-
- 1 Next generation sequencing to investigate genomic diversity in Caryophynancs

2 Boas Pucker^{1,2*}, Tao Feng^{1,3}, Samuel F. Brockington^{1,2}

1 Evolution and Diversity, Plant Sciences, University of Cambridge, Cambridge, Boas Pucker^{2,2}, Tao Feng^{2,2}, Samuel F. Brockington^{2,2}
1 Evolution and Diversity, Plant Sciences, University of
2 Genetics and Genomics of Plants, CeBiTec & Fa
5 Germany
3 Wuhan Botanical Garden, Chinese Academy of Sc
- 3 1 Cenetics and Genomics of Plants, CeBiTec & Faculty of Biology, Bielefeld University, Bielefeld University, Bielefeld University, Bielefeld University, Bielefeld University, Bielefeld Iniversity, Bielefeld Sciences, Wuh
-
- 4 2 Genetics and Genomics of Plants, CeBiTec & Faculty of Biology, Bielefeld University, Bielefeld, 5 Germany
5 Germany + Correspond
5 Germany BP: bpuck
5 Germany SFB: sb77
6 Germany SFB: sb77
2
-
-
-
-
-
-
- ⁶ 3 The Starting author: Boas Pucker, bpucker@cebitec.uni-bielefeld.de

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

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

5 SFB: sb771@cam.ac.uk, ORC 8
19 BP: bpucker@cebitec.uni-bielefeld.de, ORCID: 0000-0002-3321-7471
19 TF: fengtao@wbgcas.cn, ORCID: 0000-0002-0489-2021
19 SFB: sb771@cam.ac.uk, ORCID: 0000-0003-1216-219X
2
19 Key words: whole genome sequencing, genome 9012345 9 BE: https://www.marker.unit.project.unit.project.unit.project.unit.project.unit.project.unit.project.unit.project.unit.project.unit.project.unit.project.unit.project.unit.project.unit.project.unit.project.unit.project.un
-
-

11 SFB: sb771@cam.ac.uk, ORCID: 0000-0003-1216-219X
12 Key words: whole genome sequencing, genome assem
Macarthuria australis, Pharnaceum exiguum, Caryophy.
15 **Abstract** Caryophyllales are a highly diverse and large order 12

12

13 Key words: whole genome sequencing, genome asser

14 Macarthuria australis, Pharnaceum exiguum, Caryophy

15

15

16 Abstract

17 Caryophyllales are a highly diverse and large order of

18 species are important 13 14 15 16 17 18 19 20 21 13 Key words: whole genome sequencing, genome assembly, anthocyanin, betalain, kewa caespitosa,
13 Macarthuria australis, Pharmaceum exiguum, Caryophyllales
15 Abstract
17 Caryophyllales are a highly diverse and large orde Macurinum ussirums, Pharmaceum exiguum, Caryophymates

15

16 Abstract

17 Caryophyllales are a highly diverse and large order of plant

18 species are important crops like *Beta vulgaris*, many others

19 This order is we 16 17 18 19 20 11 21 22 23 24 25 17 Caryophy
18 species a
19 This orde
20 and beta
21 genome
22 Pharnace
23 Caryophy
24 e.g. with
25 18 Suppose are important crops like *Beta vulgaris*, many others can survive under extreme conditions.

17 Suppose are important crops like *Beta vulgaris*, many others can survive under extreme conditions.

17 and betalai Species are important crops like Beta vulgaris, many others can survive dilder extreme conditions.
This order is well known for the complex pigment evolution, because the red pigments anthocyanin
and betalain occur with mu 19 This order is an experimental exclusion in species of the Caryophyllales. Here we report about

19 Thermoceum exiguum (Molluginaceae) (*Mocenthuria australis* (Macarthuriaceae), and

19 *Pharnaceum exiguum* (Molluginace 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 22 Pharmaceum exiguum (Molluginaceae) which are representing different taxonomic groups in the
23 Pharmaceum exiguum (Molluginaceae) which are representing different taxonomic groups in the
23 Capacy of these assemblies en 22 Pharmaceum exiguum (Molnuginaceae) 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 24 e.g. with respect to certain genes of interest.
25
26 25
26
27

- --
26
27
- $\frac{27}{1}$

Example 10

29 Caryophyllale

30 to colonise e

31 fast radiatior

22 conditions. C:

23 repeated shii

24 extreme envi

25 to protect ph

26 repeated evc

27 mechanisms

38 mechanisms 29 Caryophyllales form the largest flowering plant of the largest flowering density, the colonise extreme environments. Examples are the evolution of Cactaceae in deserts, extremely fast radiation [1–3] e.g. in and-adapted 33 to colonise extreme environmental and the greatest concentration of halophytic plant species in nitrogen-poor
32 conditions. Caryophyllales harbor the greatest concentration of halophytic plant species and display
33 re 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 environme Example in the greated 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 phot

33 extreme environments, species exhibit many adaptations [2-4] such as specialized betalain pigments
33 to protect photosystems in high salt and high light conditions [5]. There are several examples for
35 repeated evolut 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 133 To protect photosystems in the Caryophyllales e.g. leaf and stem succulence for water storage, various
133 mechanisms for salt tolerance, arid-adapted C₄ and CAM photosynthesis [4], and insect trapping
138 mechanisms 37

38 mechanisms for salt tolerance, arid-adapted C₄ and CAM photosynthesis [4], and insect trapping

38 mechanisms to acquire nitrogen [6].

39 in addition, to their fascinating trait evolution, the Caryophyllales are 38 mechanisms to acquire nitrogen [6].

39 In addition, to their fascinating trait evolution, the Caryophyllales are well known for important crops

39 and horticultural species like sugar beet, quinoa and spinach. Most pr 39 In addition, to their fascinating trait (and horticultural species like suga sequence of *Beta vulgaris* [7] which [7–10]. In addition, genomes of *Spi* hypochondriacus [13], *Chenopodium* several other cacti [15], rece 39 and horticultural species like sugar beet, quinoa and spinach. Most prominent is the genome
39 sequence of *Beta vulgaris* [7] which was often used as a reference for studies within Caryophyllales
39 [7-10]. In addition 44

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 *h* 44 sequence of Beta vulgaris [7] which was often used as a reference for stadies within Caryophyliaes

42 [7-10]. In addition, genomes of *Spinacia oferacea* [7,11], *Dianthus caryophyllus* [12], *Amaranthus*

