# Next generation sequencing to investigate genomic diversity in Caryophyllales

Boas Pucker[1,2*], Tao Feng[1,3], Samuel F. Brockington[1,2]

1 Evolution and Diversity, Plant Sciences, University of Cambridge, Cambridge, United Kingdom

2 Genetics and Genomics of Plants, CeBiTec & Faculty of Biology, Bielefeld University, Bielefeld, Germany

3 Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, China

* corresponding author: Boas Pucker, bpucker@cebitec.uni-bielefeld.de

BP: bpucker@cebitec.uni-bielefeld.de, ORCID: 0000-0002-3321-7471

TF: fengtao@wbgcas.cn, ORCID: 0000-0002-0489-2021

SFB: sb771@cam.ac.uk, ORCID: 0000-0003-1216-219X

Key words: whole genome sequencing, genome assembly, anthocyanin, betalain, *Kewa caespitosa*, M*acarthuria australis*, *Pharnaceum exiguum*, *Caryophyllales*

## Abstract

Caryophyllales are a highly diverse and large order of plants with a global distribution. While some species are important crops like *Beta vulgaris*, many others can survive under extreme conditions. This order is well known for the complex pigment evolution, because the red pigments anthocyanin and betalain occur with mutual exclusion in species of the Caryophyllales. Here we report about genome assemblies of *Kewa caespitosa* (Kewaceae), *Macarthuria australis* (Macarthuriaceae), and *Pharnaceum exiguum* (Molluginaceae) which are representing different taxonomic groups in the Caryophyllales. The availability of these assemblies enhances molecular investigation of these species e.g. with respect to certain genes of interest.

## Introduction

Caryophyllales form the largest flowering plant order and are recognized for their outstanding ability to colonise extreme environments. Examples are the evolution of Cactaceae in deserts, extremely fast radiation [1–3] e.g. in arid-adapted Aizoaceae and in carnivorous species in nitrogen-poor conditions. Caryophyllales harbor the greatest concentration of halophytic plant species and display repeated shifts to alpine and arctic habitats in Caryophyllaceae and Montiaceae. Due to these extreme environments, species exhibit many adaptations [2–4] such as specialized betalain pigments to protect photosystems in high salt and high light conditions [5]. There are several examples for repeated evolution in the Caryophyllales e.g. leaf and stem succulence for water storage, various mechanisms for salt tolerance, arid-adapted $C_4$ and CAM photosynthesis [4], and insect trapping mechanisms to acquire nitrogen [6].

In addition, to their fascinating trait evolution, the Caryophyllales are well known for important crops and horticultural species like sugar beet, quinoa and spinach. Most prominent is the genome sequence of *Beta vulgaris* [7] which was often used as a reference for studies within Caryophyllales [7–10]. In addition, genomes of *Spinacia oleracea* [7,11], *Dianthus caryophyllus* [12], *Amaranthus hypochondriacus* [13], *Chenopodium quinoa* [14] were sequenced. Besides *Carnegiea gigantea* and several other cacti [15], recent genome sequencing projects were focused on crops due to their economical relevance. However, genome sequences of other species within the Caryophyllales, are needed to provide insights into unusual patterns of trait evolution.

The evolution of pigmentation is known to be complex within the Caryophyllales [8] with a single origin of betalain and at least three reversals to anthocyanin pigmentation. The biosynthetic pathways for betalain and anthocyanin pigmentation are both well characterized. While previous studies have demonstrated that the genes essential for anthocyanin synthesis persists in betalain pigmented taxa, the fate of the betalain pathway in the multiple reversals to anthocyanin pigmentations is unknown. Here, we sequenced three species from different families to contribute to the genomic knowledge about Caryophyllales: *Kewa caespitosa* (Kewaceae), *Macarthuria australis* (Macarthuriaceae), and *Pharnaceum exiguum* (Molluginaceae) were selected as representatives of anthocyanic lineages within the predominantly betalain pigmented Caryophyllales. *K. caespitosa* and *P. exiguum* are examples of putative reversals from betalain pigmentation to anthocyanic pigmentation, while *Macarthuria* is a lineage that diverged before the inferred origin of betalain pigmentation [8].

Several transcript sequences of the three plants investigated here were assembled as part of the 1KP project [16]. Since the sampling for this transcriptome project was restricted to leaf tissue, available sequences are limited to genes expressed there. Here we report three draft genome sequences to complement the available gene set and to enable analysis of untranscribed sequences like promoters, regulatory elements, pseudogenes, and transposable elements.

