 assembly. BUSCO results were similar for the two haploid and primary assemblies as all assemblies were relatively complete.

Comparison with Hi-C assembly

A dotplot analysis of 14 pseudo-molecules of *M. jansinii* Hi-C assembly against the HiFiasm assembly is shown in **Figure 1**. The dotplot of contigs >1Mb in size showed a complete match of 25 contigs (out of total 30) with the 14 Hi-C pseudo-molecules (Sharma et al., 2021b) (**Figure 1(A)**). The remaining five large contigs did not contribute to the genome assembly. They were composed of nuclear ribosomal RNA sequences. Chromosomes 1, 4, 5, 6, 10, 11, 13 and 14 were covered by single contig of the HiFiasm assembly (**Figure 2 (A)**), two chromosomes (Chr 8 and 9) were covered by two contigs and chromosomes 2, 3 and 12 were covered by three contigs and chromosome seven by four contigs (**Figure 2(B) & Table 2 & 3**).

Organelle genome analysis

Dotplot analysis of a 159 Mb full length chloroplast genome against the HiFiasm genome assembly indicated the insertion of small fragments of chloroplast sequences in the nuclear genome assembly (**Figure 3 (A), Figure S1 (A)**), which also align with the Hi-C assembly results (Sharma et al., 2021b) (**Figure S1 (B)**). Secondly, among the middle size contig set, only one contig (ptg0000186) of 183 Mb aligned with the chloroplast genome (**Figure 3 (B)**). Contig ptg0000186 covered the complete chloroplast genome including the two inverted repeat regions of the chloroplast (**Figure S4**). Another HiFiasm middle size contig, ptg000066, also showed some similarity with the chloroplast assembly and also aligned with the terminal end of Hi-C chromosome 14 (**Figure S5**). Analysis of the smaller size contigs showed the majority of these contigs contains some fragments of the chloroplast assembly (**Figure 3 (C)**).

Mitochondrial sequence analysis was performed. The size of the *de-novo* mitochondrial assembly was 351 Kb. Analysis against the HiFiasm assembly, indicated the presence of mitochondrial sequences in the smallest set of contigs. The majority of these contigs cover

small fragments of the mitochondria genome (**Figure 4 (C)**). Whereas, in the larger contig set, (>1 Mb), only a few contigs showed some similarity with mitochondrial sequences. These represent the mitochondria sequences inserted in the nuclear the genome (**Figure 4 (A)**), which aligns with the dotplot result of Hi-C assembly (**Figure S1 (B) b**). The middle size contigs did not show the presence of any mitochondria sequences in the dotplot analysis (**Figure 4 (B)**).

Nuclear ribosomal RNA gene sequences analysis

Dotplot analysis of nuclear ribosomal RNAs sequences showed matches with the majority of the middle size contigs with a small number of contigs from the smaller set of contigs having ribosomal RNAs sequence (**Figure 5 (b) & (c)**).

Analysis of repeat elements

The Hifiasm assemblies were larger than the corresponding Hi-C pseudomolecules (**Table 3**). This is probably because the HiFiasm contigs included a larger proportion of repetitive elements than the corresponding Hi-C pseudomolecules (**Table 4**). The longer chromosome had a generally higher content of repetitive elements suggesting that the presence of these repeat regions explained their greater size.

Discussion

This era of genomics is continuing to advance with improved sequencing technologies and the potential to sequence all the species on earth (Lewin et al., 2018).

Accurate chromosome level genome assembly requires accurate reads, high genome coverage and long read length. This has typically involved the use of very high coverage and data from multiple sequencing platforms along with mapping of Hi-C technologies to achieve chromosome level assemblies. However, the combination of high sequence accuracy in a long read in HiFi reads (99.8% accuracy at around 15 Kb average length) provides the option to assemble a complete genome using a single sequencing technology (Cheng et al., 2021) and with a more readily obtainable genome coverage (Wenger et al., 2019a).

