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 pigments anthocyanins and betalains 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 one of 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], and *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 [16,17], 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 [18]. 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

62 complement the available gene set and to enable analysis of untranscribed sequences like
63 promoters, regulatory elements, pseudogenes, and transposable elements.

64

65

66 **Material & Methods**

67 **Plant material**

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

78

79 **Sequencing**

80 Libraries of *K. caespitosa*, *M. australis*, and *P. exiguum* were sequenced on an Illumina HiSeq X-Ten
81 generating 2x150nt reads (AdditionalFile 1). Trimmomatic v0.36 [19] was applied for adapter removal
82 and quality trimming as described previously [20]. Due to remaining adapter sequences, the last 10
83 bases of each read were clipped. FastQC [21] was applied to check the quality of the reads.

84

85 **Genome size estimation**

86 The size of all three investigated genomes was estimated based on k-mer frequencies in the
87 sequencing reads. Jellyfish v2 [22] was applied for the construction of a k-mer table with parameters
88 described by [23]. The derived histogram was further analyzed by GenomeScope [23] to predict a
89 genome size. This process was repeated for all odd k-mer sizes between 17 and 25 (AdditionalFile 2).
90 Finally, an average value was selected from all successful analyses.

91

92 **Genome assembly**

93 The performance of different assemblers on the data sets was tested (AdditionalFile 3, AdditionalFile
94 4, AdditionalFile 5). While CLC Genomics Workbench performed best for the *M. australis* assembly,
95 SOAPdenovo2 [24] showed the best results for *K. caespitosa* and *P. exiguum* and was therefore
96 selected for the final assemblies. To optimize the assemblies, different k-mer sizes were tested as
97 this parameter can best be adjusted empirically [25]. First, k-mer sizes from 67 to 127 in steps of 10
98 were evaluated, while most parameters remained on default values (AdditionalFile 6). Second,
99 assemblies with k-mer sizes around the best value of the first round were tested. In addition,
100 different insert sizes were evaluated without substantial effect on the assembly quality. In
101 accordance with good practice, assembled sequences shorter than 500 bp were discarded prior to
102 downstream analyses. Custom Python scripts [20,26] were deployed for assembly evaluation based
103 on simple statistics (e.g. N50, N90, assembly size, number of contigs), number of genes predicted by
104 AUGUSTUS v3.2 [27] *ab initio*, average size of predicted genes, and number of complete BUSCOs
105 [28]. Scripts are available on github: https://github.com/bpucker/GenomeAssemblies2018.

106 BWA-MEM v0.7 [29] was used with the –m flag to map all sequencing reads back against the
107 assembly. REAPR v1.0.18 [30] was applied on the selected assemblies to identify putative assembly
108 errors through inspection of paired-end mappings and to break sequences at those points.

109 The resulting assemblies were further polished by removal of non-plant sequences. First, all
110 assembled sequences were subjected to a BLASTn [31] against the sugar beet reference genome
111 sequence RefBeet v1.5 [7,32] and the genome sequences of *Chenopodium quinoa* [14], Carnegiea
112 gigantea [15], *Amaranthus hypochondriacus* [13], and *Dianthus caryophyllus* [12]. Hits below the e-
113 value threshold of 10^{-10} were considered to be of plant origin. Second, all sequences without hits in
114 this first round were subjected to a BLASTn search against the non-redundant nucleotide database
115 nt. Sequences with strong hits against bacterial and fungal sequences were removed as previously
116 described [20,26]. BLASTn against the *B. vulgaris* plastome (KR230391.1, [33]) and chondrome
117 (BA000009.3, [34]) sequences was performed to identify and remove sequences from these
118 organelle subgenomes.

119

120 **Assembly quality assessment**

121 Mapping of sequencing reads against the assembly and processing with REAPR [30] was the first
122 quality control step. RNA-Seq reads (AdditionalFile 7) were mapped against the assemblies to assess
123 completeness of the gene space and to validate the assembly with an independent data set. STAR
124 v2.5.1b [35] was used for the RNA-Seq read mapping as previously described [26].