## Material & Methods

### Plant material

The seeds of *Kewa caespitosa* (Friedrich) Christenh., *Marcarthuria australis* Hügel ex Endl., and *Pharnaceum exiguum* Adamson were obtained from Millennium Seed Bank (London, UK) and were germinated at the Cambridge University Botanic Garden. The plants were grown in controlled glasshouse under conditions: long-day (16 h light and 8 h dark), 20 °C, 60% humidity. About 100 mg fresh young shoots were collected and immediately frozen in liquid nitrogen. Tissue was ground in liquid nitrogen using a mortar and pestle. DNA was extracted using the QIAGEN DNeasy Plant Mini Kit (Hilden, Germany) and RNA was removed by the QIAGEN DNase-Free RNase Set. DNA quantity and quality were assessed by Nanodrop (Thermofisher Scientific, Waltham, MA, USA) and agarose gel electrophoresis. DNA samples were sent to BGI Technology (Hongkong) for library construction and Illumina sequencing.

### Sequencing

Libraries of *K. caespitosa*, *M. australis*, and *P. exiguum* were sequenced on an Illumina HiSeq X-Ten generating 2x150nt reads (AdditionalFile 1). Trimmomatic v0.36 [17] was applied for adapter removal and quality trimming as described previously [18]. Due to remaining adapter sequences, the last 10 bases of each read were clipped. FastQC [19] was applied to check the quality of the reads.

### Genome size estimation

The size of all three investigated genomes was estimated based on k-mer frequencies in the sequencing reads. Jellyfish v2 [20] was applied for the construction of a k-mer table with parameters described by [21]. The derived histogram was further analyzed by GenomeScope [21] to predict a genome size. This process was repeated for all odd k-mer sizes between 17 and 25 (AdditionalFile 2). Finally, an average value was selected from all successful analyses.

## Genome assembly

The performance of different assemblers on the data sets was tested (AdditionalFile 3, AdditionalFile 4, AdditionalFile 5). While CLC Genomics Workbench performed best for the *M. australis* assembly, SOAPdenovo2 [22] showed the best results for *K. caespitosa* and *P. exiguum* and was therefore selected for the final assemblies. To optimize the assemblies, different k-mer sizes were tested as this parameter can best be adjusted empirically [23]. First, k-mer sizes from 67 to 127 in steps of 10 were evaluated, while most parameters remained on default values (AdditionalFile 6). Second, assemblies with k-mer sizes around the best value of the first round were tested. In addition, different insert sizes were evaluated without substantial effect on the assembly quality. In accordance with good practice, assembled sequences shorter than 500 bp were discarded prior to downstream analyses. Custom Python scripts [18,24] were deployed for assembly evaluation based on simple statistics (e.g. N50, N90, assembly size, number of contigs), number of genes predicted by AUGUSTUS v3.2 [25] *ab initio*, average size of predicted genes, and number of complete BUSCOs [26]. Scripts are available on github: https://github.com/bpucker/GenomeAssemblies2018.

BWA-MEM v0.7 [27] was used with the –m flag to map all sequencing reads back against the assembly. REAPR v1.0.18 [28] was applied on the selected assemblies to identify putative assembly errors through inspection of paired-end mappings and to break sequences at those points.

The resulting assemblies were further polished by removal of non-plant sequences. First, all assembled sequences were subjected to a BLASTn [29] against the sugar beet reference genome sequence RefBeet v1.5 [7,30] and the genome sequences of *Chenopodium quinoa* [14], Carnegiea gigantea [15], *Amaranthus hypochondriacus* [13], and *Dianthus caryophyllus* [12]. Hits below the e-value threshold of $10^{-10}$ were considered to be of plant origin. Second, all sequences without hits in this first round were subjected to a BLASTn search against the non-redundant nucleotide database nt. Sequences with strong hits against bacterial and fungal sequences were removed as previously described [18,24]. BLASTn against the *B. vulgaris* plastome (KR230391.1, [31]) and chondrome (BA000009.3, [32]) sequences was performed to identify and remove sequences from these organelle subgenomes.

## Assembly quality assessment

Mapping of sequencing reads against the assembly and processing with REAPR [28] was the first quality control step. RNA-Seq reads (AdditionalFile 7) were mapped against the assemblies to assess completeness of the gene space and to validate the assembly with an independent data set. STAR v2.5.1b [33] was used for the RNA-Seq read mapping as previously described [24].

## Genome annotation

RepeatMasker [34] was applied using crossmatch [35] to identify and mask repetitive regions prior to gene prediction. Masking was performed in sensitive mode (-s) without screening for bacterial IS elements (-no_is) and skipping interspersed repeats (-noint). Repeat sequences of the Caryophyllales (-species caryophyllales) were used and the GC content was calculated per sequence (-gccalc). Protein coding sequences of transcriptome assemblies (AdditionalFile 7) were mapped to the respective genome assembly via BLAT [36] to generate hints for the gene prediction process as previously described [37]. BUSCO v3 [26] was deployed to optimize species-specific parameter sets for all three species based on the sugar beet parameter set [38]. AUGUSTUS v.3.2.2 [25] was applied to incorporate all available hints with previously described parameter settings to optimize the prediction of non-canonical splice sites [37]. Different combinations of hints and parameters were evaluated to achieve an optimal annotation of all three assemblies. A customized Python script was deployed to remove all genes with premature termination codons in their CDS or spanning positions with ambiguous bases. Representative transcripts and peptides per locus were identified based on maximization of the encoded peptide length. INFERNAL (cmscan) [39] was used for the prediction of non-coding RNAs based on models from Rfam13 [40].