Here, we have combined the benefit of this highly accurate reads along with an improved assembly tool HiFiasm (Cheng et al., 2021). HiFi read genome coverage of 28-40 X, for plant genomes within the range of 700 -1000 Mb size, was sufficient to generate high quality assemblies with Mb contig sizes (Sharma et al., 2021b). The extracted DNA from *Macadamia jansonii* contained some impurities, as might be expected for many samples. Two SMRT cells were required to generate 28 X genome coverage with CCS reads. For some samples this may be possible with one single run providing the required coverage if sufficient DNA purity is achieved reducing the cost of obtaining sufficient sequence. When the two individual CCS runs of 10 X and 18 X, were assembled separately using HiFiasm, the final assembly was very fragmented (N50 of 0.38 Mb and 4.4 Mb, respectively) for *M. jansonii* (Table S1), whereas the combined 28X gave a highly continuous assembly with N50 of 46.1 Mb and 99.6 % BUSCO results. Suggesting if the DNA isolation methods is highly efficient in removing impurities, then a single run with less coverage would be sufficient to assemble the genome. Which implies, lesser the data volume, so do the computational requirements. The higher base-calling accuracy by HiFi improves the assembly accuracy by bypassing many time consuming and

heavy computational requirement steps in the assembly workflow. The *M. jansanii* assembly from HiFiasm using HiFi sequencing data produced a near chromosome level assembly, with eight contigs covering eight complete Hi-C pseudo-molecules and other six chromosomes, being covered by only one to four breaks by total of 17 contigs. The gaps in those six chromosomes were found to be due to the presence of large numbers of ribosomal RNA gene sequences (need to look into Dotplot paf file, with Hi-C and HiFiasm). For a plant with ~800 Mb data, we estimate a high-quality chromosome level assembly could be produced within a week from the plant material, if the DNA extraction step is well established.

This highly continuous *M. jansanii* chromosome level assembly will help in better understanding of the genome of macadamias. All four species of *Macadamia* are listed as threatened under Australian legislation (Mast et al., 2008), but *M. jansanii* is particularly endangered, due to its very low population size (less than 200 plants in the wild) (Shapcott and Powell, 2011). The highly accurate genome assembly will facilitate its conservation and use in breeding. *M. jansanii* has small inedible nuts and (Gross and Weston, 1992). However, due to its small tree size and narrow root spread, it is being tested as a rootstock and in hybrids with the commercial species, *M. integrifolia* (Alam et al., 2018). The HiFiasm assembly (BUSCO 99 %) is much better than the Hi-C assembly (BUSCO 97%) (Sharma et al., 2021b) suggesting the incorporation of some regions missing in the Hi-C assembly.

The initiative to complete the genome assembly of almost all living organisms (Lewin et al., 2018; Koepfli et al., 2015), requires a highly efficient assembly method with sustainable financial, computational and time requirements without compromising on genome accuracy. Continuity and completeness should be taken into consideration (Rhie et al., 2021). Our analysis suggests that HiFiasm assembly with the HiFi reads may require almost no further scaffolding for the plants with similar genome size ~800 Mb. Analysis of the nature of the few remaining regions of the genome that are not assembled in these analyses may allow the

development of targeted strategies to complete these assemblies. Analysis of the sequences at the ends of the contigs formed by HiFiasm assembly of HiFi reads may identify those contigs that have been interrupted by repetitive sequences that cannot be assembled *de novo*. This technology is successfully assembling regions with high levels of the repeat sequences that make up more than 50% of the *M. jansanii* genome (Sharma et al., 2021b). It may be that the very high accuracy of the HiFi reads detects minor variations in repeat sequences that allow their unique assembly and that only perfect repeats that are longer than the HiFi reads create a barrier to assembly. This study suggests that the large ribosomal gene clusters in the genome of plants may be one of the few limitations to complete assembly. This would suggest that manual analysis could be used to infer that contigs terminating in ribosomal gene repeats might be linked to other contigs ending in ribosomal contigs in the genome. However, additional information may be required for the plants with very large and complex genomes. This approach will be useful, for producing plant genomes, generating high quality *de novo* chromosome level assemblies, especially for laboratories with limited financial, technical and computational resources.

Experimental procedures

Sequencing data Short read (Illumina) sequencing data analysed was from (Murigneux et al., 2020) and long read data (PacBio HiFi) was from (Sharma et al., 2021a).