125

126 **Genome annotation**

127 RepeatMasker [36] was applied using crossmatch [37] to identify and mask repetitive regions prior to
128 gene prediction. Masking was performed in sensitive mode (-s) without screening for bacterial IS
129 elements (-no_is) and skipping interspersed repeats (-noint). Repeat sequences of the Caryophyllales
130 (-species caryophyllales) were used and the GC content was calculated per sequence (-gccalc).
131 Protein coding sequences of transcriptome assemblies (AdditionalFile 7) were mapped to the
132 respective genome assembly via BLAT [38] to generate hints for the gene prediction process as
133 previously described [39]. BUSCO v3 [28] was deployed to optimize species-specific parameter sets
134 for all three species based on the sugar beet parameter set [40]. AUGUSTUS v.3.2.2 [27] was applied
135 to incorporate all available hints with previously described parameter settings to optimize the
136 prediction of non-canonical splice sites [39]. Different combinations of hints and parameters were
137 evaluated to achieve an optimal annotation of all three assemblies. A customized Python script was
138 deployed to remove all genes with premature termination codons in their CDS or spanning positions
139 with ambiguous bases. Representative transcripts and peptides per locus were identified based on
140 maximization of the encoded peptide length. INFERNAL (cmscan) [41] was used for the prediction of
141 non-coding RNAs based on models from Rfam13 [42].

142 Functional annotation was transferred from *Arabidopsis thaliana* (Araport11) [43] via reciprocal best
143 BLAST hits as previously described [26]. In addition, GO terms were assigned to protein coding genes
144 through an InterProScan5 [44]-based pipeline [26].

145

146 **Comparison between transcriptome and genome assembly**

147 The assembled genome sequences were compared against previously published transcriptome
148 assemblies (AdditionalFile 7) in a reciprocal way to assess completeness and differences. BLAT [38]
149 was used to align protein coding sequences against each other. This comparison was limited to the
150 protein coding sequences to avoid biases due to UTR sequences, which are in general less reliably
151 predicted or assembled, respectively [39]. The initial alignments were filtered via filterPSL.pl [45]
152 based on recommended criteria for gene prediction hint generation to remove spurious hits and to
153 reduce the set to the best hit per locus e.g. caused by multiple splice variants.

154

155

156

157 **Results**

158 **Genome size estimation and genome sequence assembly**

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

163 Different assembly tools and parameters were evaluated to optimize the assembly process
164 (AdditionalFile 3, AdditionalFile 4, AdditionalFile 5). Sizes of the final assemblies ranged from
165 254.5 Mbp (*P. exiguum*) to 531 Mbp (*K. caespitosa*) (Table 1, AdditionalFile 8). The best continuity
166 was achieved for *P. exiguum* with an N50 of 57 kbp.

167 **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

168

169

170 **Assembly validation**

171 The mapping of sequencing reads against the assembled sequences resulted in mating rates of 99.5%
172 (*K. caespitosa*), 98% (*M. australis*), and 94.8% (*P. exiguum*). REAPR identified between 1390 (*P.*
173 *exiguum*) and 16181 (*M. australis*) FCD errors which were corrected by breaking assembled
174 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 (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

226 probably underestimated in the genome size estimation as well as in the assembly process, this
227 agreement between estimated genome size and final assembly size indicates a high assembly quality.
228 The continuity of the *P. exiguum* assembly is comparable to the assembly continuity of *Dianthus*
229 *caryophyllus* [12] with a scaffold N50 of 60.7 kb. Additional quality indicators are the high proportion
230 of detected BUSCOs in the final assemblies as well as the high mapping rate of reads against the
231 assemblies. The percentage of complete BUSCOs is in the same range as the value of the
232 *D. caryophyllales* genome assembly which revealed 88.9% complete BUSCOs based on our BUSCO
233 settings. We demonstrate a cost-effective generation of draft genome assemblies of three different
234 plant species. Investing into more paired-end sequencing based on Illumina technology would not
235 substantially increase the continuity of the presented assemblies. This was revealed by initial
236 assemblies for *M. australis* performed with less than half of all generated sequencing reads. Although
237 the total assembly size increased when doubling the amount of incorporated sequencing reads, the
238 continuity is still relatively low. No direct correlation between the sequencing depth and the
239 assembly quality was observed in this study. Genome properties seem to be more influential than
240 the amount of sequencing data. Even very deep sequencing with short reads in previous studies
241 [12,20] was unable to compete with the potential of long reads in genome assembly projects [13,14].
242 No major breakthroughs were achieved in the development of publicly available short read
243 assemblers during the last years partly due to the availability of long reads which made it less
244 interesting.