Functional annotation was transferred from *Arabidopsis thaliana* (Araport11) [41] via reciprocal best BLAST hits as previously described [24]. In addition, GO terms were assigned to protein coding genes through an InterProScan5 [42]-based pipeline [24].

## Comparison between transcriptome and genome assembly

The assembled genome sequences were compared against previously published transcriptome assemblies (AdditionalFile 7) in a reciprocal way to assess completeness and differences. BLAT [36] was used to align protein coding sequences against each other. This comparison was limited to the protein coding sequences to avoid biases due to UTR sequences, which are in general less reliably predicted or assembled, respectively [37]. The initial alignments were filtered via filterPSL.pl [43] based on recommended criteria for gene prediction hint generation to remove spurious hits and to reduce the set to the best hit per locus e.g. caused by multiple splice variants.

## Results

### Genome size estimation and genome sequence assembly

Prior to the *de novo* genome assembly, the genome sizes of *Kewa caespitosa*, *Macarthuria australis*, and *Pharnaceum exiguum* were estimated from the sequencing reads (Table 1, AdditionalFile 1). The estimated genome sizes range from 265 Mbp (*P. exiguum*) to 623 Mbp (*M. caespitosa*). Based on these genome sizes, the sequencing coverage ranges from 111x (*K. caespitosa*) to 251x (*M. australis*).

Different assembly tools and parameters were evaluated to optimize the assembly process (AdditionalFile 3, AdditionalFile 4, AdditionalFile 5). Sizes of the final assemblies ranged from 254.5 Mbp (*P. exiguum*) to 531 Mbp (*K. caespitosa*) (Table 1, AdditionalFile 8). The best continuity was achieved for *P. exiguum* with an N50 of 515 Mbp.

**Table 1: Genome size estimation and *de novo* assembly statistics**.

| | *Kewa caespitose* | *Macarthuria australis* | *Pharnaceum exiguum* |
|---|---|---|---|
| Accession | GCA_900322205 | GCA_900322265 | GCA_900322385 |
| Estimated genome size [Mbp] | 623 | 497.5 | 265 |
| Sequencing coverage | 111x | 251x | 206x |
| Assembly size (-N) | 531,205,354 | 525,292,167 | 254,526,612 |
| Number of sequence | 55,159 | 271,872 | 16,641 |
| N50 | 28,527 | 2,804 | 56,812 |
| Max. sequence length | 340,297 | 211,626 | 514,701 |
| GC content | 38.1% | 36.6% | 37.4% |
| Complete BUSCOs | 83.6% | 44.4% | 84.3% |
| Assembler | SOAPdenovo2 | CLC Genomics Workbench v9 | SOAPdenovo2 |
| k-mer size | 79 | Automatic | 117 |

### Assembly validation

The mapping of sequencing reads against the assembled sequences resulted in mating rates of 99.5% (*K. caespitosa*), 98% (*M. australis*), and 94.8% (*P. exiguum*). REAPR identified between 1390 (*P. exiguum*) and 16181 (*M. australis*) FCD errors which were corrected by breaking assembled sequences. The mapping of RNA-Seq reads to the polished assembly resulted in mapping rates of

53.9% (*K. caespitosa*) and 43.1% (*M. australis*), respectively, when only considering uniquely mapped reads. Quality assessment via BUSCO revealed 83.6% (*K. caespitosa*), 44.4% (*M. australis*), and 84.3% (*P. exiguum*) complete benchmarking universal single copy ortholog genes (n=1440). In addition, 6.5% (*K. caespitosa*), 21.7% (*M. australis*), and 4.0% (*P. exiguum*) fragmented BUSCOs as well as 9.9% (*K. caespitosa*), 33.9% (*M. australis*), and 11.7% (*P. exiguum*) missing BUSCOs were identified. The proportion of duplicated BUSCOs ranges from 1.5% (*K. caespitosa*) to 2.1% (*P. exiguum*). The number of duplicated BUSCOs was high in *M. australis* (11.8%) compared to both other genome assemblies (1.5% and 2.1%, respectively).

## Genome annotation

After intensive optimization (AdditionalFile 9), the polished structural annotation contains between 26,155 (*P. exiguum*) and 80,236 (*M. australis*) protein encoding genes per genome (Table 2). The average number of exons per genes ranged from 2.9 (*M. australis*) to 6.6 (*K. caespitosa*). Predicted peptide sequence lengths vary between 241 (*M. australis*) and 447 (*K. caespitosa*) amino acids. High numbers of recovered BUSCO genes support the assembly quality (Fig. 1). Functional annotations were assigned to between 50% (*K. caespitosa*) and 70% (*P. exiguum*) of the predicted genes per species. These assemblies revealed between 598 (P. exiguum) and 1604 (M. australis) putative rRNA, 821 (*K. caespitosa*) to 1492 (*M. australis*) tRNA genes, and additional non-protein-coding RNA genes (Table 2).