44 *several*

42 [7–10]. In addition, genomes of Spinacia of defaced [7,11], Dianthus caryophylus [12], Mindialitius hypochondriacus [13], Chenopodium quinoa [14] were sequenced. Besides *Carnegiea gigantea* and several other cacti [15] Example of the react in the multiple and the reacted and the reacted and the reacted of the reacted of the reacted of the reacted of provide insights into unusual patterns of trait evolution.
The evolution of pigmentation 44 seconomical relevance. However, genome sequences of other species within the Caryophyllales, are

44 seconomical relevance. However, genome sequences of other species within the Caryophyllales, are

44 rhe evolution of The evolution of pigmentation is known to be complex within the Caryophyllales [8] with a single

are evolution of pigmentation is known to be complex within the Caryophyllales [8] with a single

origin of betalain and at The evolution of pigmentation is known to be complex within th

48 origin of betalain and at least three reversals to anthocyanin

49 origins in the trait evolution and anthocyanin pigmentation are both w

50 studies have 19 The evolution of betalain and at least three reversals to anthocyanin pigmentation. The biosynthetic

19 origin of betalain and anthocyanin pigmentation are both well characterized. While previous

19 studies have demon 49

48 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 pigmentations 49 studies have demonstrated that the genes essential for anthocyanin synthesis persists in betalain

51 pigmentations is unknown. Here, we sequenced three species from different families to contribute to

52 pigmentations For the studies have demonstrated that the general formulation is performed that the formulations is unknown. Here, we sequenced three species from different families to contribute to
the genomic knowledge about Caryophyll For pigmentations is unknown. Here, we sequenced three species from different families to contribute to
the genomic knowledge about Caryophyllales: *Kewa coespitosa* (Kewaceae), *Macarthuria australis*
(Macarthuriaceae), a The genomic knowledge about Caryophyllales: *Kewa caespitosa* (Kewaceae), *Macarthuria australis*

(Macarthuriaceae), and *Pharnaceum exiguum* (Molluginaceae) were selected as representatives of

anthocyanic lineages withi Improper about Caryophymetes: Kewa cuespitosa (Kewaceae), Moccintaino dostroins

(Macarthuriaceae), and *Pharnaceum exiguum* (Molluginaceae) were selected as representatives of

anthocyanic lineages within the predominantl macarinamizedae), and *Tharnaceum exiguum* (Montghildeae) were selected as representatives of
anthocyanic lineages within the predominantly betalain pigmented Caryophyllales. *K. caespitosa* and
pigmentation, while *Mocart* 25 anthocyanic integes within the predominantly ocealant pigmented caryophyllales. K. caespitosa and

26 P. exigium are examples of putative reversals from betalain pigmentation to anthocyanic

25 pigmentation [8].

36 Sev

F. exigually are examples of putative reversals from betalain pignientation to anthocyante
pigmentation, while Macarthuria is a lineage that diverged before the inferred origin of betalain
pigmentation [8].
Several transcr big 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
p 59 Several transcript
59 Several transcript
60 project [16]. Since
61 sequences are lim 59 Several transcript sequences of the transcription project was restricted to leaf tissue, available
61 sequences are limited to genes expressed there. Here we report three draft genome sequences to
61 sequences are limit 61 sequences are limited to genes expressed there. Here we report three draft genome sequences to
discussed to leaf the sample of the sample of the sequences to leaf the sample of the sequences to leaf the sequence of the 61 sequences are limited to genes expressed there. Here we report three draft genome sequences to

-
-
-

Examplement the available gene set and the available elements.

For promoters, regulatory elements, pseudogenes, and transposable elements.

For Material & Methods

Plant material

The seeds of *Kewa caespitosa* (Friedrich material **& Methods**
65 **Material & Methods**
66 **Material & Methods**
67 **Plant material**
68 The seeds of *Kewa caespitosa* (Friedrich) Christenh., *Marcarthuria austi
69 <i>Pharnaceum exiguum* Adamson were obtained from Mill 65 66 67 68 69 70 71 72 73 66 67 68 69 70 71 72 73 74 -France Internal Source

66 Material The seeds of Kewa

69 Pharnaceum exiguum

70 germinated at the C

71 glasshouse under cor

72 fresh young shoots w

73 liquid nitrogen using

74 (Hilden, Germany) an

91 quality were ass France The seeds of

68 The seeds of

69 *Pharnaceum ex*

70 germinated at

71 glasshouse unc

72 fresh young sh

73 liquid nitrogen

74 (Hilden, Germa

75 quality were a

electrophoresis

77 Illumina sequen Fine seeds of Kewa caespitosa (Friedrich) Christenh., Marcarthura dastruis Hügel ex Endl., and
 Pharnaceum exiguum Adamson were obtained from Millennium Seed Bank (London, UK) and were

germinated at the Cambridge Univer Framaceum exiguality Rounson were obtained from Millennium Seed Bank (London, OK) and were
germinated at the Cambridge University Botanic Garden. The plants were grown in controlled
glasshouse under conditions: long-day (1 Framination at the 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 Fresh young shoots were collected and immediately frozen in liquid nitrogen. Tissue was ground in

173 liquid nitrogen using a mortar and pestle. DNA was extracted using the QIAGEN DNeasy Plant Mini Kit

174 (Hilden, Germa The Unitropen using a mortar and pestle. DNA was extracted using the QIAGEN DNeasy Plant Mini Kit

173 liquid nitrogen using a mortar and pestle. DNA was extracted using the QIAGEN DNeasy Plant Mini Kit

174 (Hilden, Germa (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 s

The quality were assessed by Nanodrop (Thermofisher Scientific, Waltham, MA, USA) and agarose gel

relectrophoresis. DNA samples were sent to BGI Technology (Hongkong) for library construction and

Illumina sequencing.
 S electrophoresis. DNA samples were sent to BGI Technology (Hongkong) for library construction and

177 Illumina sequencing.

179 Sequencing

179 Sequencing

189 Libraries of *K. caespitosa, M. australis,* and *P. exiguum* w The Ulumina sequencing.

78 **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] w TREAD Sequencing

79 Sequencing

80 Libraries of *K. caespi*

81 generating 2x150nt read

82 and quality trimming

83 bases of each read we

84

85 **Genome size estimat**

86 The size of all three 79 80 81 82 83 84 85 86 87 For Sequencing

80 Libraries of a

81 generating 2

82 and quality 1

83 bases of eac

84

85 **Genome size**

86 The size of

87 sequencing lescribed by

Bottom is the caes of K. caes pinsols, M. australis, and P. exiguum were sequenced on an infinition insequence

81 generating 2x150nt reads (AdditionalFile 1). Trimmomatic v0.36 [17] was applied for adapter removal

82 and 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 esti 82 and quality in minimity of destribute per first computer of the minimity of the reads.

