1   **Next generation sequencing to investigate genomic diversity in Caryophyllales**

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

3   1 Evolution and Diversity, Plant Sciences, University of Cambridge, Cambridge, United Kingdom

4   2 Genetics and Genomics of Plants, CeBiTec & Faculty of Biology, Bielefeld University, Bielefeld,
5   Germany

6   3 Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, China

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

8

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

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

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

12

15

16   **Abstract**

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

25

26

27

28   **Introduction**

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

39   In addition, to their fascinating trait evolution, the Caryophyllales are well known for important crops
40   and horticultural species like sugar beet, quinoa and spinach. Most prominent is the genome
41   sequence of *Beta vulgaris* [7] which was often used as a reference for studies within Caryophyllales
42   [7–10]. In addition, genomes of *Spinacia oleracea* [7,11], *Dianthus caryophyllus* [12], *Amaranthus*
43   *hypochondriacus* [13], *Chenopodium quinoa* [14] were sequenced. Besides *Carnegiea gigantea* and
44   several other cacti [15], recent genome sequencing projects were focused on crops due to their
45   economical relevance. However, genome sequences of other species within the Caryophyllales, are
46   needed to provide insights into unusual patterns of trait evolution.

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

59   Several transcript sequences of the three plants investigated here were assembled as part of the 1KP
60   project [16]. Since the sampling for this transcriptome project was restricted to leaf tissue, available
61   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 [17] was applied for adapter removal
82   and quality trimming as described previously [18]. Due to remaining adapter sequences, the last 10
83   bases of each read were clipped. FastQC [19] 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 [20] was applied for the construction of a k-mer table with parameters
88   described by [21]. The derived histogram was further analyzed by GenomeScope [21] 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

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].

125

126   **Genome annotation**

127   RepeatMasker [34] was applied using crossmatch [35] 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 [36] to generate hints for the gene prediction process as
133   previously described [37]. BUSCO v3 [26] was deployed to optimize species-specific parameter sets
134   for all three species based on the sugar beet parameter set [38]. AUGUSTUS v.3.2.2 [25] was applied
135   to incorporate all available hints with previously described parameter settings to optimize the
136   prediction of non-canonical splice sites [37]. 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) [39] was used for the prediction of
141   non-coding RNAs based on models from Rfam13 [40].

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

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 [36]
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 [37]. The initial alignments were filtered via filterPSL.pl [43]
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 515 Mbp.

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

175 53.9% (*K. caespitosa*) and 43.1% (*M. australis*), respectively, when only considering uniquely mapped

176 reads. Quality assessment via BUSCO revealed 83.6% (*K. caespitosa*), 44.4% (*M. australis*), and 84.3%

177 (*P. exiguum*) complete benchmarking universal single copy ortholog genes (n=1440). In addition,

178 6.5% (*K. caespitosa*), 21.7% (*M. australis*), and 4.0% (*P. exiguum*) fragmented BUSCOs as well as 9.9%

179 (*K. caespitosa*), 33.9% (*M. australis*), and 11.7% (*P. exiguum*) missing BUSCOs were identified. The

180 proportion of duplicated BUSCOs ranges from 1.5% (*K. caespitosa*) to 2.1% (*P. exiguum*). The number

181 of duplicated BUSCOs was high in *M. australis* (11.8%) compared to both other genome assemblies

182 (1.5% and 2.1%, respectively).

183

184 **Genome annotation**

185 After intensive optimization (AdditionalFile 9), the polished structural annotation contains between

186 26,155 (*P. exiguum*) and 80,236 (*M. australis*) protein encoding genes per genome (Table 2). The

187 average number of exons per genes ranged from 2.9 (*M. australis*) to 6.6 (*K. caespitosa*). Predicted

188 peptide sequence lengths vary between 241 (*M. australis*) and 447 (*K. caespitosa*) amino acids. High

189 numbers of recovered BUSCO genes support the assembly quality (Fig. 1). Functional annotations

190 were assigned to between 50% (*K. caespitosa*) and 70% (*P. exiguum*) of the predicted genes per

191 species. These assemblies revealed between 598 (P. exiguum) and 1604 (M. australis) putative rRNA,

192 821 (*K. caespitosa*) to 1492 (*M. australis*) tRNA genes, and additional non-protein-coding RNA genes

193 (Table 2).

194

195 **Fig. 1. Assembly completeness.**

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

197

198 **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)		

199

200

201 **Comparison between transcriptome and genome assemblies**

202 Previously released transcriptome assemblies were compared to the genome assemblies to assess
203 completeness and to identify differences. In total 44,169 of 65,062 (67.9%) coding sequences of *the*
204 *K. caespitose* transcriptome assembly were recovered in the corresponding genome assembly. This
205 recovery rate is lower for both *M. australis* assemblies, where only 27,894 of 58,953 (47.3%) coding
206 sequences were detected in the genome assembly. The highest rate was observed for *P. exiguum*,
207 where 37,318 of 42,850 (87.1%) coding sequences were found in the genome assembly. When
208 screening the transcriptome assemblies for transcript sequences predicted based on the genome
209 sequences, the recovery rate was lower (Fig. 2). The number of predicted representative coding
210 sequences with best hits against the transcriptome assembly ranged from 16.3% in *K. caespitosa* to
211 29.7% in *P. exiguum* thus leaving most predicted coding sequences without a good full length hit in
212 the transcriptome assemblies.