**Fig. 1. Assembly completeness.**

Assembly completeness was assessed based on the proportion of complete, fragmented, and missing BUSCOs.

**Table 2: Assembly annotation statistics**.

| | *Kewa caespitosa* | *Macarthuria australis* | *Pharnaceum exiguum* |
|---|---|---|---|
| Final gene number | 50661 | 80236 | 26,155 |
| Functional annotation assigned | 25,058 (49.46%) | 50,536 (62.98%) | 18,372 (70.24%) |
| Average gene lengths [bp] | 5494 | 1936 | 5090 |
| Average mRNA length [bp] | 2143 | 1018 | 2154 |
| Average peptide length [aa] | 447 | 241 | 435 |
| RBHs vs. BeetSet2 | 9,968 | 10,568 | 10,045 |
| Average number of exons per | 6.6 | 2.9 | 6 |

| | | | |
|---|---|---|---|
| gene | | | |
| Number of predicted tRNAs | 821 | 1491 | 1260 |
| Number of predicted rRNAs | 720 | 1604 | 598 |
| Link to data set | https://docs.cebitec.uni-bielefeld.de/s/pZ4kGpPEDtTPgjW<br>(TEMPORARY LINK FOR PEER-REVIEW) | | |

### Comparison between transcriptome and genome assemblies

Previously released transcriptome assemblies were compared to the genome assemblies to assess completeness and to identify differences. In total 44,169 of 65,062 (67.9%) coding sequences of *the K. caespitose* transcriptome assembly were recovered in the corresponding genome assembly. This recovery rate is lower for both *M. australis* assemblies, where only 27,894 of 58,953 (47.3%) coding sequences were detected in the genome assembly. The highest rate was observed for *P. exiguum*, where 37,318 of 42,850 (87.1%) coding sequences were found in the genome assembly. When screening the transcriptome assemblies for transcript sequences predicted based on the genome sequences, the recovery rate was lower (Fig. 2). The number of predicted representative coding sequences with best hits against the transcriptome assembly ranged from 16.3% in *K. caespitosa* to 29.7% in *P. exiguum* thus leaving most predicted coding sequences without a good full length hit in the transcriptome assemblies.

**Fig. 2. Recovery of sequences between transcriptome and genome assemblies.**

The figure displays the percentage of sequences present in one assembly that are recovered or missing in the other assembly type.

## Discussion

An almost perfect match between the predicted genome size and the final assembly size was observed for *P. exiguum*. When taking gaps within scaffolds into account the *K. caespitosa* assembly size reached the estimated genome size. High heterozygosity could be one explanation for the assembly size exceeding the estimated haploid genome size of *M. australis*. The two independent genome size estimations for *M. australis* based on different read data sets indicate almost perfect reproducibility of this method. Although centromeric regions and other low complexity regions were probably underestimated in the genome size estimation as well as in the assembly process, this agreement between estimated genome size and final assembly size indicates a high assembly quality. The continuity of the *P. exiguum* assembly is similar to the assembly continuity of *Dianthus caryophyllus* [12] with a scaffold N50 of 60.7 kb. Additional quality indicators are the high proportion of detected BUSCOs in the final assemblies as well as the high mapping rate of reads against the assemblies. The percentage of complete BUSCOs is in the same range as the value of the *D. caryophyllales* genome assembly which revealed 88.9% complete BUSCOs based on our BUSCO settings. We demonstrate a cost-effective generation of draft genome assemblies of three different plant species. Investing into more paired-end sequencing based on Illumina technology would not substantially increase the continuity of the presented assemblies. This was revealed by initial assemblies for *M. australis* performed with less than half of all generated sequencing reads. Although the total assembly size increased when doubling the amount of incorporated sequencing reads, the continuity is still relatively low. No direct correlation between the sequencing depth and the assembly quality was observed in this study. Genome properties seem to be more influential than the amount of sequencing data. Even very deep sequencing with short reads in previous studies [12,18] was unable to compete with the potential of long reads in genome assembly projects [13,14]. No major breakthroughs were achieved in the development of publicly available assemblers during the last years partly due to the availability of long reads which made it less interesting.

The number of predicted genes in *P. exiguum* is in the range expected for most plants [44,45]. While the predicted gene numbers for *K. caespitosa* and *M. australis* are much higher, they are only slightly exceeding the number of genes predicted for other plants [44,45]. Nevertheless, the assembly continuity and the heterozygosity of *M. australis* are probably the most important factors for the artificially high number of predicted genes. The high percentage of duplicated BUSCOs (11.8%) indicates the presence of both alleles for several genes. As the average gene length in *M. australis* is shorter than in both other assemblies, some gene model predictions might be too short. This gene prediction could be improved by an increase in assembly continuity.