245 The number of predicted genes in *P. exiguum* is in the range expected for most plants [46,47]. While
246 the predicted gene numbers for *K. caespitosa* and *M. australis* are much higher than that for
247 *P. exiguum*, they are only slightly exceeding the number of genes predicted for other plants [46,47].
248 Nevertheless, the assembly continuity and the heterozygosity of *M. australis* are probably the most
249 important factors for the artificially high number of predicted genes. The high percentage of
250 duplicated BUSCOs (11.8%) indicates the presence of both alleles for several genes. As the average
251 gene length in *M. australis* is shorter than in both other assemblies, some gene model predictions
252 might be too short. This gene prediction could be improved by an increase in assembly continuity.

253 There is a substantial difference between the transcriptome sequences and the predicted transcripts
254 of the genome assembly. The presence of alternative transcripts and fragmented transcripts in the
255 transcriptome assemblies are one explanation why not all transcripts were assigned to a genomic
256 locus. Another explanation is the intraspecific variation, as the genome and transcriptome
257 assemblies were generated using different individuals. Some transcripts probably represent genes
258 which are not properly resolved in the genome assemblies. This is especially the case for *M. australis*.
259 The high percentage of complete BUSCOs of the *K. caespitosa* and *P. exiguum* genome assemblies
260 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, such as gene family investigations from *B. vulgaris* to more representatives 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

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

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

297 4. Kadereit G, Ackerly D, Pirie MD. A broader model for C4 photosynthesis evolution in plants
298 inferred from the goosefoot family (Chenopodiaceae s.s.). Proc R Soc B Biol Sci. 2012;279:
299 3304–3311. doi:10.1098/rspb.2012.0440

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

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

305 7. Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F, Tafer H, et al. The
306 genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). Nature. 2014;505:
307 546–549. doi:10.1038/nature12817

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

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

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

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

318 12. Yagi M, Kosugi S, Hirakawa H, Ohmiya A, Tanase K, Harada T, et al. Sequence Analysis of the
319 Genome of Carnation (Dianthus caryophyllus L.). DNA Res Int J Rapid Publ Rep Genes Genomes.
320 2014;21: 231–241. doi:10.1093/dnares/dst053

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

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

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

330 16. Shimada S, Takahashi K, Sato Y, Sakuta M. Dihydroflavonol 4-reductasecDNA from non-
331 Anthocyanin-Producing Species in the Caryophyllales. Plant Cell Physiol. 2004;45: 1290–1298.
332 doi:10.1093/pcp/pch156

333 17. Shimada S, Inoue YT, Sakuta M. Anthocyanidin synthase in non-anthocyanin-producing
334 Caryophyllales species. Plant J. 2005;44: 950–959. doi:10.1111/j.1365-313X.2005.02574.x

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

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

340 20. Pucker B, Holtgräwe D, Sörensen TR, Stracke R, Viehöver P, Weisshaar B. A *De Novo* Genome
341 Sequence Assembly of the *Arabidopsis thaliana* Accession Niederzenz-1 Displays
342 Presence/Absence Variation and Strong Synteny. PLOS ONE. 2016;11: e0164321.
343 doi:10.1371/journal.pone.0164321

344 21. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. 2010
345 [cited 14 Dec 2017]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

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

348 23. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al.
349 GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017;33:
350 2202–2204. doi:10.1093/bioinformatics/btx153

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

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

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

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

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

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

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

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

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

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

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

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

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

383 37. Green P. Consed--A Finishing Package [Internet]. [cited 11 Feb 2019]. Available:
384 http://www.phrap.org/consed/consed.html#howToGet

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

387 39. Pucker B, Holtgräwe D, Weisshaar B. Consideration of non-canonical splice sites improves gene
388 prediction on the *Arabidopsis thaliana* Niederzenz-1 genome sequence. BMC Res Notes.
389 2017;10. doi:https://doi.org/10.1186/s13104-017-2985-y

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

393 41. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics.
394 2013;29: 2933–2935. doi:10.1093/bioinformatics/btt509

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

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

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

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

46. 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

47. 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 [23] 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.