HiFiasm Assembly

The HiFiasm genome assembly (Cheng et al., 2021) was generated using High Performance Computing (HPC) facility at University of Queensland. For assembly, 24 core processing units (CPUs) and 120 Gb of memory was employed. Default settings of the HiFiasm assembler were used to assemble heterozygous genomes with built-in duplication purging parameters. HiFiasm

output directory consists of two haploid (1 & 2), one primary contig and one alternate haplotig GFA graph files. Each haploid and one primary contig GFA file was converted to FASTA format using awk command.

Analysis of Assembly

The primary HiFiasm assembly of *M. jansanii* included 779 contigs that were categorised into three subsets a) contigs less than 1Mb size, b) Contigs less than 1Mb and more than 100Kb size and c) contigs less than 100 Kb size. Along with the main primary and two haploid assemblies and all three sets of primary contig subsets were passed through analysis using QUAST (Gurevich et al., 2013), BUSCO (Simão et al., 2015) and RepeatModeler (Humann et al., 2019).

Comparison with Hi-C assembly

The Hifiasm contigs were compared with the *M. jansanii* 14 pseudo-molecules from the Hi-C assembly (Sharma et al., 2021b) using online interactive dotplot tool (Cabanettes and Klopp, 2018).

Characterisation of organelle genomes content of Hifiasm contigs

A reference mitochondrial genome, chloroplast genome and nuclear ribosomal RNA sequence from this sample were assembled from Illumina raw reads (Murigneux et al., 2020) using the GetOrganelle (Jin et al., 2020) with default parameters. The HiFiasm contigs (779) were compared with the organellar and ribosomal sequences in dotblots.

Accession numbers

The Hifiasm assembly, chloroplast assembly, mitochondria assembly and nuclear ribosomal RNA sequence of *M. jansanii* has been deposited under NCBI bioproject PRJNA694456.

Acknowledgements

This project was funded by the Hort Frontiers Advanced Production Systems Fund as part of the Hort Frontiers strategic partnership initiative developed by Hort Innovation, with co-investment from The University of Queensland, and contributions from the Australian Government. We thank the Research Computing Centre (RCC), University of Queensland for support and providing high performance computing resources.

Short legends for Supporting Information

Figure S1: Comparison of HiFiasm and Hi-C pseudo-molecules against the Chloroplast, mitochondria and nuclear ribosomal RNAs.

Figure S2: (a) Dotplots of Hi-C pseudo-molecules against HiFiasm contigs (Longest contigs >1Mb). (b) Dotplots of Hi-C pseudo-molecules against HiFiasm contigs (Longest and Middle size contigs).

Figure S3: (a) Dotplots of Hi-C pseudo-molecules against HiFiasm contigs (Longest contigs >1Mb). S2 (b) Dotplots of Hi-C pseudo-molecules against HiFiasm contigs (Longest and Middle size contigs).

Figure S4: Chloroplast assembly covered by a single HiFiasm Contig (Ptg0000186|) and small parts by Ptg000066|.

Figure S5: Chloroplast sequence (Ptg0000186| & Ptg000066|) insertions in the Hi-C assembly

Table S1: HiFiasm contigs

271 Table S2: Middle size HiFiasm contigs (< 1Mb and > 100 Kb) which are part of Hi-C pseudo-
272 molecule assembly.

273 Table S3: HiFiasm contigs (biggest contigs & middle size contigs) corresponding to *M. jansinii*
274 Hi-C 14 pseudo-molecules.

275 Table S4: IPA and HiFiasm assembly details

276 **Competing Interests**

277 The authors declare that they have no competing interests.

278 **Author contributions**

279 Contributions of authors were as follows: Designed the study and supervised the project: RJH,
280 AF, AM and BT. Genome assembly and analysis: PS, AM. Data analysis: PS, AF, AM & RJH.
281 Tables and Figures: PS, RJH, and AF. Drafted the manuscript: PS and RJH. Data deposition:
282 PS. All authors edited and approved the final manuscript.