There is a substantial difference between the transcriptome sequences and the predicted transcripts of the genome assembly. The presence of alternative transcripts and fragmented transcripts in the transcriptome assemblies are one explanation why not all transcripts were assigned to a genomic locus.  Some transcripts probably represent genes which are not properly resolved in the genome assemblies. This is especially the case for *M. australis*. The high percentage of complete BUSCOs of the *K. caespitosa* and *P. exiguum* genome assemblies indicate that missing sequences in the genome assemblies account only for a minority of the differences. The complete BUSCO percentage of the *P. exiguum* genome assembly even exceeds the value assigned to the corresponding transcriptome assembly. Although BUSCOs are selected in a robust way, it is likely that some of these genes are not present in the genomes investigated here, since *B. vulgaris* is the closest relative with an almost completely sequenced genome [7]. Our genome assemblies provide additional sequences of genes which are not expressed in the tissues sampled for the generation of the transcriptome assembly. In addition, coding sequences might be complete in the genome assemblies, while low expression caused a fragmented assembly based on RNA-Seq reads. This explains why only a small fraction of the predicted coding sequences of the genome assemblies was mapped to the coding sequences derived from the corresponding transcriptome assembly.

The availability of assembled sequences as well as large sequencing read data sets enables the investigation of repeats e.g. transposable elements across a large phylogenetic distance within the Caryophyllales. It also allows the extension of genome-wide analysis like gene family investigations from *B. vulgaris* across Caryophyllales. As all three species produce anthocyanins, we provide the basis to study the underlying biosynthetic genes. Due to the huge evolutionary distance to other anthocyanin producing species, the availability of these sequences could facilitate the identification of common and unique features of the involved enzymes.

## Author contribution

TF isolated DNA. BP and TF performed data processing, assembly, and annotation. BP, TF, and SFB interpreted the results. BP wrote the initial draft. All authors read and approved the final version of this manuscript.

## Acknowledgements

We thank the CeBiTec Bioinformatic Resource Facility team for great technical support.

## References

1. Brockington SF, Walker RH, Glover BJ, Soltis PS, Soltis DE. Complex pigment evolution in the Caryophyllales. New Phytol. 2011;190: 854–864. doi:10.1111/j.1469-8137.2011.03687.x

2. Yang Y, Moore MJ, Brockington SF, Soltis DE, Wong GK-S, Carpenter EJ, et al. Dissecting Molecular Evolution in the Highly Diverse Plant Clade Caryophyllales Using Transcriptome Sequencing. Mol Biol Evol. 2015;32: 2001–2014. doi:10.1093/molbev/msv081

3. Smith SA, Brown JW, Yang Y, Bruenn R, Drummond CP, Brockington SF, et al. Disparity, diversity, and duplications in the Caryophyllales. New Phytol. 2018;217: 836–854. doi:10.1111/nph.14772

4. Kadereit G, Ackerly D, Pirie MD. A broader model for C4 photosynthesis evolution in plants inferred from the goosefoot family (Chenopodiaceae s.s.). Proc R Soc B Biol Sci. 2012;279: 3304–3311. doi:10.1098/rspb.2012.0440

5. Jain G, Schwinn KE, Gould KS. Betalain induction by l-DOPA application confers photoprotection to saline-exposed leaves of Disphyma australe. New Phytol. 2015;207: 1075–1083. doi:10.1111/nph.13409

6. Thorogood CJ, Bauer U, Hiscock SJ. Convergent and divergent evolution in carnivorous pitcher plant traps. New Phytol. 2018;217: 1035–1041. doi:10.1111/nph.14879

7. Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F, Tafer H, et al. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). Nature. 2014;505: 546–549. doi:10.1038/nature12817

8. Brockington SF, Yang Y, Gandia-Herrero F, Covshoff S, Hibberd JM, Sage RF, et al. Lineage-specific gene radiations underlie the evolution of novel betalain pigmentation in Caryophyllales. New Phytol. 2015;207: 1170–1180. doi:10.1111/nph.13441

9. Stevanato P, Trebbi D, Saccomani M. Single nucleotide polymorphism markers linked to root elongation rate in sugar beet. Biol Plant. 2017;61: 48–54. doi:10.1007/s10535-016-0643-1

10. Kong W, Yang S, Wang Y, Bendahmane M, Fu X. Genome-wide identification and characterization of aquaporin gene family in Beta vulgaris. PeerJ. 2017;5. doi:10.7717/peerj.3747

11. Xu C, Jiao C, Sun H, Cai X, Wang X, Ge C, et al. Draft genome of spinach and transcriptome diversity of 120 Spinacia accessions. Nat Commun. 2017;8. doi:10.1038/ncomms15275

12. Yagi M, Kosugi S, Hirakawa H, Ohmiya A, Tanase K, Harada T, et al. Sequence Analysis of the Genome of Carnation (Dianthus caryophyllus L.). DNA Res Int J Rapid Publ Rep Genes Genomes. 2014;21: 231–241. doi:10.1093/dnares/dst053