213

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

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

217

218

219 **Discussion**

220 An almost perfect match between the predicted genome size and the final assembly size was
221 observed for *P. exiguum*. When taking gaps within scaffolds into account the *K. caespitosa* assembly
222 size reached the estimated genome size. High heterozygosity could be one explanation for the
223 assembly size exceeding the estimated haploid genome size of *M. australis*. The two independent
224 genome size estimations for *M. australis* based on different read data sets indicate almost perfect
225 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 similar 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   [12,18] 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 assemblers during
243   the last years partly due to the availability of long reads which made it less interesting.

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

252   There is a substantial difference between the transcriptome sequences and the predicted transcripts
253   of the genome assembly. The presence of alternative transcripts and fragmented transcripts in the
254   transcriptome assemblies are one explanation why not all transcripts were assigned to a genomic
255   locus. Some transcripts probably represent genes which are not properly resolved in the genome
256   assemblies. This is especially the case for *M. australis*. The high percentage of complete BUSCOs of
257   the *K. caespitosa* and *P. exiguum* genome assemblies indicate that missing sequences in the genome
258   assemblies account only for a minority of the differences. The complete BUSCO percentage of the
259   *P. exiguum* genome assembly even exceeds the value assigned to the corresponding transcriptome
260   assembly. Although BUSCOs are selected in a robust way, it is likely that some of these genes are not

261 present in the genomes investigated here, since *B. vulgaris* is the closest relative with an almost
262 completely sequenced genome [7]. Our genome assemblies provide additional sequences of genes
263 which are not expressed in the tissues sampled for the generation of the transcriptome assembly. In
264 addition, coding sequences might be complete in the genome assemblies, while low expression
265 caused a fragmented assembly based on RNA-Seq reads. This explains why only a small fraction of
266 the predicted coding sequences of the genome assemblies was mapped to the coding sequences
267 derived from the corresponding transcriptome assembly.

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

275

276 **Author contribution**

280

281 **Acknowledgements**

283

284 **References**

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

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

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

293   4.   Kadereit G, Ackerly D, Pirie MD. A broader model for C4 photosynthesis evolution in plants
294   inferred from the goosefoot family (Chenopodiaceae s.s.). Proc R Soc B Biol Sci. 2012;279:
295   3304–3311. doi:10.1098/rspb.2012.0440

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

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

301   7.   Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F, Tafer H, et al. The
302   genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). Nature. 2014;505:
303   546–549. doi:10.1038/nature12817

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

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

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

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

314   12.   Yagi M, Kosugi S, Hirakawa H, Ohmiya A, Tanase K, Harada T, et al. Sequence Analysis of the
315   Genome of Carnation (Dianthus caryophyllus L.). DNA Res Int J Rapid Publ Rep Genes Genomes.
316   2014;21: 231–241. doi:10.1093/dnares/dst053

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

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

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

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

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

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

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

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

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

339 21. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al.
340   GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017;33:
341   2202–2204. doi:10.1093/bioinformatics/btx153

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

378   37.   Pucker B, Holtgräwe D, Weisshaar B. Consideration of non-canonical splice sites improves gene
379   prediction on the *Arabidopsis thaliana* Niederzenz-1 genome sequence. BMC Res Notes.
380   2017;10. doi:https://doi.org/10.1186/s13104-017-2985-y

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

384   39.   Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics.
385   2013;29: 2933–2935. doi:10.1093/bioinformatics/btt509

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

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

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

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

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

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

403

404

405   **Supporting Information**

406   **AdditionalFile 1. Sequencing result overview.**

407   **AdditionalFile 2. Genome size estimation results.** Genome size estimations with GenomeScope [21]

408   are listed for various k-mer sizes. Two different read sets of *M. australis* were used for the genome

409   size estimation (1=ERR2401802, 2=ERR2401614) to check the reproducibility.

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

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

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

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

414   **AdditionalFile 7. Gene prediction hint sources.** These RNA-Seq read data sets and transcriptome

415   assemblies were incorporated in the gene annotation process as hints.

416   **AdditionalFile 8. Assembly attempt evaluation results.** Statistics of raw assemblies were calculated

417   to identify the best parameter settings. Since k-mer size was previously reported as the most

418   important parameter, extensive optimization was performed. In addition, different settings for insert

419   sizes were evaluated for *P. exiguum* (phe001-phe006). Parameter optimization for *M. australis* was

420   performed on a subset of all reads due to availability.

421   **AdditionalFile 9. Gene prediction statistics.** Different gene prediction approaches were performed

422   during the optimization process. Results of these predictions include *ab initio* gene prediction and

423   hint-based approaches. RNA-Seq reads and coding sequences derived from previous transcriptome

424   assemblies are two incorporated hint types. In addition, we assessed the impact of repeat masking

425   prior to gene prediction.

426