Improvements in the Sequencing and Assembly of Plant Genomes

Priyanka Sharma[1], Othman Al-Dossary[1,2], Bader Alsubaie[1,2], Ibrahim Al-Mssallem[2], Onkar Nath[1], Neena Mitter[1], Gabriel Rodrigues Alves Margarido[1,4], Bruce Topp[1], Valentine Murigneux[3], Ardy Kharabian Masouleh[1], Agnelo Furtado[1], Robert J Henry[1,5]

[1]Queensland Alliance for Agriculture and Food Innovation, University of Queensland, Brisbane 4072 Australia

[2]College of Agriculture and Food Sciences, King Faisal University, Al Hofuf, Saudi Arabia

[3]Genome Innovation Hub, University of Queensland Brisbane 4072 Australia

[4]Departamento de Genética, Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, São Paulo 13418-900, Brazil

[5]Centre of Excellence for Plant Success in Nature and Agriculture, University of Queensland, Brisbane 4072 Australia

Abstract

Background Advances in DNA sequencing have reduced the difficulty of sequencing and assembling plant genomes. A range of methods for long read sequencing and assembly have been recently compared and we now extend the earlier study and report a comparison with more recent methods.

Results Updated Oxford Nanopore Technology software supported improved assemblies. The use of more accurate sequences produced by repeated sequencing of the same molecule (PacBio HiFi) resulted in much less fragmented assembly of sequencing reads. The use of more data to give increased genome coverage resulted in longer contigs (higher N50) but reduced the total length of the assemblies and improved genome completeness (BUSCO). The original model species, *Macadamia jansenii*, a basal eudicot, was also compared with the 3 other *Macadamia* species and with avocado (*Persea americana*), a magnoliid, and jojoba (*Simmondsia chinensis*) a core eudicot. In these phylogenetically diverse angiosperms, increasing sequence data volumes also caused a highly linear increase in contig size, decreased assembly length and further improved already high completeness. Differences in genome size and sequence complexity apparently influenced the success of assembly from these different species.

Conclusions Advances in long read sequencing technology have continued to significantly improve the results of sequencing and assembly of plant genomes. However, results were consistently improved by greater genome coverage (using an increased number of reads) with the amount needed to achieve a particular level of assembly being species dependant.

Keywords: long read sequencing, assembly, plant, Pacific Biosciences, HiFi reads, Oxford Nanopore Technology

Background

Recent advances in DNA sequencing technology have facilitated the sequencing and assembly of plant genomes with a rapid growth in reports of high quality chromosome level assemblies [1]. A basal eudicot, *Macadamia jansenii*, was used to compare the range of long read sequencing and assembly technologies available in 2019[2]. The Pac Bio (Sequel), Oxford Nanopore Technology (PromethION) and BGI (single-tube Long Fragment Read) platforms were applied to the analysis of the same sample. Assembly tools were evaluated for these data sets and the contribution of short reads to improving assemblies assessed [2]. Technology improvements had delivered ongoing increases in the length and quality of sequence reads delivered by these platforms. Since the original study, significant further advances have been made with the use of repeated sequencing of the same molecule to greatly increase sequence accuracy for long reads. This technology allows the generation of long reads (10-25kb) with greater than 99.5% accuracy [3]. Comparison of long read technologies demonstrates the pros and cons of different platforms in relation to contiguity, accuracy of sequence and data analysis time [4]. We now update the earlier study to demonstrate the impact of these improvements on genome assemblies. Factors such as the volume of data (bp) used in the assembly were explored for the *Macadamia* genome, related species and other diverse species with similar sized genomes.

Long read versus HiFi assemblies

Comparison of assemblies based upon long reads [5] and circular consensus sequence (CCS) reads from HiFi [3], showed the greater accuracy of the CCS reads resulted in greatly improved assemblies for the *Macadamia jansenii* genome (Table 1).

Table 1 Improvement in long read sequencing (Pac Bio) for *Macadamia jansenii* when using higher accuracy sequencing.

	Long reads*[2]	HIFi
Total data	65.2Gb	28Gb
Contig N50	1.57Mb	4.49Mb
Assembly length	758Mb	738Mb
Number of contigs	762	284
BUSCO	97%	98%

*phased Falcon assembly

The assembly with the high quality HiFi reads was less fragmented with slightly reduced total genome length and improved completeness (BUSCO). The use of around 20Gb of high quality (HiFi) data gave N50 values of 4 Mb and resulted in assemblies with fewer than 300 contigs required to cover the genome. This represents a significant advance over the assemblies possible when this sample was used previously to compare different long read platforms and assembly tools, many of which required long computing times to assemble contigs [2]. The high quality IPA assembly had a run time of 20 h with 120 Gb peak memory on the FlashLite computer cluster. This analysis requirement compares favourably with the results for a large number of earlier assembly tools reported for the same sample[2], but

provides a much higher quality assembly. Assembly of the HiFi data with other recent tools was also compared. Flye resulted in a highly fragment genome of 993 Mb with an N50 of 459 kb, while Hifiasm produced a genome of 827 Mb composed of 779 contigs but with an N50 of 46.1 Mb and a L75 of 14.

Results for other *Macadamia* species

Macadami jansenii is an endangered species and is one of four species in the *Macadamia* genus. Sequences of all four species were obtained using the same HiFi techniques and all gave similar high quality outcomes when assembled (Table 2).