283

Reference

- Alam, M. M., Wilkie, J. & Topp, B. L. Early growth and graft success in macadamia seedling and cutting rootstocks. 2018. *International Society for Horticultural Science (ISHS), Acta Horticulturae*, pp 637-644.
- Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E. & Gouil, Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, 21(1), pp 30.
- Cabanettes, F. & Klopp, C. 2018. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, 6(e4958-e4958).
- Chen, Q., Lan, C., Zhao, L., Wang, J., Chen, B. & Chen, Y.-P. P. 2017. Recent advances in sequence assembly: principles and applications. *Briefings in Functional Genomics*, 16(6), pp 361-378.
- Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, 18(2), pp 170-175.
- Driguez, P., Bougouffa, S., Carty, K., Putra, A., Jabbari, K., Reddy, M., Soppe, R., Cheung, M.S., Fukasawa, Y. and Ermini L. 2021. LeafGo: Leaf to Genome, a quick workflow to produce high-quality de novo plant genomes using long-read sequencing technology. *Genome Research* 22, 256.
- Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S. & Aiden, A. P. J. S. 2017. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*, 356(6333), pp 92-95.
- Fierst, J. L. 2015. Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. *Frontiers in Genetics*, 6(220).
- Gross, C. & Weston, P. H. J. A. S. B. 1992. Macadamia janseni (Proteaceae), a new species from central Queensland. *Australian Systematic Botany*, 5(6), pp 725-728.
- Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), pp 1072-5.
- Hon, T., Mars, K., Young, G., Tsai, Y.-C., Karalius, J. W., Landolin, J. M., Maurer, N., Kudrna, D., Hardigan, M. A., Steiner, C. C., Knapp, S. J., Ware, D., Shapiro, B., Peluso, P. & Rank, D. R. 2020. Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data*, 7(1), pp 399.

- Humann, J. L., Lee, T., Ficklin, S. & Main, D. 2019. Structural and Functional Annotation of Eukaryotic Genomes with GenSAS. *In: Kollmar, M. (ed.) Gene Prediction: Methods and Protocols*. pp 29-51.
- Jin, J.-J., Yu, W.-B., Yang, J.-B., Song, Y., dePamphilis, C. W., Yi, T.-S. & Li, D.-Z. 2020. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology*, 21(1), pp 241.
- Koepfli, K.-P., Paten, B., Genome, K. C. o. S. & O'Brien, S. J. 2015. The Genome 10K Project: a way forward. *Annual review of animal biosciences*, 3(57-111).
- Levy, S. E. & Myers, R. M. 2016. Advancements in Next-Generation Sequencing. *Annual Review of Genomics and Human Genetics*, 17(1), pp 95-115.
- Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., van Sluys, M.-A., Soltis, P. S., Xu, X., Yang, H. & Zhang, G. 2018. Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, 115(17), pp 4325.
- Liao, X., Li, M., Zou, Y., Wu, F.-X., Yi, P. & Wang, J. 2019. Current challenges and solutions of de novo assembly. *Quantitative Biology*, 7(2), pp 90-109.
- Mast, A. R., Willis, C. L., Jones, E. H., Downs, K. M. & Weston, P. H. 2008. A smaller Macadamia from a more vagile tribe: inference of phylogenetic relationships, divergence times, and diaspore evolution in Macadamia and relatives (tribe Macadamieae; Proteaceae). *Am J Bot*, 95(7), pp 843-70.
- Murigneux, V., Rai, S. K., Furtado, A., Bruxner, T. J. C., Tian, W., Harliwong, I., Wei, H., Yang, B., Ye, Q., Anderson, E., Mao, Q., Drmanac, R., Wang, O., Peters, B. A., Xu, M., Wu, P., Topp, B., Coin, L. J. M. & Henry, R. J. 2020. Comparison of long-read methods for sequencing and assembly of a plant genome. *GigaScience*, 9(12), gaa146.
- Phillippy, A. M. 2017. New advances in sequence assembly. *Genome Research*, 27(5), pp xi-xiii.
- Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Functammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., Haase, B., Mountcastle, J., Winkler, S., Paez, S., Howard, J., Vernes, S. C., Lama, T. M.,