82 bases of each read were clipped. Fast QC [19] was applied to check the quality of the reads.

84

84

85 **Genome size estimation** Sa bases of each read were computed. The process of the size of all three investigated genomes was estimated based on k-mer frequencial sequencing reads. Jellyfish v2 [20] was applied for the construction of a k-mer table 85 86 87 88 99 91 91 So Scholle size estimation
86 The size of all three in
87 sequencing reads. Jellyfis
88 described by [21]. The d
89 genome size. This proces
90 Finally, an average value
91 Source the size of all three investigated 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 r 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 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 89 Finally, an average value was selected from all successful analyses.
91 91 Finally, an avege value was selected from all successful analyses.
91

93 The performance o
94 4, AdditionalFile 5)
95 SOAPdenovo2 [22]
96 selected for the fir
197 this parameter can
98 were evaluated, w
199 assemblies with k-
100 different insert si:
101 accordance with go
102 downstream ana 94 4, AdditionalFile 5). While CLC Genomics Workbench performed best for the *M. australis* assembly,
95 SOAPdenovo2 [22] showed the best results for *K. caespitosa* and *P. exiguum* and was therefore
96 selected for the f 4, AdditionalFile 3). While CLC Genomics Workbench performed best for the M. australis assembly,
95 SOAPdenovo2 [22] showed the best results for *K. caespitosa* and *P. exiguum* and was therefore
96 selected for the final 99 Sold cellovez [22] showed the best results for K. caespitosa and Y. exigually and was therefore
96 selected for the final assemblies. To optimize the assemblies, different k-mer sizes were tested as
97 this parameter c 99 this parameter can best be adjusted empirically [23]. 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,
39 assemblies with k 97 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,

99 assemblies with k-mer s 99 assemblies with k-mer sizes around the best value of the first round were tested. In addition,

99 different insert sizes were evaluated without substantial effect on the assembly quality. In

901 accordance with good 99 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 Pyt 101 accordance with good practice, assembled sequences shorter than 500 bp were discarded prior to
102 downstream analyses. Custom Python scripts [18,24] were deployed for assembly evaluation based
103 on simple statistics downstream analyses. Custom Python scripts [18,24] 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 [25]

102 on simple statistics (e.g. N50, N90, assembly size, number of contigs), number of genes predicted by
103 on simple statistics (e.g. N50, N90, assembly size, number of contigs), number of complete BUSCOs
105 [26]. Scrip 203 and number of statistics (e.g. N120, N120, N120, N120, N120, N120, Pomplete BUSCOs

104 AUGUSTUS v3.2 [25] *ab initio*, average size of predicted genes, and number of complete BUSCOs

105 BWA-MEM v0.7 [27] was used wit 104 AUGUSTUS V3.2 [25] *ab initio,* average size of predicted genes, and number of complete BUSCOS
105 [26]. Scripts are available on github: https://github.com/bpucker/GenomeAssemblies2018.
106 BWA-MEM v0.7 [27] was used 106 BWA-MEM v0.7 [27] was used with the $-m$ flag to map all sequencing reads back a assembly. REAPR v1.0.18 [28] was applied on the selected assemblies to identify putative errors through inspection of paired-end mappings 106 BWA-MEM v1.0.18 [28] was applied on the selected assemblies to identify putative assembly
106 errors through inspection of paired-end mappings and to break sequences at those points.
109 The resulting assemblies were f errors through inspection of paired-end mappings and to break sequences at those points.

103 The resulting assemblies were further polished by removal of non-plant sequences. First, all

110 assembled sequences were subje The resulting assemblies were further polished by removal of non-plant sequences

110 assembled sequences were subjected to a BLASTn [29] against the sugar beet reference

111 sequence RefBeet v1.5 [7,30] and the genome s 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 hypochon* 111 sequence RefBeet v1.5 [7,30] 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 threshol signatea [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 Examble the shold of 10⁻¹⁰ were considered to be of plant origin. Second, all sequences without hits in

113 value threshold of 10⁻¹⁰ were considered to be of plant origin. Second, all sequences without hits in

115 th 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 st 115 nt. Sequences with strong hits against bacterial and fungal sequences were removed as previously
115 nt. Sequences with strong hits against the *B. vulgaris* plastome (KR230391.1, [31]) and chondrome
117 (BA000009.3, [

described [18,24]. BLASTn against the *B. vulgaris* plastome (KR230391.1, [31]) and chondrome
117 (BA000009.3, [32]) sequences was performed to identify and remove sequences from these
118 organelle subgenomes.
119 Assembl 118 described [18,24]. BBSTn against the B. vulgaris plastome (KR2303911, [31]) and chondrome
116 (BA000009.3, [32]) sequences was performed to identify and remove sequences from these
118 organelle subgenomes.
119 **Assemb** 118 organelle subgenomes.

118 organelle subgenomes.

119 Assembly quality assessment

121 Mapping of sequencing reads against the assembly and processing with REAPR [28] was the first

121 Mapping of sequencing reads agai 119
120 Assembly quality asses:
121 Mapping of sequencing
122 quality control step. RN
123 completeness of the ge
124 v2.5.1b [33] was used fo ---
120
121
122
123 120 Assembly quality assessment
121 Mapping of sequencing reads
122 quality control step. RNA-Seq
123 completeness of the gene spa
124 v2.5.1b [33] was used for the F 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 [33] wa 123 completeness of the gene space and to validate the assembly with an independent data set. STAR
124 v2.5.1b [33] was used for the RNA-Seq read mapping as previously described [24]. 124 v2.5.1b [33] was used for the RNA-Seq read mapping as previously described [24]. 124 v2.5.1b [33] was used for the RNA-Seq read mapping as previously described for the RNA-Seq read mapping as
The RNA-Seq read mapping as previously described for the RNA-Seq read mapping as previously described for the
T

125

126
127
128
129
130
131
132
133
134 127 RepeatMasker [34] w

128 gene prediction. Ma:

129 elements (-no_is) and

130 (-species caryophylla

131 Protein coding sequ

132 respective genome a

133 previously described

134 for all three species b

135 to incor 128 Repeatment (134) and skipping and masker mode (-s) without screening for bacterial IS

128 elements (-no_is) and skipping interspersed repeats (-noint). Repeat sequences of the Caryophyllales

1329 (-species caryophyll 129 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).

131 Protein coding seq 139 elements (many metapology metapology (content was calculated per sequence (-gccalc).

139 (-species caryophyllales) were used and the GC content was calculated per sequence (-gccalc).