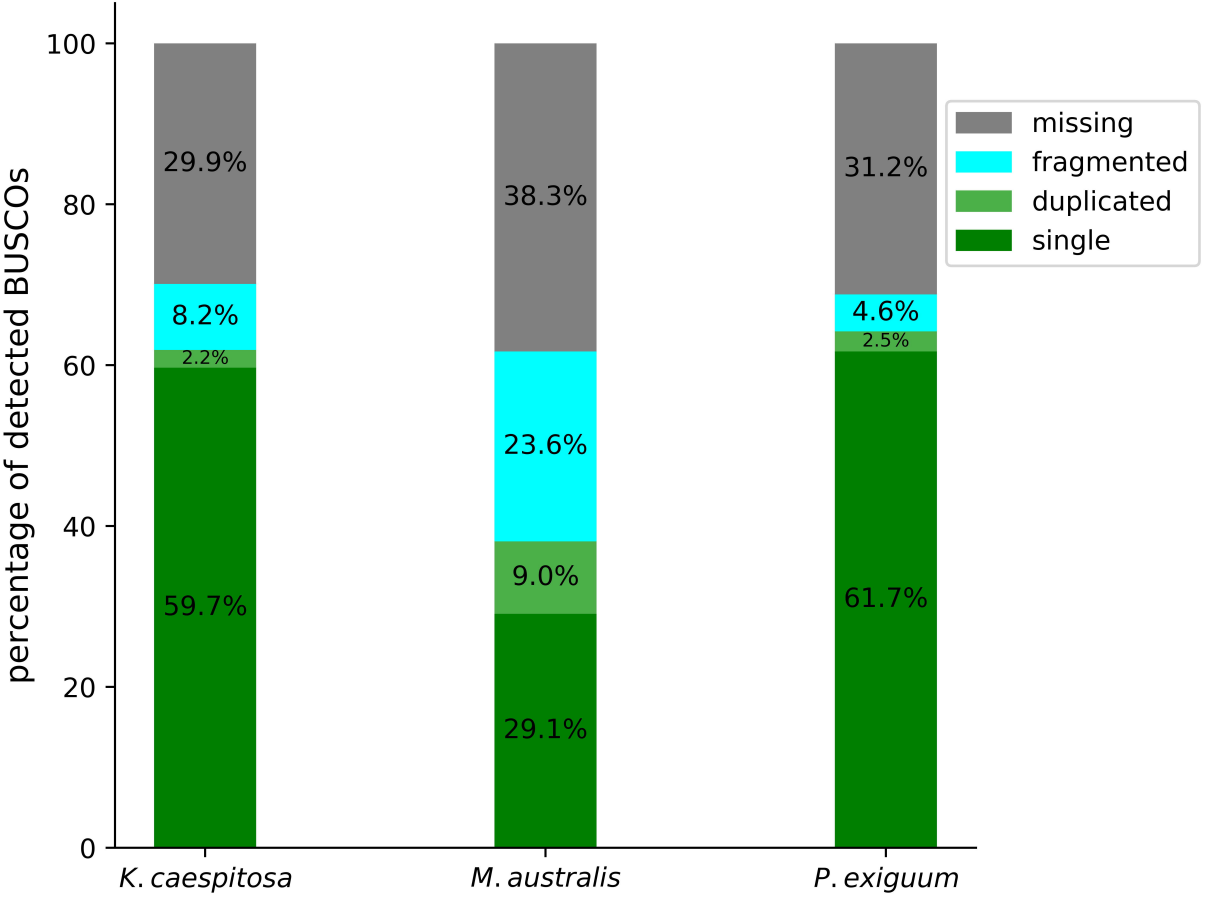
AdditionalFile 7. Gene prediction hint sources. These RNA-Seq read data sets and transcriptome assemblies were incorporated in the gene annotation process as hints.