13. Lightfoot DJ, Jarvis DE, Ramaraj T, Lee R, Jellen EN, Maughan PJ. Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (Amaranthus hypochondriacus) chromosomes provide insights into genome evolution. BMC Biol. 2017;15: 74. doi:10.1186/s12915-017-0412-4

14. Jarvis DE, Ho YS, Lightfoot DJ, Schmöckel SM, Li B, Borm TJA, et al. The genome of *Chenopodium quinoa*. Nature. 2017;542: 307–312. doi:10.1038/nature21370

15. Copetti D, Búrquez A, Bustamante E, Charboneau JLM, Childs KL, Eguiarte LE, et al. Extensive gene tree discordance and hemiplasy shaped the genomes of North American columnar cacti. Proc Natl Acad Sci. 2017;114: 12003–12008. doi:10.1073/pnas.1706367114

16. Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci U S A. 2014;111: E4859–E4868. doi:10.1073/pnas.1323926111

17. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma Oxf Engl. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170

18. Pucker B, Holtgräwe D, Sörensen TR, Stracke R, Viehöver P, Weisshaar B. A *De Novo* Genome Sequence Assembly of the *Arabidopsis thaliana* Accession Niederzenz-1 Displays

Presence/Absence Variation and Strong Synteny. PLOS ONE. 2016;11: e0164321. doi:10.1371/journal.pone.0164321

19. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data. 2010. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

20. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27: 764–770. doi:10.1093/bioinformatics/btr011

21. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017;33: 2202–2204. doi:10.1093/bioinformatics/btx153

22. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. GigaScience. 2012;1: 18. doi:10.1186/2047-217X-1-18

23. Cha S, Bird DM. Optimizing k-mer size using a variant grid search to enhance de novo genome assembly. Bioinformation. 2016;12: 36–40. doi:10.6026/97320630012036

24. Haak M, Vinke S, Keller W, Droste J, Rückert C, Kalinowski J, et al. High Quality *de Novo* Transcriptome Assembly of *Croton tiglium*. Front Mol Biosci. 2018;5. doi:https://doi.org/10.3389/fmolb.2018.00062

25. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinforma Oxf Engl. 2011;27: 757–763. doi:10.1093/bioinformatics/btr010

26. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinforma Oxf Engl. 2015;31: 3210–3212. doi:10.1093/bioinformatics/btv351

27. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio. 2013; Available: http://arxiv.org/abs/1303.3997

28. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 2013;14: R47. doi:10.1186/gb-2013-14-5-r47

29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410. doi:10.1016/S0022-2836(05)80360-2

30. Holtgräwe D, Rosleff Sörensen T, Parol-Kryger R, Pucker B, Kleinbölting N, Viehöver P, et al. Low coverage re-sequencing in sugar beet for anchoring assembly sequences to genomic positions [Internet]. 2017. Available: https://jbrowse.cebitec.uni-bielefeld.de/RefBeet1.5/

31. Stadermann KB, Weisshaar B, Holtgräwe D. SMRT sequencing only de novo assembly of the sugar beet (Beta vulgaris) chloroplast genome. BMC Bioinformatics. 2015;16. doi:10.1186/s12859-015-0726-6

32. Kubo T, Nishizawa S, Sugawara A, Itchoda N, Estiati A, Mikami T. The complete nucleotide sequence of the mitochondrial genome of sugar beet (Beta vulgaris L.) reveals a novel gene for tRNACys(GCA). Nucleic Acids Res. 2000;28: 2571–2576.

33. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinforma Oxf Engl. 2013;29: 15–21. doi:10.1093/bioinformatics/bts635

34. Smit A, Hubley R, Green P. RepeatMasker Frequently Open-4.0 [Internet]. 2015. Available: http://www.repeatmasker.org/

35. Green P. Consed--A Finishing Package [Internet]. Available: http://www.phrap.org/consed/consed.html#howToGet

36. Kent WJ. BLAT—The BLAST-Like Alignment Tool. Genome Res. 2002;12: 656–664. doi:10.1101/gr.229202

37. Pucker B, Holtgräwe D, Weisshaar B. Consideration of non-canonical splice sites improves gene prediction on the *Arabidopsis thaliana* Niederzenz-1 genome sequence. BMC Res Notes. 2017;10. doi:https://doi.org/10.1186/s13104-017-2985-y

38. Minoche AE, Dohm JC, Schneider J, Holtgräwe D, Viehöver P, Montfort M, et al. Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. Genome Biol. 2015;16: 184. doi:10.1186/s13059-015-0729-7

39. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29: 2933–2935. doi:10.1093/bioinformatics/btt509

40. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 2018;46: D335–D342. doi:10.1093/nar/gkx1038

41. Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. Plant J. 2017;89: 789–804. doi:10.1111/tpj.13415

42. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30: 1236–1240. doi:10.1093/bioinformatics/btu031

43. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: *ab initio* prediction of alternative transcripts. Nucleic Acids Res. 2006;34: W435–W439. doi:10.1093/nar/gkl200

44. Sterck L, Rombauts S, Vandepoele K, Rouzé P, Van de Peer Y. How many genes are there in plants (… and why are they there)? Curr Opin Plant Biol. 2007;10: 199–203. doi:10.1016/j.pbi.2007.01.004

45. Pucker B, Brockington SF. Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes. BMC Genomics. 2018;19: 980. doi:10.1186/s12864-018-5360-z

## Supporting Information

**AdditionalFile 1. Sequencing result overview.**

**AdditionalFile 2. Genome size estimation results.** Genome size estimations with GenomeScope [21] are listed for various k-mer sizes. Two different read sets of *M. australis* were used for the genome size estimation (1=ERR2401802, 2=ERR2401614) to check the reproducibility.