Table 2 Comparison of assemblies of *Macadamia* species*

	M. jansenii	*M. integrifolia*	*M. tetraphylla*	*M. ternifolia*
Contig N50	4.5Mb	5.3Mb	10.0Mb	6.4Mb
Longest Contig	16.6Mb	26.4Mb	32.1Mb	21.2Mb
Assembly Length	738Mb	742Mb	707Mb	716Mb
Number of contigs	284	249	153	211
BUSCO	98%	98%	97%	98%

*Primary assemblies shown. For details of associate assemblies see supplementary Table 1.

Results for other diverse plant species

Methods for sequencing plant genomes need to be applied to genomes with a wide range of sizes and complexities. Macadamia is a basal eudicot. Other diverse flowering plant genomes were sequenced to determine how widely applicable the results of this study would be in

plant genome sequencing. Jojoba (*Simmondsia chinensis*), a core eudicot from the Caryopyllales, and avocado, a magnoliid, were compared. The three diverse genomes were all similar in size (700-1000Mb). Many important plant genomes are in or near this size range [6]. *M. jansenii* is an endangered species with relatively low heterozygosity, avocado has much greater heterozygosity [7] and jojoba has been reported to be a tetraploid[8]. Heterozygosity and polyploidy both complicate assembly [9,10]. The quality of the assemblies was more contiguous (fewer contigs required to cover the genome) or similar (avocado) with less data in each of these cases when HiFi reads were used instead of the earlier continuous long reads (Table 3). The macadamia and jojoba genomes gave N50 values that were larger when using the HiFi (CCS) reads than with long reads (CLR). However, the N50 for the slightly larger genome of avocado was greater when using the long reads compared to that obtained with the HiFi reads. This suggest that the larger genome may have longer repeat regions that limit contig assembly in some parts of the genome with HiFi reads.

Table 3 Long read versus HiFi sequencing of other diverse species

	Long reads		HiFi	
	Jojoba	Avocado	Jojoba	Avocado
Total data	152Gb	159Gb	41.4Gb	44Gb
Contig N50	4.73Mb	6.7Mb	4.89Mb	4.3Mb
Assembly length	1260Mb	787Mb	780Mb	749Mb
Number of contigs	762	308	284	298
BUSCO	99%	99%	98%	98%

Impact of the sequencing coverage on the assemblies

The length of the contigs assembled (Figure1) was directly related to the volume of sequence data used. Analysis of four related Macadamia species gave a similar linear relationship between data volume and contig N50 for input of between 10 and 40 times genome coverage. The size of the contigs assembled showed a similar dependence upon the amount of sequence data (genome coverage) across species with the slope of the relationship varying for different species (Figure 2). The macadamia genomes could be assembled with lower coverage. This may be a function of genome size with their smaller genomes requiring less coverage to

achieve a given N50. The larger genomes may contain a higher proportion of repetitive sequences that are difficult to assemble.

Assemblies based upon more data were slightly shorter in total length (Figure 3). This reduction was probably due to removal of duplicated end sequences as contigs were joined. The high quality of these assemblies was confirmed by BUSCO values of more than 95%. Genome completeness was high in all cases but increased slightly when more data was used in the assembly (Figure 4).

These results were confirmed when applying these methods to sequencing the other phylogenetically diverse plant genomes with slightly larger genomes with greater genome complexity. In each case N50 and completeness increased with data volume and genome size declined.

Impact of the read length on the assemblies

The length of sequence reads was also expected to influence the assembly. Examination of size distribution of the 6 species showed that the length of the sequences varied slightly within the expected range around 15kb for HiFi data. The minor differences in mean read length and numbers of longer reads did not explain the differences in the size of the contigs assembled (Supplementary Figure 1). This suggest that the different amounts of sequence data required to drive assembly to a particular level may be associated more with the complexity of the sequence. The close relationship between sequence volume and N50 for the four *Macadamia* species may reflect the similar sequence complexity of the species in this group. The jojoba and avocado required more sequence data to reach the same level of assembly. The slightly larger genome size of these two species may be enough to explain this difference due to the likely higher proportion of repetitive sequence in the somewhat larger genomes.

Oxford Nanopore Technologies updates

Oxford Nanopore Technologies (ONT) regularly releases updated basecalling software to convert the raw electrical signal into sequence data. We repeated the basecalling of the ONT raw data of *M. jansenii* using different versions of the Guppy basecaller released in March 2019 (v2.3.7), April 2019 (v3.0.3) and June 2020 (v4.0.11). The assembly quality improved as shown by an increase in the assembly contiguity and in the number of complete BUSCOs before any polishing (Table 4). The assembly size decreased from 817 Mb to 798 Mb. Two versions of the Flye assembler were applied to the same basecalled sequence dataset, which resulted in a significant increase in genome contiguity and completeness as well as a reduced genome assembly size.