139 rotein coding sequences of tr 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 132 Proticel contribution of the substitute of 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 sug 133 previously described [37]. BUSCO v3 [26] was deployed to optimize species-specific parameter sets
133 previously described [37]. BUSCO v3 [26] was deployed to optimize species-specific parameter sets
133 for all three 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 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 ac 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 g 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. R 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 139 with ambiguous bases. Representative transcripts and peptides per locus were identified based on
139 with ambiguous bases. Representative transcripts and peptides per locus were identified based on
140 mor-coding RNAs

maximization of the encoded peptide length. INFERNAL (cmscan) [39] was used for the prediction of
non-coding RNAs based on models from Rfam13 [40].
142 Functional annotation was transferred from *Arabidopsis thaliana* (Ara

141 non-coding RNAs based on models from Rfam13 [40].

142 Functional annotation was transferred from *Arabidopsis the encoded* (Araport11) [41] via reciprocal best

143 BLAST hits as previously described [24]. In addition Eunctional annotation was transferred from *Arabidop*

143 BLAST hits as previously described [24]. In addition, G

144 through an InterProScan5 [42]-based pipeline [24].

145 **Comparison between transcriptome and genome a** 142 Functional annotation was transferred from Arabidopsis transferred from Arabidopsis thalian (Araport11) [41] via reciprocal best
143 BLAST hits as previously described [24]. In addition, GO terms were assigned to prote 144 Interaction, 1913 Blased pipeline [24].

144 Interactions and therefore and genome assembly

145 Comparison between transcriptome and genome assembly

147 The assembled genome sequences were compared against previously 145

146 **Comparison between transcriptome and genome a**

147 The assembled genome sequences were compa

148 assemblies (AdditionalFile 7) in a reciprocal way to

149 was used to align protein coding sequences agains

150 146
147
148
149
150
151
152
153
154 146 Comparison between transcriptome and genome assembly
148 assemblies (AdditionalFile 7) in a reciprocal way to assess c
149 was used to align protein coding sequences against each of
150 protein coding sequences to avoi 148 assemblies (AdditionalFile 7) in a reciprocal way to assess completeness and differences. BLAT [36]

148 assembles (AdditionalFile 7) in a reciprocal way to assess completeness and differences. BLAT [36]

149 was used 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 assem 149 was used to align protein term, graphene spanned with a term was used to the protein coding sequences to avoid biases due to UTR sequences, which are in general less reliably predicted or assembled, respectively [37]. 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 reluce the set 152 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.
154
155
155 153 reduce the set to the best hit per locus e.g. caused by multiple splice variants.
153
154
155 154
155
156

-- ·
155
156 155

- 157 **Kesants**
158 **Genome**
159 Prior to
160 and *Pho*
161 estimate
162 these ge
163 Differen
164 (Additio
165 254.5 M
166 was ach 159 Prior to the *de novo* genome assembly, the genome sizes and *Pharnaceum exiguum* were estimated from the sequenties estimated genome sizes range from 265 Mbp (*P. exiguum* these genome sizes, the sequencing coverage r
-
-
-
-
-
-
-

168

169

Exampler SOAPdenovo2 CLC Genomics SOAPd

Workbench v9

K-mer size 79 Automatic 117

Assembly validation

The mapping of sequencing reads against the assembled sequences resulted in ma

(K. caespitosa), 98% (M. australis), Workbench v9

Morkbench v9

Automatic

Automatic

Assembly validation

The mapping of sequencing reads against the assembled sequences result

(K. caespitosa), 98% (M. australis), and 94.8% (P. exiguum). REAPR id

exiguum) Automatic
Automatic
Model sequence
P. exiguum). R
which were
e polished asse 117
In mating rate
Itified betweer
by breaking
Ited in mappir Assembly validation
The mapping of sequencing reads against the assembled sequences resulted in r
(K. caespitosa), 98% (M. australis), and 94.8% (P. exiguum). REAPR identifie
exiguum) and 16181 (M. australis) FCD errors wh 169
170
171
172
173
174 170
171
172
173
174 170 Assembly validation
171 The mapping of sequences
172 (*K. caespitosa*), 98%
173 *exiguum*) and 1618
174 sequences. The map 172 (*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 172 (K. caespitosa), 98% (M. dastralis), and 94.8% (P. exiguum). REAPR identified between 1996 (P.
173 exiguum) and 16181 (M. australis) FCD errors which were corrected by breaking assembled
174 sequences. The mapping of Exiguum) and 16161 (M. distribute) 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 174 september 174 septembe
174 september 174 septemb

175 53.53% (K. caespitosa) and 43.1% (M. australis), respectively, when only considering uniquely inapped
176 reads. Quality assessment via BUSCO revealed 83.6% (*K. caespitosa*), 44.4% (*M. australis*), and 84.3%
179 (*P.* 177 reads. Quality assessment via BUSCO revealed 83.6% (K. caespitosa), 44.4% (M. australis), and 94.3%

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

178 (K. caespitos 178 (P. exiguinity complete benchmarking universal single copy ortholog genes (n=1440). In addition,
178 6.5% (K. caespitosa), 33.9% (M. australis), and 11.7% (P. exiguum) fragmented BUSCOs as well as 9.9%
179 (K. caespito 178 6.5% (K. caespitosa), 21.7% (M. australis), and 4.0% (P. exiguum) missing BUSCOs were identified. The
179 (K. caespitosa), 33.9% (M. australis), and 11.7% (P. exiguum) missing BUSCOs were identified. The
180 proportion

179 (K. caespitosa), 33.9% (M. australis), and 11.7% (P. exigually 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 180 proportion of duplicated BUSCOs was high in *M. australis* (11.8%) compared to both other genome assemblies
181 of duplicated BUSCOs was high in *M. australis* (11.8%) compared to both other genome assemblies
183 **Geno** 181 of duplicated BUSCOs was inglimitative customis (11.8%) compared to both other genome assemblies

182 (1.5% and 2.1%, respectively).