- Grutzner, F., Warren, W. C., Balakrishnan, C. N., Burt, D., George, J. M., Biegler, M. T., Iorns, D., Digby, A., Eason, D., Robertson, B., Edwards, T., Wilkinson, M., Turner, G., Meyer, A., Kautt, A. F., Franchini, P., Detrich, H. W., Svoldal, H., Wagner, M., Naylor, G. J. P., Pippel, M., Malinsky, M., Mooney, M., Simbirsky, M., Hannigan, B. T., Pesout, T., Houck, M., Misuraca, A., Kingan, S. B., Hall, R., Kronenberg, Z., Sović, I., Dunn, C., Ning, Z., Hastie, A., Lee, J., Selvaraj, S., Green, R. E., Putnam, N. H., Gut, I., Ghurye, J., Garrison, E., Sims, Y., Collins, J., Pelan, S., Torrance, J., Tracey, A., Wood, J., Dagnew, R. E., Guan, D., London, S. E., Clayton, D. F., Mello, C. V., Friedrich, S. R., Lovell, P. V., Osipova, E., Al-Ajli, F. O., Secomandi, S., Kim, H., Theofanopoulou, C., Hiller, M., Zhou, Y., Harris, R. S., Makova, K. D., Medvedev, P., Hoffman, J., Masterson, P., Clark, K., Martin, F., Howe, K., Flicek, P., Walenz, B. P., Kwak, W., Clawson, H., et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, 592(7856), pp 737-746.
- Shapcott, A. & Powell, M. 2011. Demographic structure, genetic diversity and habitat distribution of the endangered, Australian rainforest tree *Macadamia janseni* help facilitate an introduction program. *Australian Journal of Botany*, 59(3), pp 215-225.
- Sharma, P., Aldossary, O., Alsubaie, B., Al-Mssallem, I., Nath, O., Mitter, N., Alves Margarido, G. R., Topp, B., Murigneux, V., Masouleh, A. K., Furtado, A. & Henry, R. J. 2021a. Improvements in the Sequencing and Assembly of Plant Genomes. *Gigabyte*, 1, 2021.
- Sharma P, Murigneux V, Haimovitz J, Nock CJ, Masouleh AK, Topp B, Furtado A, Henry RJ 2021b. The genome of the endangered *Macadamia janseni* displays little diversity but represents an important genetic resource for plant breeding. *BioRxiv*//2021/459545
- Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A. & Waterston, R. H. 2017. DNA sequencing at 40: past, present and future. *Nature*, 550(7676), pp 345-353.
- Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), pp 3210-3212.
- Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Functammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., Ruan, J., Marschall, T., Sedlazeck, F. J., Zook, J. M., Li, H., Koren, S., Carroll, A., Rank, D. R. & Hunkapiller, M. W. 2019a. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology*, 37(10), pp 1155-1162.
- Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Functammasan, A., Kolesnikov, A. & Olson, N. D. J. N. b. 2019b. Accurate

circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10), pp 1155–1162.

Yu, A., Li, F., Xu, W., Wang, Z., Sun, C., Han, B., Wang, Y., Wang, B., Cheng, X. & Liu, A. 2019. Application of a high-resolution genetic map for chromosome-scale genome assembly and fine QTLs mapping of seed size and weight traits in castor bean. *Scientific Reports*, 9(1), pp 11950.

423 **Table 1: (A) HiFiasm contigs in different size categories. (B) Comparison of Primary and Haploid assemblies generated from HiFiasm genome assembler**
 424 **tool.**

(A) HiFiasm Assembly	Number of contigs	Assembly length	N50	N75	BUSCO
Total Contigs	779	826 Mb	46 Mb	25 Mb	99.60%
Contigs >10 Mb	19	746 Mb	48 Mb	39 Mb	93.90%
Contigs > 42 Mb	23	772 Mb	48 Mb	30 Mb	99.10%
Contigs >1 Mb	30	784 Mb	46 Mb	30 Mb	99.10%
Contigs > 100 Kb	94	805 Mb	46 Mb	27 Mb	99.00%
Between <1 Mb to 100kb	64	20 Mb	0.49 Mb	0.22 Mb	0.20%
Between <100 Kb to 10 Kb	685	22 Mb	0.032 Mb	0.028 Mb	0.00%
(B) HiFiasm Primary and Haploid assemblies' comparison					
Primary assembly	779	827 Mb	46.1 Mb	25 Mb	99.60%
Hap 1_assembly	879	816 Mb	24.4 Mb	8.9 Mb	98.80%
Hap 2_assembly	363	776 Mb	14.3 Mb	5.4 Mb	97.90%
Hap 1 > 1 Mb	96	736 Mb	16.4 Mb	6.8 Mb	96.70%
Hap 2 > 1Mb	72	766 Mb	24.5 Mb	12.3 Mb	98.10%