Table 4 ONT genome assembly statistics of *M. jansenii* using the Flye assembler, the pass reads and different Guppy basecallers versions

Basecaller	Guppy v2.3.7	Guppy v3.0.3		Guppy v4.0.11
Assembler	Flye v2.5	Flye v2.4.2	Flye v2.5	Flye v2.5
Number of reads	1,597,353	1,592,919		1,594,802
Contig N50 (Mb)	1.44	0.94	1.51	1.79
Assembly length (Mb)	817	845	811	798
Number of contigs	2,996	4,242	2,855	2,841
Number of contigs (>10 kb)	2,300	3,275	2,088	1,913
BUSCO complete (%)	66.8	51.4	75.1	79.1

Conclusions

These assemblies represent significant advances over the highly fragmented genomes previously reported for these species [11-14]. Advances in long read sequencing using different platforms provide improving options for plant genome sequencing and assembly [15]. A recent comparison of these methods applied to rice genome sequencing showed strengths and weaknesses of both with greater sequence accuracy in the Pac Bio assemblies and more contiguity in the ONT assemblies [4]. The resulting genome sequences can be evaluated for the best combination of sequence and assembly accuracy [16]. The results presented here show that contig size can be increased by adding more sequence reads to achieve a linear increase in N50. This extra data will result in slightly shorter total assembly lengths and improved completeness of the genomes. The improved methods when combined with higher level assembly tools[17] will support routine, rapid and efficient generation of highly accurate chromosome level genome sequences of plant species[18].

Methods

DNA extraction

All local, national and international guidelines and legislation was observed in obtaining samples for this study. *Macadamia jansenii* DNA was prepared as described earlier[19]. Three other Macadamia species (*M. tetraphylla*, *M. ternifolia* and *M. integrifolia*) and Jojoba (*Simmondsia chinensis*) were also extracted using this method with minor modifications where phenol was excluded from the extraction method[20]. Avocado (*Persea americana*) DNA was isolated by a modified CTAB (cetyl-trimethyl ammonium bromide) DNA extraction protocol [21,22]. Leaf tissue (0.2 g) was ground and added to 15 ml of 2% CTAB buffer, pH 8.0 followed by 15 min incubation at 65 °C. The supernatant after centrifugation at 10 g for 15 min was treated with RNAse A (10ng/µl) and incubated at 37°C for 30 min.

Chloroform: Isoamyl alcohol (24:1) washes were performed followed by precipitation with isopropanol and 70% ethanol washes. The DNA was resuspended in ultrapure DNAse and RNAse free water for sequencing.

DNA sequencing and assembly

Long read sequencing was as previously described[19]. Long reads (CLR) were assembled using Falcon [2] for *M. jansenii* and Canu v 2.0 for the other genomes. HiFi gDNA libraries were prepared using sheared genomic DNA (~15-20 kb) was sequenced on a PacBio Sequel II (software/chemistry v9.0.0) following diffusion loading. Sequence data was processed to generate CCS reads using the default settings of the CCS application (v4.2.0) in SMRT Link (v9.0.0); minimum parameters for passes (3), accuracy (0.99), CCS read length (10) and maximum CCS read length (50000). CCS reads were assembled using the Improved Phased Assembly (IPA) method (PacBio). The IPA assembly tool (https://github.com/PacificBiosciences/pbbioconda/wiki/Improved-Phased-Assembler) uses the HiFi sequencing reads (high-quality consensus reads) and generates phased assembly. This produces a primary contig folder, including the main assembly and an associated contig, containing haplotigs and duplications. For all assemblies, 24 CPUs and 120Gb of memory was employed.

Assessment of completeness

The completeness of genome assemblies was evaluated using benchmarking universal single-copy orthologues (BUSCO) analysis (v4.1.2), using genome mode and lineage Eukaryota_odb10 dataset.

Availability of data

Sequence data from Pac Bio (Sequel), Oxford Nanopore Technology (PromethION) and BGI (single-tube Long Fragment Read) analysis of *M. jansenii* was described by Murigneux et al [2]. BGI, PacBio, ONT, and Illumina sequencing data generated in that study were deposited in the SRA under BioProject PRJNA609013and BioSample SAMN14217788. Accession numbers are as follows: BGI (SRR11191908), PacBio (SRR11191909), ONT PromethION (SRR11191910), ONT MinION (SRR11191911), and Illumina (SRR11191912). Assemblies and other supporting data are available from the *GigaScience* GigaDB repository [25]. Pac Bio HiFi reads described in this paper are deposited as CCS reads under NCBI BioProject ID Macadamia : PRJNA694456; Avocado: PRJNA694184 and Jojoba: PRJNA694450

Additional files

Figure S1:Size distribution of reads sequenced

Competing interests

The authors declare that they have no competing interests.

Funding

The macadamia and avocado components of this project were 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 University of Queensland and contributions from the Australian Government. The research on jojoba was funded by King Faisal University.

Author's contributions

Contributions of authors were as follows. RH, AF, VM, AM, IA Conceptualization; PS, OA, BA, ON, VM, AF Data curation; PS, OA, BA,ON, GM, VM ,AM, AF Formal Analysis; RH, AF, IA, AM Funding acquisition; PS, OA, BA, ON ,NM, VM, AM, AF,RH Investigation; IA, BT, Resources; IA, BT, NM, RH, AM, AF Supervision; RH, PS, ON, AM Writing – original draft; All authors Writing – review & editing

Acknowledgements

The project was supported by the University of Queensland Research Computing Centre (RCC) and the University of Queensland Genome Innovation Hub.