183 **Genome annotation**

185 After intensive optimization (AdditionalFile 9), the po 183

184 **Genome annotation**

185 After intensive optimization (*A*

186 26,155 (*P. exiguum*) and 80,2

187 average number of exons per

188 peptide sequence lengths vary

189 numbers of recovered BUSCO

190 were assigned ---
184
185
186
187
188
189
191
192
193 185 After intensive optim
186 26,155 (*P. exiguum*)
187 average number of e
188 peptide sequence ler
189 numbers of recovere
190 were assigned to be
191 species. These assem
192 821 (*K. caespitosa*) to
193 (Table 2). 26, 155 (*P. exiguum*) and 80, 236 (*M. australis*) protein encoding genes per genome (Table 2). The

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

287 average number 26,155 (P. exiguality and 80,256 (M. dustrialis) protein encoding genes per genome (Table 2). The
average number of exons per genes ranged from 2.9 (M. australis) and 447 (K. caespitosa) amino acids. High
numbers of recove average number of exons per genes ranged from 2.9 (M. dustralis) to 6.0 (K. caespitosa). Predicted
peptide sequence lengths vary between 241 (M. australis) and 447 (K. caespitosa) amino acids. High
numbers of recovered BUS 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 189 numbers of the assigned to between 50% (*K. caespitosa*) and 70% (*P. exiguum*) of the predicted genes per
1890 were assigned to between 50% (*K. caespitosa*) and 70% (*P. exiguum*) and 1604 (*M.* australis) putative r 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 (Table 2).
193 Fig. 1. 192 821 (*K. caespitosa*) to 1492 (*M. australis*) tRNA genes, and additional non-protein-coding RNA genes
193 (Table 2).
194 **Fig. 1. Assembly completeness.**
196 Assembly completeness was assessed based on the proportion

-- 1
195
196
197
198

198
|
|
|

199

200

Number of predicted rRNAs 720

Link to data set https://docs.cebitec.uni-bielefeld.de/s/pZ4kGpPEDt

(TEMPORARY LINK FOR PEER-REVIEW)

(TEMPORARY LINK FOR PEER-REVIEW)

Comparison between transcriptome and genome assemblies Eink to data set

In the state of the

Comparison between transcriptome and genome assemblies

Previously released transcriptome as CHEMPORARY LINK FOR PEER-REVIEW)

Comparison between transcriptome and genome assemblies

Previously released transcriptome assemblies were compared to the genome assemblies to

Completeness and to identify differences. I (TEMPORARY ENTERTATION CONTENTATION)

assemblies were compared to the generics. In total 44,169 of 65,062 (67.9

bly were recovered in the correspond

australis assemblies, where only 27,89

enome assembly. The highest rat 200
201
202
203
204
205
206
207
208 ---
201
202
203
204
205
206
207
208
209 202 Previously released transcriptome and genome assembles
202 Previously released transcriptome assemblies were compared
204 K. *caespitose* transcriptome assembly were recovered in the
205 recovery rate is lower for both 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

2025 recovery r 203 completeness and to identity differences. In total 44,169 of 65,062 (67.5%) coding sequences of the K. *coespitose* transcriptome assembly were recovered in the corresponding genome assembly. This recovery rate is lowe 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 wer Ecovery rate is lower for both M. Bustralis assembles, where only 27,894 of 58,953 (47.3%) coung
sequences were detected in the genome assembly. The highest rate was observed for *P. exiguum*,
where 37,318 of 42,850 (87.1% 207 sequences were detected in the genome assembly. The highest rate was observed for P. exiguum,
208 streening the transcriptome assemblies for transcript sequences predicted based on the genome
209 sequences, the recover 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 hit 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. coespitoso* to
212 29.7% in *P. ex* 212 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 the transcr 210 sequences with best hits against the transcriptome assembly ranged from 26.3% in K. caespitosa to
211 29.7% in P. exiguum thus leaving most predicted coding sequences without a good full length hit in
212 fig. 2. Recov

212 213
212 the transcriptome assemblies.
213 Tig. 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 mis 213

214 Fig. 2. Recovery of sequences be

215 The figure displays the percentag

216 other assembly type.

217

218 Discussion

220 An almost perfect match bet

221 observed for 8 aviguum Who

214
215
216
217
218
220
220
222 **Fig. 2. Recovery of sequences between transcriptome and genome assemblies.**

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

217 **Discussion**

218 **Discussion**

2 215 The figure displays the percentage of sequences present in one assembly that are recovered or missing in the

216 other assembly type.

217

217

220 An almost perfect match between the predicted genome size and the fi 216 other assembly type.
217
218 **Discussion**
220 An almost perfect
221 observed for *P. exi*
222 size reached the ϵ
223 assembly size exce
224 genome size estim
225 reproducibility of tl 218
219
220
221
222
223
224
225 219
220
221
222
223
224
225 219 **Discussion**
220 An almost
221 observed fo
222 size reache
223 assembly si
224 genome siz
225 reproducibi 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 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 estimati 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 reproducibil 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 224 genome size estimations for *M. districties* based on different read data sets indicate almost perfect
225 reproducibility of this method. Although centromeric regions and other low complexity regions were 225 reproducibility of this method. Although centromeric regions and other low complexity regions were

External, *probably intertianed in the genome size* and final assembly size indicates a high assembly quality.

228 are continuity of the *P. exiguum* assembly is similar to the assembly continuity of *Dianthus*

229 *cary* 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 229 The continuity of the P. exigualized assembly is similar to the assembly continuity of Diantinus
229 *caryophyllus* [12] with a scaffold N50 of 60.7 kb. Additional quality indicators are the high proportion
230 of dete 229 caryophyllus [12] with a scaffold N50 of 60.7 kb. Additional quality indicators are the migh proportion
230 of detected BUSCOs in the final assemblies as well as the high mapping rate of reads against the
232 assemblie 231 of the percentage of complete BUSCOs is in the same range as the value of the
231 of detections. The percentage of complete BUSCOs is in the same range as the value of the
232 of detings. We demonstrate a cost-effectiv 232 *D. caryophyllales* genome assembly which revealed 88.9% complete BUSCOs based on our BUSCO
232 *o. caryophyllales* genome assembly which revealed 88.9% complete BUSCOs based on our BUSCO
233 settings. We demonstrate a 232 D. caryophylmaes 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. Investi 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. a* 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 2235 substantially increased when doubling the amount of incorporated sequencing reads. Although
235 assemblies for *M. australis* performed with less than half of all generated sequencing reads. Although
237 the total ass 236 assemblies for M. dustrialis performed with less than half of all generated sequencing reads. Although

236 continuity is still relatively low. No direct correlation between the sequencing depth and the

238 continuity 233

238 continuity is still relatively low. No direct correlation between the sequencing depth and the

237 assembly quality was observed in this study. Genome properties seem to be more influential than

240 the amount o

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 unabl 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 break 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 number 242 Interaction and the development of publicly available assemblers during
242 No major breakthroughs were achieved in the development of publicly available assemblers during
243 Interaction of predicted genes in *P. exig* 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 244 The number of predicted genes in *P. exiguum* is in the range expected for most plants
245 the predicted gene numbers for *K. caespitosa* and *M. australis* are much higher, they a
246 exceeding the number of genes pre 244 The number of predicted genes in P. exiguum is in the range expected for most plants [44,45]. When
245 the predicted gene number of genes predicted for other plants [44,45]. Nevertheless, the assembly
244 exceeding the 245 the predicted gene numbers for K. caespitosa and M. australis are model ingite, they are only slightly
245 exceeding the number of genes predicted for other plants [44,45]. Nevertheless, the assembly
243 continuity and

247

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 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 pres 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 percentaion 249 indicates the presence of both alleles for several genes. As the average gene length in M. australis 3
250 shorter than in both other assemblies, some gene model predictions might be too short. This gene
251 There is a 251 prediction could be improved by an increase in assembly continuity.