425
 426
 427
 428
 429
 430
 431

Table 2: Chromosomal location of HiFiasm contigs > 1 Mb

Contig id > 1 Mb	Length in bp	Hi-C pseudo-molecule corresponding HiFiasm contigs
ptg000016l	71935981	Chr 1 + Ribo RNA
ptg000003l	57251071	Chr 6
ptg000017l	57081251	Chr 4
ptg000011l	56513637	Chr 5
ptg000004l	49863231	Chr 10
ptg000012l	48320516	Chr 11
ptg000023l	47997562	Chr 13
ptg000010l	46138073	Chr 2 + Ribo RNA
ptg000008l	46131124	Chr 14
ptg000014l	43049961	Chr 9
ptg000009l	39279660	Chr 3
ptg000002l	29700554	Chr 8
ptg000001l	26771894	Chr 12
ptg000006l	25189511	Chr 2
ptg000007l	23138637	Chr 7 + Ribo RNA
ptg000013l	22539440	Chr 8
ptg000020l	22399594	Chr 7
ptg000052l	20335125	Chr 12
ptg000021l	13354688	Chr 3
ptg000019l	8098418	Chr 7
ptg000022l	6676624	Chr 3
ptg000005l	6127021	Chr 9
ptg000072l	4271045	Chr 12
ptg000018l	2743534	Ribo RNA
ptg000025l	2713795	Part of Chr 2
ptg000062l	1651603	Ribo RNA
ptg000074l	1299006	Ribo RNA
ptg000034l	1171806	Ribo RNA
ptg000036l	1154310	Ribo RNA
ptg000033l	1122141	Part of Chr 7

435 **Table 3: Hifiasm contigs (> 1 Mb) covering each of the Hi-C pseudo-molecules.**

<i>M. jansinii</i> HiC pseudo- molecules (A)	Size of HiC Pseudo- molecules (B)	Hifiasm Contigs Corresponding to HiC scaffolds (C)	Hifiasm contigs length (with explanation) (D)	Hifiasm Combined contigs length (E)	Extra Hifiasm length (Hifiasm contig length - HiC scaffold length) (E-B)
Chr 1	67682215	ptg000016l	71935981	71935981	4253766
Chr 2	63669590	ptg000006l + ptg000025l + ptg000010l	74041379 (= 25189511 + 2713795 + 46138073)	74041379	10371789
Chr 3	58143993	ptg000021l + ptg000009l + ptg000022l	59310972 (= 13354688 + 39279660 + 6676624)	59310972	1166979
Chr 4	56076407	ptg000017l	57081251	57081251	1004844
Chr 5	55220784	ptg000011l	56513637	56513637	1292853
Chr 6	53595462	ptg000003l	57251071	57251071	3655609
Chr 7	52077970	ptg000019l + ptg000020l + ptg000033l + ptg000007l	54758790 (= 8098418 + 22399594 + 1122141 + 23138637)	54758790	2680820
Chr 8	49563658	ptg000013l + ptg000002l	52239994 (= 22539440 + 29700554)	52239994	2676336
Chr 9	49085581	ptg000014l + ptg000005	49176982 (= 43049961 + 6127021)	49176982	91401
Chr 10	48974653	ptg000004l	49863231	49863231	888578
Chr 11	47698009	ptg000012l	48320516	48320516	622507
Chr 12	46713600	ptg000001l + ptg000072l + ptg000052l	51378064 (= 26771894 + 4271045 + 20335125)	51378064	4664464
Chr 13	45610911	ptg000023l	47997562	47997562	2386651
Chr 14	45288529	ptg000008l	46131124	46131124	842595

436

437 **Table 4:** Comparative repetitive elements of Hi-C pseudo-molecules and HiFiasm assembly