Figure legends

Figure1 Influence of data volume on assembly for Macadamia species

N50 of contigs is plotted against the genome coverage. Genome sizes used to calculate coverage were; *M.integrifolia*, 895Mb [12]; *M janseni*, 780Mb [2]; *M. tetraphylla*, 758Mb [23] and *M. ternifolia*, 758Mb (not known but assumed the same as *M. tetraphylla* due to similar assembly size).

Figure 2 Influence of data volume on assembly for diverse species. N50 of contigs is plotted against the genome coverage. Genome sizes used to calculate coverage were, jojoba, 1003Mb[14]; avocado, 920Mb [24] and as in figure 1 for *Macadamia* species

Figure 3 Decrease in length of total assembly as more genome coverage is used in the assembly

Figure 4 Improvement in genome completeness (BUSCO%) with genome coverage

274

275 References

276

277 1. Michael, T.P.; VanBuren, R. Building near-complete plant genomes. *Current Opinion in Plant*
278 *Biology* **2020**, *54*, 26-33, doi:10.1016/j.pbi.2019.12.009.
279 2. Murigneux, V.; Rai, S.K.; Furtado, A.; Bruxner, T.J.C.; Tian, W.; Harliwong, I.; Wei, H.; Yang, B.;
280 Ye, Q.; Anderson, E., et al. Comparison of long-read methods for sequencing and assembly of
281 a plant genome. *Gigascience* **2020**, *9*, doi:10.1093/gigascience/giaa146.
282 3. Hon, T.; Mars, K.; Young, G.; Tsai, Y.C.; Karalius, J.W.; Landolin, J.M.; Maurer, N.; Kudrna, D.;
283 Hardigan, M.A.; Steiner, C.C., et al. Highly accurate long-read HiFi sequencing data for five
284 complex genomes. *Sci Data* **2020**, *7*, doi:ARTN 399 10.1038/s41597-020-00743-4.

285 4. Lang, D.; Zhang, S.; Ren, P.; Liang, F.; Sun, Z.; Meng, G.; Tan, Y.; Li, X.; Lai, Q.; Han, L., et al.
286 Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads
287 of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. *Gigascience*
288 **2020**, *9*, doi:10.1093/gigascience/giaa123.
289 5. Amarasinghe, S.L.; Su, S.; Dong, X.Y.; Zappia, L.; Ritchie, M.E.; Gouil, Q. Opportunities and
290 challenges in long-read sequencing data analysis. *Genome Biology* **2020**, *21*, doi:ARTN 30
291 10.1186/s13059-020-1935-5.

292 6. Wendel, J.F.; Jackson, S.A.; Meyers, B.C.; Wing, R.A. Evolution of plant genome architecture.
293 *Genome Biology* **2016**, *17*, doi:ARTN 37 10.1186/s13059-016-0908-1.

294 7. Juma, I.; Geleta, M.; Nyomora, A.; Saripella, G.V.; Hovmalm, H.P.; Carlsson, A.S.; Fatih, M.;
295 Ortiz, R. Genetic diversity of avocado from the southern highlands of Tanzania as revealed by
296 microsatellite markers. *Hereditas* **2020**, *157*, 40, doi:10.1186/s41065-020-00150-0.
297 8. Tobe, H.; Yasuda, S.; Oginuma, K. Seed Coat Anatomy, Karyomorphology, and Relationships
298 of Simmondsia (Simmondsiaceae). *Bot Mag Tokyo* **1992**, *105*, 529-538, doi:Doi
299 10.1007/Bf02489427.
300 9. Kyriakidou, M.; Anglin, N.L.; Ellis, D.; Tai, H.H.; Stromvik, M.V. Genome assembly of six
301 polyploid potato genomes. *Sci Data* **2020**, *7*, doi:ARTN 88 10.1038/s41597-020-0428-4.

302 10. Kyriakidou, M.; Tai, H.H.; Anglin, N.L.; Ellis, D.; Stromvik, M.V. Current Strategies of Polyploid
303 Plant Genome Sequence Assembly. *Front Plant Sci* **2018**, *9*, doi:ARTN 1660
304 10.3389/fpls.2018.01660.

305 11. Nock, C.J.; Baten, A.; Barkla, B.J.; Furtado, A.; Henry, R.J.; King, G.J. Genome and
306 transcriptome sequencing characterises the gene space of Macadamia integrifolia
307 (Proteaceae). *Bmc Genomics* **2016**, *17*, doi:ARTN 937 10.1186/s12864-016-3272-3.