251 prediction could be improved by an increase in assembly continuity.

252 There is a substantial difference between the transcriptome sequences an There is a substantial difference between the transcriptome sequen
253 of the genome assembly. The presence of alternative transcripts a
255 transcriptome assemblies are one explanation why not all transcri
255 locus. Some 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 tran 254 of the general extendion 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 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. caesp* 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 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 ac 257 the K. caespitosa and P. exiguum genome assembles 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* genom 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 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 260 assembly. Although BUSCOs are selected in a robust way, it is likely that some of these genes are not

Example the method supposes the control of the genome assemblies provide additional sequences of genes
262 completely sequenced genome [7]. Our genome assemblies provide additional sequences of genes
263 which are not expr

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 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 seq 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 correspond 226 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 sequenc 267 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 The availability of assembled sequences as well as larger investigation of repeats e.g. transposable elements acro
270 Caryophyllales. It also allows the extension of genome-w
171 from *B*. *vulgaris* across Caryophyllales 268 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 Iron *B.* Example in the peak of repeats extension of genome-wide analysis like gene family investigations

270 Caryophyllales. It also allows the extension of genome-wide analysis like gene family investigations

271 from *B. vulga* 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 p 272 from B. vulgaris across Caryophylales. As all three species produce anthocyanims, we provide the basis to study the underlying biosynthetic genes. Due to the huge evolutionary distance to other anthocyanin producing sp

272 basis to study the unterlying unery interlated generation of common and unique features of the involved enzymes.
273 anthocyanin producing species, the availability of these sequences could facilitate the identificatio 274 and those metallity of the availability of the availability of the involved enzymes.
275 **Author contribution**
277 TF isolated DNA. BP and TF performed data processing, assembly, and annotation. BP, TF, and SFB
273 int 275
275 **Author contribution**
277 TF isolated DNA. BP and TF performed data processing, a
278 interpreted the results. BP wrote the initial draft. All auth
279 this manuscript.
280 **Acknowledgements**
282 We thank the CeBiT -- 276
277 278
279 280
281 282
283 276 Author contribution
277 TF isolated DNA. BP
278 interpreted the resul
279 this manuscript.
281 Acknowledgements
282 We thank the CeBiTe
283 References 278 Interpreted the results. BP wrote the initial draft. All authors read and approved the final version of
279 Interpreted the results. BP wrote the initial draft. All authors read and approved the final version of
282 We

279 this manuscript.
280
282 We thank the CeBiTec Bioinformatic Resource Facility team for great technical support.
283
284 References
285 1. Brockington SF, Walker RH, Glover BJ, Soltis PS, Soltis DE. Complex pigment evol

280

281 Acknowledgeme

282 We thank the Ce

283

284 References

285 1. Brockingto

285 1. Brockingto

287 2. Yang Y, Mo

288 Molecular --
281
282
283
284
285
285
287
288
289 282 We thank the CeBiTe

282 We thank the CeBiTe

283 1. Brockington SF,

285 1. Brockington SF,

287 2. Yang Y, Moore

288 Molecular Evol

290 3. Smith SA, Brow

290 3. Smith SA, Brow

291 diversity, and diversity, and 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 Div 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 Transcriptom

289 Sequencing. Mol Biol Evol. 20

283

283

282 References

285 1. Brockington SF, Walker RH, Glover BJ, Soltis PS, Soltis DE. Complex pigment evolut

286 Caryophyllales. New Phytol. 2011;190: 854–864. doi:10.1111/j.1469-8137.2011.03

287 2. Yang Y, Moore 284
285
286
287
288
289
291
291 285 1. Brocki

285 1. Brocki

286 Caryop

287 2. Yang Y

288 Molec

290 3. Smith

291 doi:10

292 doi:10 288 Molecular Evolution in the Highly Diverse Plant Clade Caryophyllales Using Transcripton
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 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 doi:10.1111/nph.14772
292 doi:10.1111/nph.14772 290 3. Smith SA, Brown JW, Yang Y, Bruenn R, Drummond CP, Brockington SF, et al. D
diversity, and duplications in the Caryophyllales. New Phytol. 2018;217: 836–8
doi:10.1111/nph.14772 291 diversity, and duplications in the Caryophyllales. New Phytol. 2018;217: 836–854.
doi:10.1111/nph.14772
292 doi:10.1111/nph.14772 292 doi:10.1111/nph.14772
 $\frac{1}{2}$ 292 doi:10.1277
292 de juny 1920
2021 de juny 1920

-
-
-
- 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 I-DOPA application confers p 3304 -3311. doi:10.1098/rspb.2012.0440

296 5. Jain G, Schwinn KE, Gould KS. Betalain induction by l-DOPA application confers photoprote

297 to saline-exposed leaves of Disphyma australe. New Phytol. 2015;207: 1075–1083.
 296 5. Jain G, Schwinn KE, Gould KS. Betalain ind

297 to saline-exposed leaves of Disphyma aus

298 doi:10.1111/nph.13409

299 6. Thorogood CJ, Bauer U, Hiscock SJ. Conve

200 plant traps. New Phytol. 2018;217: 1035–

301 298 doi:10.1111/nph.13409

299 6. Thorogood CJ, Bauer U, Hiscock SJ. Convergent and divergent evolution in carnivoro

2013 Thorogood CJ, Bauer U, Hiscock SJ. Convergent and divergent evolution in carnivoro

2013 Thorogood 299 6. Thorogood CJ, Bauer U,

300 plant traps. New Phytol.