**AdditionalFile 3. Evaluation of assembly attempts of *K. caespitosa*.**

**AdditionalFile 4. Evaluation of assembly attempts of *M. australis*.**

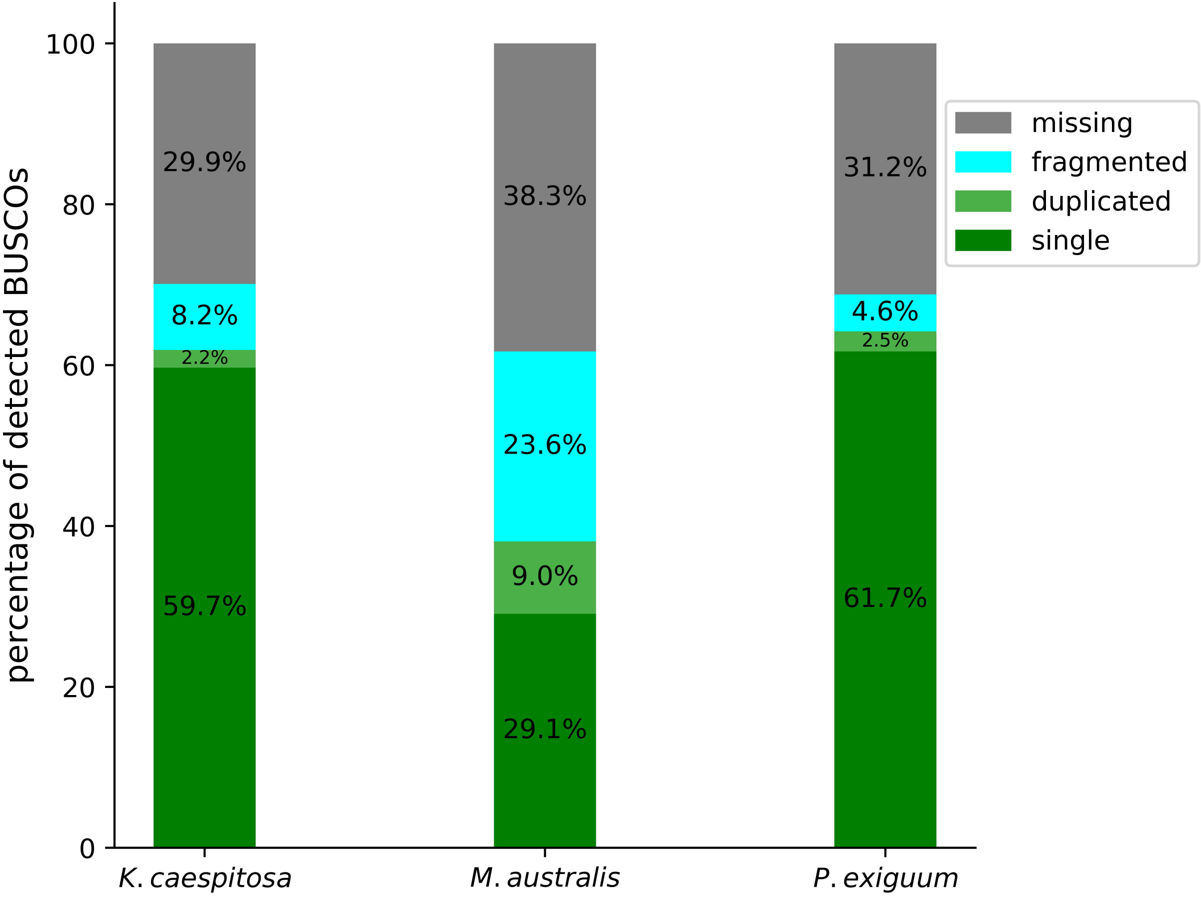
**AdditionalFile 5. Evaluation of assembly attempts of *P. exiguum*.**