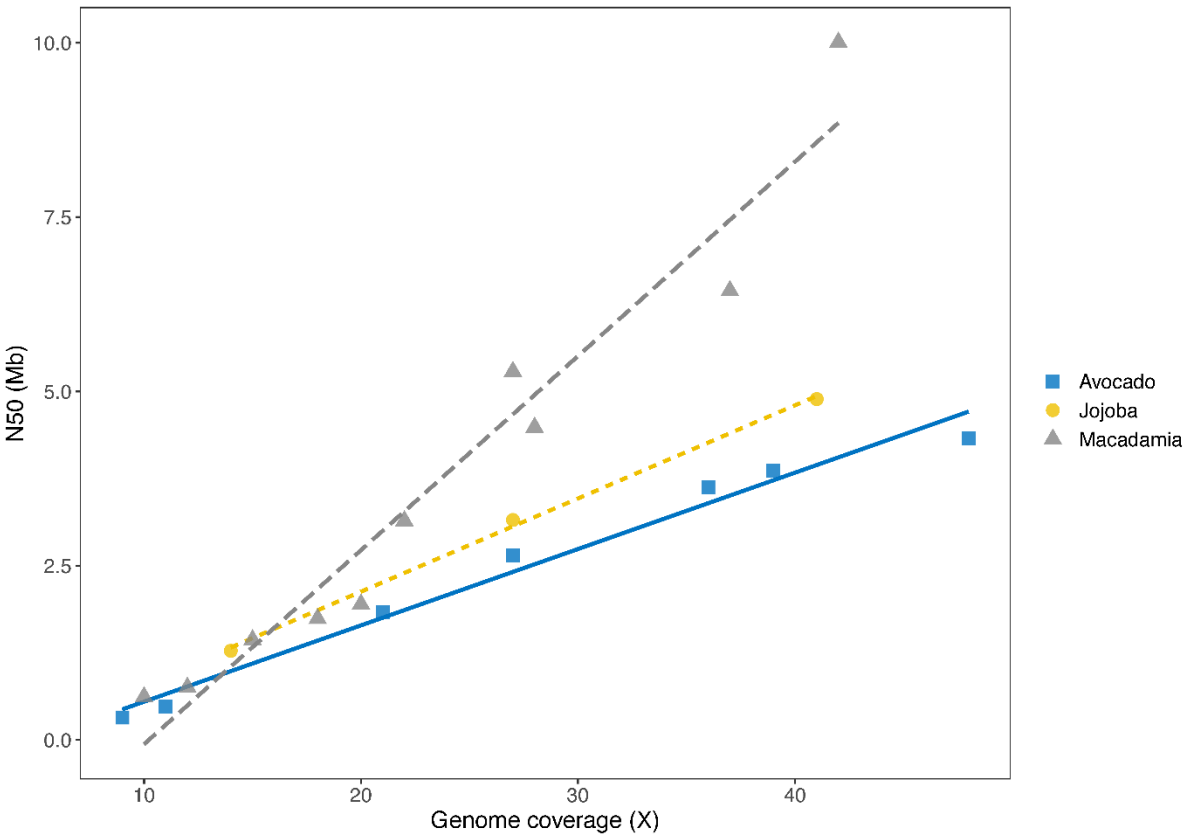
308 12. Nock, C.J.; Baten, A.; Mauleon, R.; Langdon, K.S.; Topp, B.; Hardner, C.; Furtado, A.; Henry,
309 R.J.; King, G.J. Chromosome-Scale Assembly and Annotation of the Macadamia Genome
310 (Macadamia integrifoliaHAES 741). *G3-Genes Genomes Genetics* **2020**, *10*, 3497-3504,
311 doi:10.1534/g3.120.401326.
312 13. Rendon-Anaya, M.; Ibarra-Laclette, E.; Mendez-Bravo, A.; Lan, T.Y.; Zheng, C.F.; Carretero-
313 Paulet, L.; Perez-Torres, C.A.; Chacon-Lopez, A.; Hernandez-Guzman, G.; Chang, T.H., et al.
314 The avocado genome informs deep angiosperm phylogeny, highlights introgressive
315 hybridization, and reveals pathogen-influenced gene space adaptation. *Proceedings of the*
316 *National Academy of Sciences of the United States of America* **2019**, *116*, 17081-17089,
317 doi:10.1073/pnas.1822129116.
318 14. Sturtevant, D.; Lu, S.P.; Zhou, Z.W.; Shen, Y.; Wang, S.; Song, J.M.; Zhong, J.S.; Burks, D.J.;
319 Yang, Z.Q.; Yang, Q.Y., et al. The genome of jojoba (Simmondsia chinensis): A taxonomically

isolated species that directs wax ester accumulation in its seeds. *Sci Adv* **2020**, *6*, doi:ARTN eaay3240 10.1126/sciadv.aay3240.