301 7. Dohm JC, Minoche AE, H

302 genome of the recently

303 546–549. doi:10.1038/n;

304 8. Brockington SF, Yang Y, (

305 specific gene radiations

306 New P 209 6. Thomas. New Phytol. 2018;217: 1035–1041. doi:10.1111/nph.14879

201 7. Dohm J.C, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F, Tafer H, et al. The

202 genome of the recently domesticated crop plant su
- 1977 to saline-exposed leaves of Disphyma australe. New Phytol. 2015;207: 1075–1083.

298 6. Thorogood CJ, Bauer U, Hiscock SJ. Convergent and divergent evolution in carnivorous pitcher

299 6. Thorogood CJ, Bauer U, Hisco Example of the recently domesticated crop plant sugar beet (*Beta vulgaris*). Nature. 2014

302 genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). Nature. 2014

304 8. Brockington SF, Yang Y, Gand 303 546–549. doi:10.1038/nature12817
303 546–549. doi:10.1038/nature12817
303 546–549. doi:10.1038/nature12817
305 specific gene radiations underlie the evolution of novel betalain pigmentation in Caryophyllale
306 specifi
-
- Franch transported transformation of the recently domesticated crop plant sugar beet (*Beta vulgar* 546–549. doi:10.1038/nature12817
302 genome of the recently domesticated crop plant sugar beet (*Beta vulgar* 546–549. doi 8. Brockington SF, Yang Y, Gandia-Herr

specific gene radiations underlie the

New Phytol. 2015;207: 1170–1180. doit

303 546 549. Stevanato P, Trebbi D, Saccomani M

808 649. Stevanato P, Trebbi D, Saccomani M

809 10. Ko
-

306 See 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 selongation rate in sugar beet. Biol Plant. 2017;61: 48-54 9. Stevanato P, Trebbi D, Saccomani M. Single nucleotide poly

808 elongation rate in sugar beet. Biol Plant. 2017;61: 48–54. do

309 10. Kong W, Yang S, Wang Y, Bendahmane M, Fu X. Genome-wic

characterization of aquapori

- 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

2010 characterization of aquaporin gene family i 10. Kong W, Yang S, Ward Y, Bendahmane M, Fu X. Genome-wide identification and

110 characterization of aquaporin gene family in Beta vulgaris. Peer. 2017;5.

111 doi:10.7717/peer. 3747

12 11 Xu C, Jiao C, Sun H, Cai X, W characterization of aquaporin gene family in Beta vulgaris. PeerJ. 2017;5.

311 coi: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 transs

diversity of 120 Spinacia 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 an
313 diversity of 120 Spinacia accessions. Nat Commun. 2017;8. doi:10.1038/n
314 12. Yagi M, Kosugi S, Hirakawa 11. Xu C, Jiao C, Sun H, Cai X

12. Yagi M, Kosugi S, Hiraka

314 12. Yagi M, Kosugi S, Hiraka

315 Genome of Carnation (D

2014;21: 231–241. doi:1

317 13. Lightfoot DJ, Jarvis DE, R

318 and Hi-C-based proximit

chromoso 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 th
315 Genome of Carnation (Dianthus car 14 12. Yagi M, Kosugi S, Hirakawa H, Ohmiya A, Tanase K, Harada T, et al. Sequence Analysis of

315 Genome of Carnation (Dianthus caryophyllus L.). DNA Res Int J Rapid Publ Rep Genes Ge

2014;21: 231–241. doi:10.1093/dnare 315 Genome of Carnation (Dianthus caryophyllus L.). DNA Res Int J Rapid Publ Rep Genes Genom
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-2014;21: 231–241. doi:10.1093/dnares/dst053

315 Co14;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 proximi 13. Lightfoot DJ, Jarvis DE, Ramaraj T, Lee R, Jellen and Hi-C-based proximity-guided assembly of a

319 and Hi-C-based proximity-guided assembly of a

319 doi:10.1186/s12915-017-0412-4

321 14. Jarvis DE, Ho YS, Lightfoot and Hi-C-based proximity-guided assembly of amaranth (Amaranthus hypochondriacus)

and Hi-C-based proximity-guided assembly of amaranth (Amaranthus hypochondriacus)

chromosomes provide insights into genome evolution. BMC
-
- 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 *Cheno*
322 15. Co 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
322 *quinoa*. Nature. 2017;542: 307-312. doi:10.1038/nature21370
323 15. Copetti D, Búrquez A, Busta 321 14. Jarvis DE, Ho YS, Lightfoot DJ, Sch

322 quinoa. Nature. 2017;542: 307-3

323 15. Copetti D, Búrquez A, Bustamant

324 gene tree discordance and hemir

325 Proc Natl Acad Sci. 2017;114: 120

326 16. Wickett NJ, Mir 232 11. Jarvis De, Horsey, Hor guinou: Nature. 2017;342: 307–312: doi:10.1038/nature21370
323 15. Copetti D, Búrquez A, Bustamante E, Charboneau JLM, Childs Ki
gene tree discordance and hemiplasy shaped the genomes of N
³²⁶ Proc Natl Acad Sci. 2017;11
- Sample the discordance and hemiplasy shaped the genomes of North American columnar cact

Sample tree discordance and hemiplasy shaped the genomes of North American columnar cact

Proc Natl Acad Sci. 2017;114: 12003-12008. From Natl Acad Sci. 2017;114: 12003-12008. doi:10.1073/pnas.1706367114

325 Proc Natl Acad Sci. 2017;114: 12003-12008. doi:10.1073/pnas.1706367114

326 Individual Columnary diversification of land plants. Proc Natl Acad Sc 326 16. Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, et al. Panalysis of the origin and early diversification of land plants. Proc Natl Acad 164859–E4868. doi:10.1073/pnas.1323926111
329 17. Bolger AM analysis of the origin and early diversification of land plants. Proc Natl Acad Sci U S A. 2014;111:

528 E4859–E4868. doi:10.1073/pnas.1323926111

329 17. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for
- 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
331 228 17. Bolger AM, Lohse M, Usadel B. Trimmomatic:
330 Bioinforma Oxf Engl. 2014;30: 2114–2120. doi
331 18. Pucker B, Holtgräwe D, Sörensen TR, Stracke R
Sequence Assembly of the *Arabidopsis thalian*
- 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 thal* 331 18. Pucker B, Holtgräwe D, Sörensen TR, Stracke R, Viehöver P, Weisshaar B. A *De I*
332 Sequence Assembly of the *Arabidopsis thaliana* Accession Niederzenz-1 Display 332 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 332 Sequence Assembly of the Arabidopsis thaliana Accession Niederzenz-1 Displays