<i>M. jansinii</i> pseudo-molecules	Genome assembler	Size of Pseudo-molecules	Total repeats	LINE	LTR	DNA elements	Unclassified	Simple repeats
Chr 1	Hi-C	67682215	62%	4.13%	30.25%	0.52%	26.80%	0.64%
	HiFiasm	71935981	62.19%	3.88%	34.78%	0.58%	22.64%	0.65%
Chr 2	Hi-C	63669590	66%	3.31%	38.21%	1.12%	23.34%	0.86%
	HiFiasm	74041379	68%	2.98%	39.58%	1.22%	23.78%	0.44%
Chr 3	Hi-C	58143993	52.34%	6.15%	20.48%	1.13%	24.18%	0.67%
	HiFiasm	59310972	54.30%	7.93%	20.70%	0.97%	24.09%	0.72%
Chr 4	Hi-C	56076407	55.07%	6.26%	22.79%	0.79%	24.27%	1.13%
	HiFiasm	57081251	57.20%	7.21 %	21.57 %	0.90 %	26.58 %	1.23 %
Chr 5	Hi-C	55220784	53.10%	3.27%	31.43%	0.96%	16.99%	0.78%
	HiFiasm	56513637	54.29 %	3.47 %	31.55 %	0.75 %	17.93 %	0.85 %
Chr 6	Hi-C	53595462	55.10%	8.75%	19.94%	1.18%	25.49%	0.79%
	HiFiasm	57251071	58.78 %	9.47 %	22.60 %	1.43 %	24.01 %	1.44 %
Chr 7	Hi-C	52077970	52.89%	7.04%	21.33%	1.42%	22.38%	0.85%
	HiFiasm	54758790	51.21 %	6.65 %	21.25 %	1.25 %	21.37 %	0.79 %
Chr 8	Hi-C	49563658	43.97%	5.39%	15.47%	0.64%	21.29%	1.25%
	HiFiasm	52239994	41.81%	6.25%	21.97%	1.30%	12.29%	0%
Chr 9	Hi-C	49085581	48.13%	5.39%	18.38%	1.76%	21.97%	0.86%
	HiFiasm	49176982	45.93	5.41%	24.84%	1.65%	14.03%	0%
Chr 10	Hi-C	48974653	48.10%	6.24%	17.30%	0.90%	22.83%	1.02%
	HiFiasm	49863231	44.66%	5.91%	22.85%	2.82%	13.08%	0%
Chr 11	Hi-C	47698009	48.10%	6.24%	17.30%	0.90%	22.83%	1.02%
	HiFiasm	48320516	44.70%	4.01%	26.25%	2.82%	11.62%	0%
Chr 12	Hi-C	46713600	44.35%	4.47%	16.46%	1.39%	21.29%	0.85%
	HiFiasm	51378064	25.64	4.47%	21.17%	2.44%	11.73%	0%
Chr 13	Hi-C	45610911	42.22%	5.52%	13.67%	0.70%	21.14%	1.31%
	HiFiasm	47997562	39.68%	5.42%	18.80%	1.63%	13.76%	0%
Chr 14	Hi-C	45288529	42.53%	5.82%	12.85%	0.96%	21.53%	1.50%
	HiFiasm	46131124	41.10%	5.53%	19.60%	1.94%	13.92%	0%

Figure 1: Dotplot of *M. jansenii* Hi-C genome assembly against HiFiasm contigs. (a) HiFiasm longest contigs (> 1 Mb size). (b) HiFiasm medium size contigs (< 1 Mb and > 100 Kb) and (c) HiFiasm smallest contigs (<100 Kb).

Figure 2: Dotplots of HiFiasm contigs against Hi-C pseudo-molecules. A) Pseudo-molecules which are covered by single HiFiasm contig. B) Pseudo-molecules which are covered by more than one HiFiasm contig.

450

451

452 Figure 2(B): Pseudo-molecules which are covered by more than one HiFiasm contig.

453

454

455 Figure 3: Dotplot of *M. jansinii* chloroplast genome sequence against HiFiasm contigs. (a) HiFiasm longest contigs (> 1 Mb size). (b) HiFiasm medium size
456 contigs (< 1 Mb and > 100 Kb) and (c) HiFiasm smallest contigs (<100 Kb).

457

Figure 4: Dotplot of *M. jansinii* mitochondria genome sequence against three sets of HiFiasm contigs. (a) HiFiasm longest contigs (> 1 Mb size). (b) HiFiasm medium size contigs (< 1 Mb and > 100 Kb) and (c) HiFiasm smallest contigs (<100 Kb).

Figure 5: Dotplot of *M. jansinii* nuclear ribosomal RNA sequence against HiFiasm contigs. (a) HiFiasm longest contigs (> 1 Mb size). (b) HiFiasm medium size contigs (< 1 Mb and > 100 Kb) and (c) HiFiasm smallest contigs (<100 Kb).