15. Belser, C.; Istace, B.; Denis, E.; Dubarry, M.; Baurens, F.C.; Falentin, C.; Genete, M.; Berrabah, W.; Chevre, A.M.; Delourme, R., et al. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat Plants* **2018**, *4*, 879-+, doi:10.1038/s41477-018-0289-4.
16. Wang, W.W.; Das, A.; Kainer, D.; Schalamun, M.; Morales-Suarez, A.; Schwessinger, B.; Lanfear, R. The draft nuclear genome assembly of Eucalyptus pauciflora: a pipeline for comparing de novo assemblies. *Gigascience* **2020**, *9*, doi:ARTN giz160 10.1093/gigascience/giz160.
17. Monat, C.; Padmarasu, S.; Lux, T.; Wicker, T.; Gundlach, H.; Himmelbach, A.; Ens, J.; Li, C.D.; Muehlbauer, G.J.; Schulman, A.H., et al. TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. *Genome Biology* **2019**, *20*, doi:ARTN 284 10.1186/s13059-019-1899-5.
18. Li, F.W.; Harkess, A. A guide to sequence your favorite plant genomes. *Appl Plant Sci* **2018**, *6*, doi:ARTN e1030 10.1002/aps3.1030.
19. Cheng, B.; Furtado, A.; Henry, R.J. Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. *Gigascience* **2017**, *6*, doi:10.1093/gigascience/gix086.
20. Furtado, A. DNA Extraction from Vegetative Tissue for Next-Generation Sequencing. *Cereal Genomics: Methods and Protocols* **2014**, *1099*, 1-5, doi:10.1007/978-1-62703-715-0_1.
21. Zou, Y.; Mason, M.G.; Wang, Y.; Wee, E.; Turni, C.; Blackall, P.J.; Trau, M.; Botella, J.R. Nucleic acid purification from plants, animals and microbes in under 30 seconds. *PLoS Biol* **2017**, *15*, e2003916, doi:10.1371/journal.pbio.2003916.
22. Bienvenue, J.M.; Duncalf, N.; Marchiarullo, D.; Ferrance, J.P.; Landers, J.P. Microchip-based cell lysis and DNA extraction from sperm cells for application to forensic analysis. *J Forensic Sci* **2006**, *51*, 266-273, doi:10.1111/j.1556-4029.2006.00054.x.
23. Niu, Y.-F.; Li, G.-H.; Ni, S.-B.; He, X.-Y.; Zheng, C.; Liu, Z.Y.; Gong, L.D.; Kong, G.H.; Liu, J. Genome assembly and annotation of *Macadamia tetraphylla. bioXriv* **2020** doi: https://doi.org/10.1101/2020.03.11.987057
24. Talavera, A.; Soorni, A.; Bombarely, A.; Matas, A.J.; Hormaza, J.I. Genome-Wide SNP discovery and genomic characterization in avocado (Persea americana Mill.). *Scientific Reports* **2019**, *9*, doi:ARTN 20137 10.1038/s41598-019-56526-4.
25. Murigneux V, Rai SK, Furtado A, et al. Supporting data for "Comparison of long-read methods for sequencing and assembly of a plant genome." GigaScience Database 2020. http://dx.doi.org/10.5524/100812.

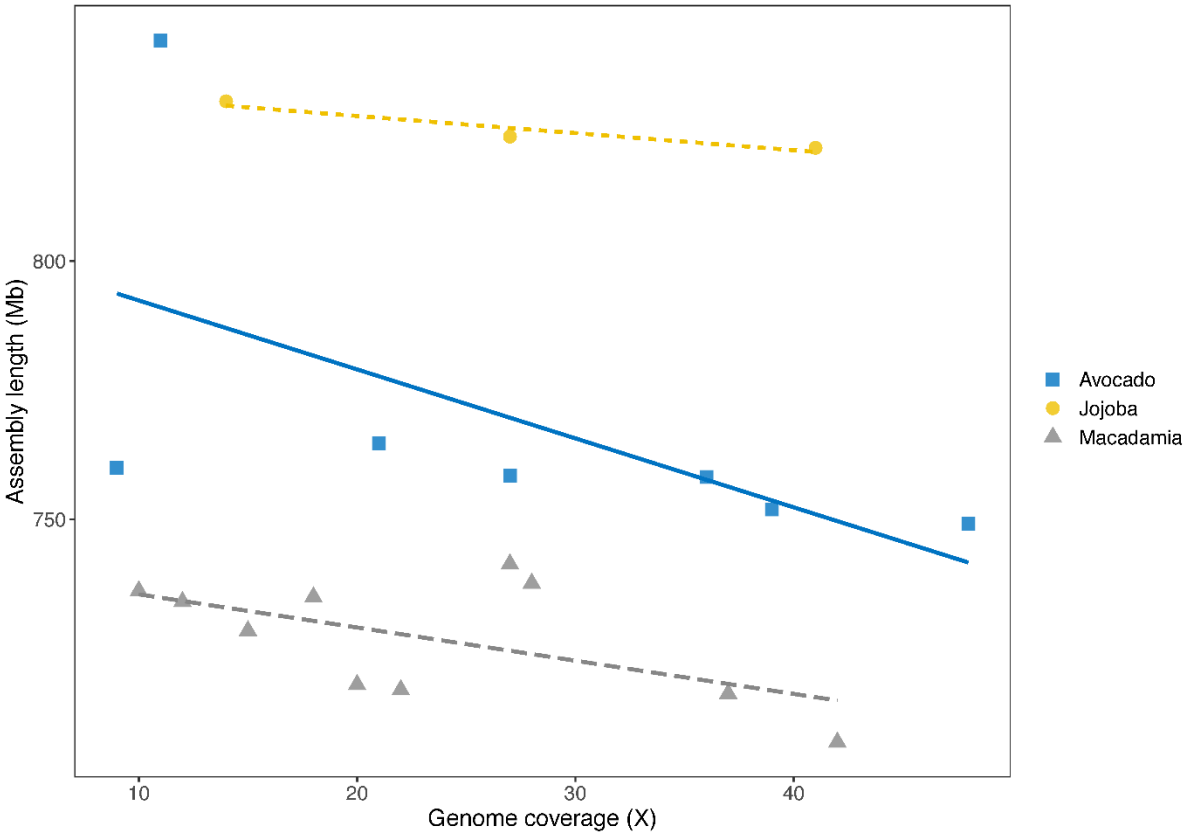
358 Figure1 Influence of data volume on assembly for Macadamia species N50 of contigs is plotted
359 against the genome coverage. Genome sizes used to calculate coverage were; *M.integrifolia*,
360 895Mb [12]; *M janseni*, 780Mb [2]; *M. tetraphylla*, 758Mb [23] and *M. ternifolia*, 758Mb
361 (not known but assumed the same as *M. tetraphylla* due to similar assembly size).