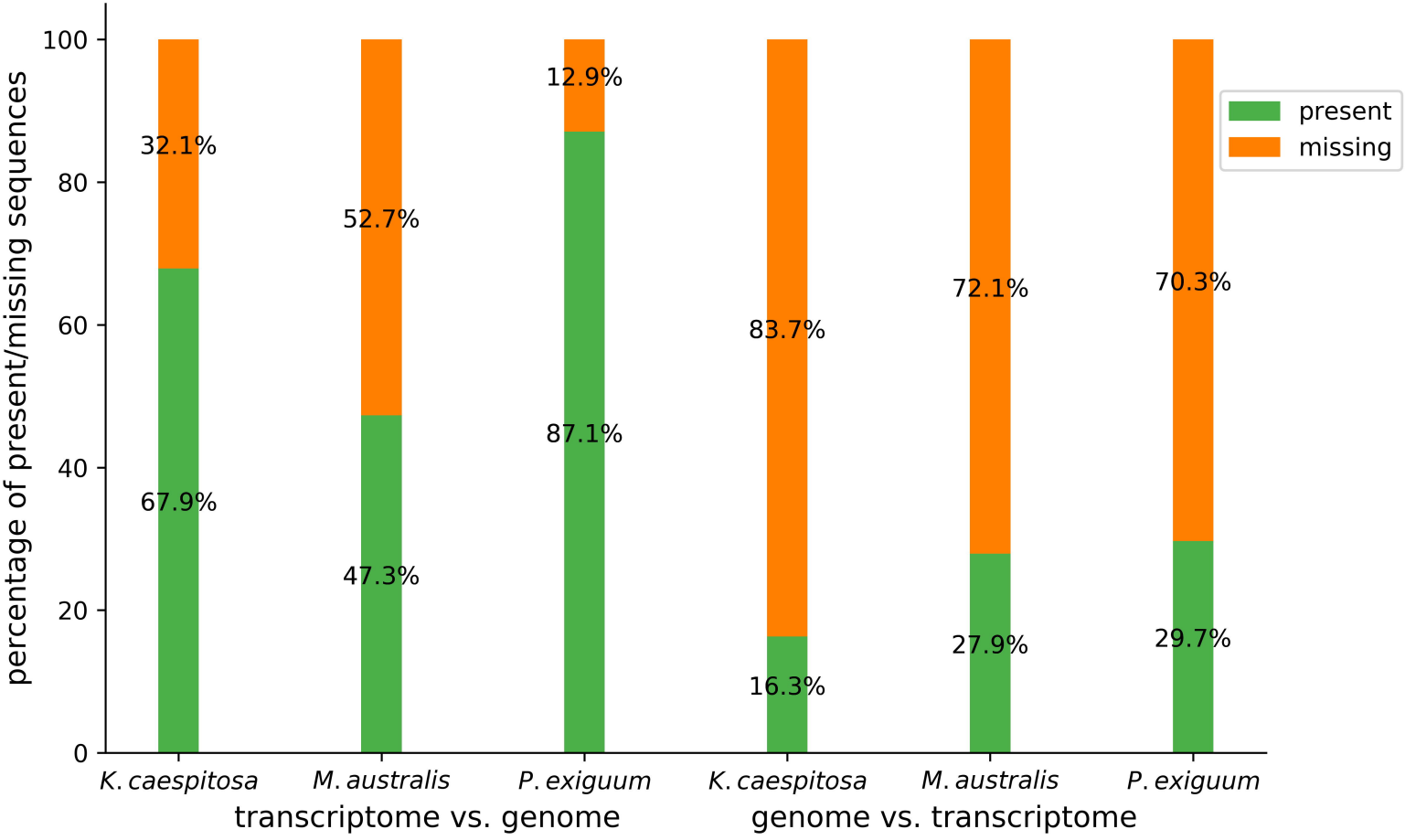
AdditionalFile 8. Assembly attempt evaluation results. Statistics of raw assemblies were calculated to identify the best parameter settings. Since k-mer size was previously reported as the most important parameter, extensive optimization was performed. In addition, different settings for insert sizes 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

432 hint-based approaches. RNA-Seq reads and coding sequences derived from previous transcriptome
433 assemblies are two incorporated hint types. In addition, we assessed the impact of repeat masking
434 prior to gene prediction.

435


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


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