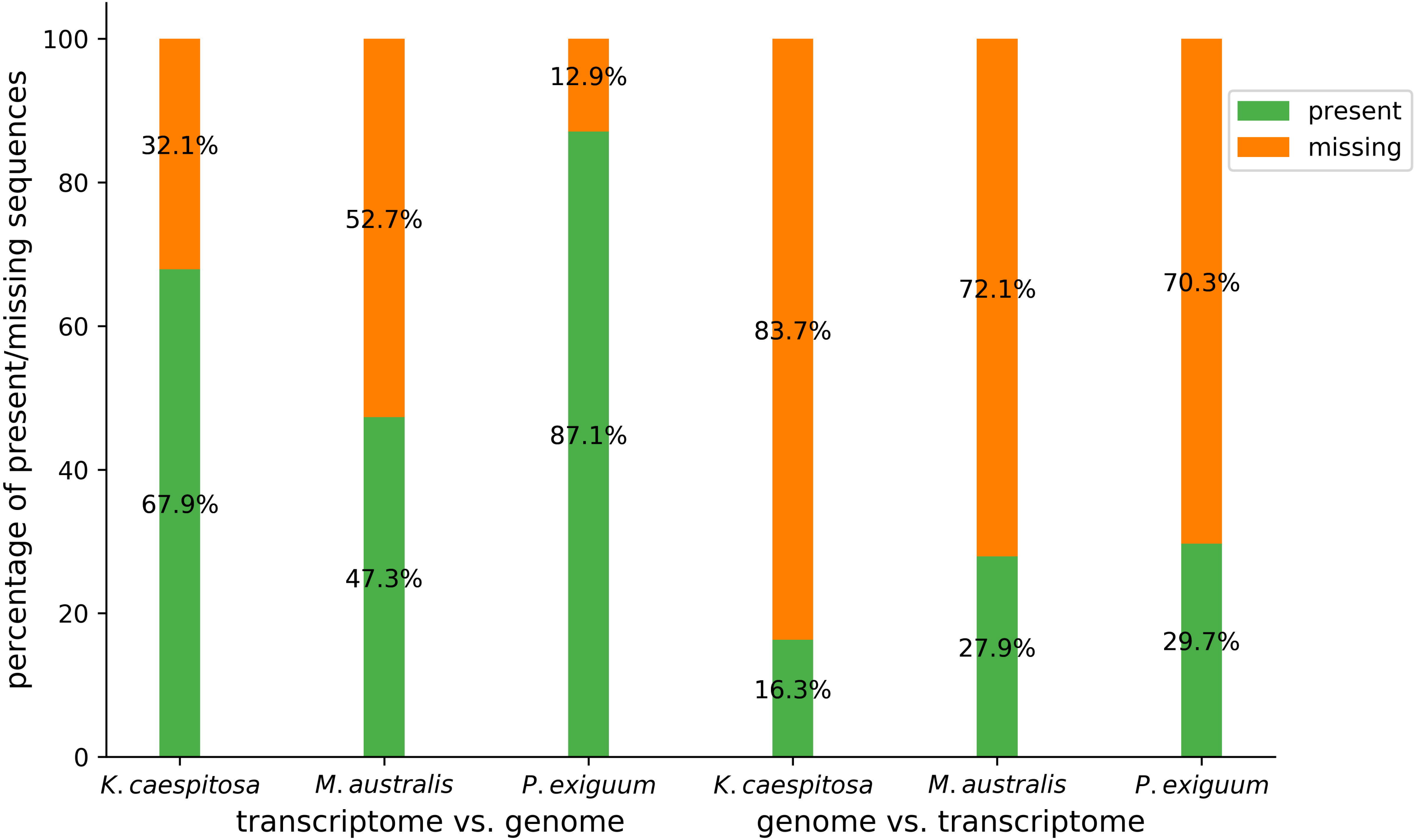
**AdditionalFile 6. Detailed list of assembly parameters.**

**AdditionalFile 7. Gene prediction hint sources.** These RNA-Seq read data sets and transcriptome assemblies were incorporated in the gene annotation process as hints.

**AdditionalFile 8. Assembly attempt evaluation results.** Statistics of raw assemblies were calculated to identify the best parameter settings. Since k-mer size was previously reported as the most important parameter, extensive optimization was performed. In addition, different settings for insert sizes were evaluated for *P. exiguum* (phe001-phe006). Parameter optimization for *M. australis* was performed on a subset of all reads due to availability.

**AdditionalFile 9. Gene prediction statistics.** Different gene prediction approaches were performed during the optimization process. Results of these predictions include *ab initio* gene prediction and hint-based approaches. RNA-Seq reads and coding sequences derived from previous transcriptome assemblies are two incorporated hint types. In addition, we assessed the impact of repeat masking prior to gene prediction.


percentage of detected BUSCOs
0
20
40
60
80
100
K. caespitosa
M. australis
P. exiguum
missing
fragmented
duplicated
single
29.9%
8.2%
2.2%
59.7%
38.3%
23.6%
9.0%
29.1%
31.2%
4.6%
2.5%
61.7%


percentage of present/missing sequences
0
20
40
60
80
100
present
missing
32.1%
67.9%
52.7%
47.3%
12.9%
87.1%
83.7%
16.3%
72.1%
27.9%
70.3%
29.7%
K. caespitosa
M. australis
P. exiguum
K. caespitosa
M. australis
P. exiguum
transcriptome vs. genome
genome vs. transcriptome