362



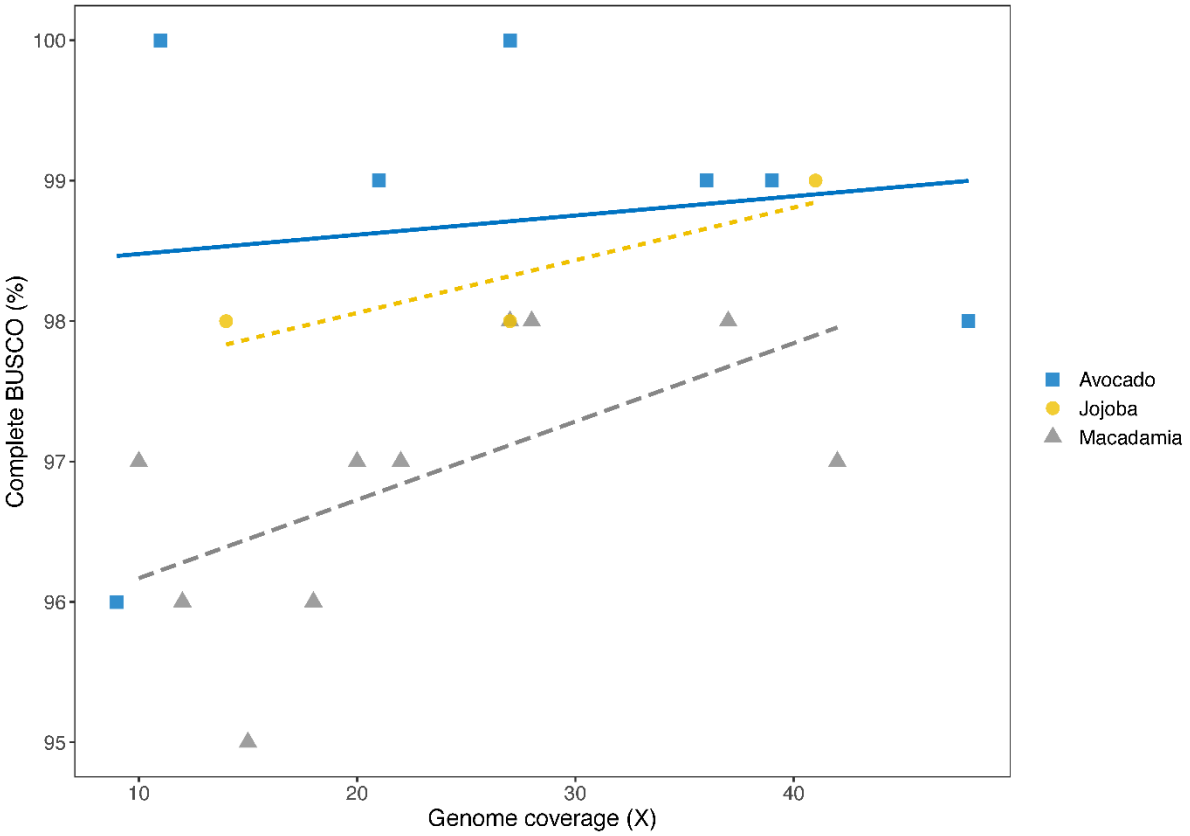
363

364 Figure 2 Influence of data volume on assembly for diverse species. N50 of contigs is plotted
365 against the genome coverage. Genome sizes used to calculate coverage were, jojoba,
366 1003Mb[14]; avocado, 920Mb [24] and as in figure 1 for *Macadamia* species



367

368 Figure 3 Decrease in length of total assembly as more genome coverage is used in the assembly



369

370

371

372

373

374

375

376

377

378

379

380

381

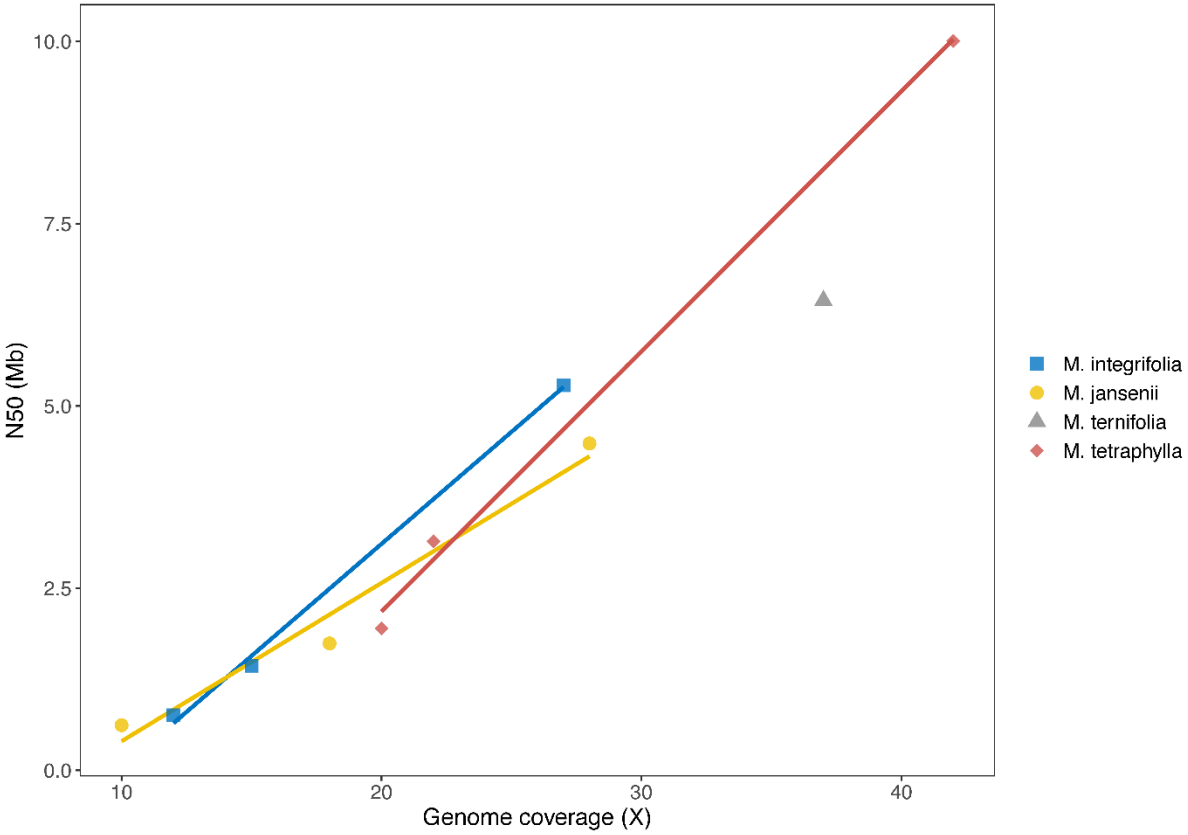
382

383

384

385

386 Figure 4 Improvement in genome completeness (BUSCO%) with genome coverage

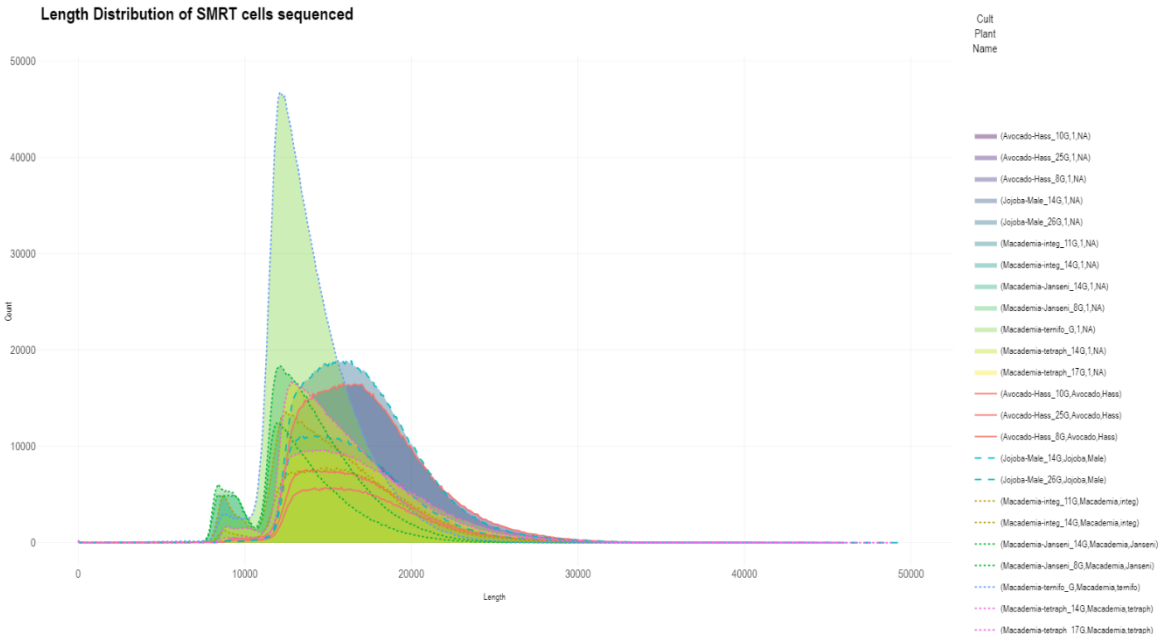


387

388 Figure S1: Size distribution of reads sequenced

389

390



391

392

393

394 Supplementary Table 1 Data for associated contigs in IPA assemblies

395

396

	M. jansenii	*M. integrifolia*	*M. ternifolia*	*M. tetraphylla*	Jojoba	Avocado
Contig N50	0.45Mb	1.23Mb	0.77Mb	1.83Mb	1.69Mb	1.53Mb
Longest Contig	5.23Mb	10.22Mb	5.68Mb	14.97Mb	8.25Mb	10.0Mb
Assembly Length	527Mb	671Mb	590Mb	655Mb	738Mb	788Mb
Number of Contigs	3966	3226	3006	2103	1999	3196

397

398

399