It is made available under a [CC-BY 4.0 International license.](http://creativecommons.org/licenses/by/4.0/) (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. bioRxiv preprint first posted online May. 24, 2019; doi: [http://dx.doi.org/10.1101/646133.](http://dx.doi.org/10.1101/646133) The copyright holder for this preprint

-
-
-
- 373 http://www.repeatmasker.org/
373 http://www.phrap.org/consed/consed.html#howToGet
376 SEMENT-The BLAST-Like Alignment Tool. Genome Res. 2002;12: 656–664.
377 doi:10.1101/gr.229202
378 37 Pucker B, Holtgräwe D, Weisshaa 374 35. Green P. Consed--A Finishing Pa

375 http://www.phrap.org/consed/org/

376 36. Kent WJ. BLAT—The BLAST-Like

377 doi:10.1101/gr.229202

378 37. Pucker B, Holtgräwe D, Weissha

380 2017;10. doi:https://doi.org/10.
 375 http://www.phrap.org/consed/consed.html#howToGet
376 36. Kent WJ. BLAT—The BLAST-Like Alignment Tool. Genome R
377 doi:10.1101/gr.229202
378 37. Pucker B, Holtgräwe D, Weisshaar B. Consideration of non-
379 prediction 376 36. Kent WJ. BLAT—The BLAST-Like Alignment Tool. Genom
377 doi:10.1101/gr.229202
378 37. Pucker B, Holtgräwe D, Weisshaar B. Consideration of nc
379 prediction on the Arabidopsis thaliana Niederzenz-1 gen
380 2017;10. doi:10.1101/gr.229202
378 37. Pucker B, Holtgräwe D, Weisshaar B. Consideration of non-canonical splice sites in
389 prediction on the Arabidopsis thaliana Niederzenz-1 genome sequence. BMC Res
380 2017;10. doi:https://doi 378 37. Pucker B, Holtgräwe D,
379 prediction on the Arabia
380 2017;10. doi:https://do
381 38. Minoche AE, Dohm JC, 5
382 single-molecule transcri
383 184. doi:10.1186/s1305
384 39. Nawrocki EP, Eddy SR. II
2013;29: 2933–
-
-

979 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,

279 prediction on the Arabidopsis indianta Weeder Echonic 2380

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. Exploitin

singl 381 38. Minoche AE, Dohm JC, Schneider J, Holtgräwe D, Viehöver
382 single-molecule transcript sequencing for eukaryotic gene
383 134. doi:10.1186/s13059-015-0729-7
384 39. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold fast 982 single-molecule transcript sequencing for eukaryotic gene prediction. Genome Biol. 2015;

184. doi:10.1186/s13059-015-0729-7

38. Mawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatic 184. doi:10.1186/s13059-015-0729-7

384 39. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics.

2013;29: 2933–2935. doi:10.1093/bioinformatics/btt509

386 40. Kalvari I, Argasinska J 384 39. Nawrocki EP, Eddy SR. Infernal 1.1: 10
385 2013;29: 2933–2935. doi:10.1093/bio
386 40. Kalvari I, Argasinska J, Quinones-Olver
387 shifting to a genome-centric resource
388 D335–D342. doi:10.1093/nar/gkx1038
389 41 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 famili 587 shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 2018

388 535 - D342. doi:10.1093/nar/gkx1038

389 41. Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Ara 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. 201

2022 2014;30: 2014;30: 2014;30: 2014;30: 2014;30: 2014;30: 3936 2013;10.1093/har/gkx1038

2035-D342. doi:10.1093/nar/gkx1038

2035-D342. doi:10.1093/nar/gkx1038

2035 -D342. doi:10.1093/nar/gkx1038

2014 2014;10.1111/tpj.1 41. Cheng C-Y, Krishnakumar V, Chan AP, T

389 41. complete reannotation of the *Arabidor*

391 804. doi:10.1111/tpj.13415

392 42. Jones P, Binns D, Chang H-Y, Fraser M,

393 protein function classification. Bioinfor

43. 42. Jones P, Binns D, Chang H-Y,

392 42. Jones P, Binns D, Chang H-Y,

393 protein function classificatic

394 doi:10.1093/bioinformatics/

395 43. Stanke M, Keller O, Gunduz

396 prediction of alternative trans

397 doi: doi:10.1093/bioinformatics/btu031

395 43. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUC

396 prediction of alternative transcripts. Nucleic Acids Res. 2006;34: W4:

397 doi:10.1093/nar/gkl200

398 44. 43. Stanke M, Keller O, Gunduz I, Hayes

ago prediction of alternative transcripts.

43. Stanke M, Keller O, Gunduz I, Hayes

prediction of alternative transcripts.

403. Sterck L, Rombauts S, Vandepoele K,

plants (... an

390 complete reannotation of the *Arabidopsis thaliana* reference genome. Plant J. 2017;89: 78
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-s 391 complete reannotation of the Arabidopsis thaliant reference genome. Frant J. 2017,09: 769
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-sc 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 395 43. Stanke M, Keller A, Somator, Hayes A, Waack S, Moscow 105. 2006;34: W435-W439.
396 prediction of alternative transcripts. Nucleic Acids Res. 2006;34: W435-W439.
397 doi:10.1093/nar/gkl200
398 44. Sterck L, Rombauts 397 doi:10.1093/nar/gkl200
398 44. Sterck L, Rombauts S, Vandepoele K, Rouzé P, Van de Peer Y. How many genes
399 plants (... and why are they there)? Curr Opin Plant Biol. 2007;10: 199–203.
400 doi:10.1016/j.pbi.2007.01.0 398 44. Sterck L, Rombauts S, Va

399 plants (... and why are th

doi:10.1016/j.pbi.2007.0

401 45. Pucker B, Brockington SF

splice sites in plant geno

403

404 Supporting Information

405 Supporting Information

406 Add 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-canoni

splice sites i

399 plants (10.1016/j.pbi.2007.01.004
399 plants (2007.01.004
399 plants (2008) splants (2018;19: 980. doi:10.1186/s1
399. doi:10.1186/s1
399. doi:10.1186/s1
399. doi:10.1186/s1
399. doi:10.1186/s1
399. doi:10.1186/s1
399. 401 45. Pucker B, Brockington SF. Genc
402 splice sites in plant genomes. B
403
404 **Supporting Information**
406 **AdditionalFile 1. Sequencing result** of splice sites in plant genomes. BMC Genomics. 2018;19: 980. doi:10.1186/s12864-018-5360-z
403
Supporting Information
AdditionalFile 1. Sequencing result overview.

404
405
406 405
406
406

403
403
405 supporting Information
406 AdditionalFile 1. Sequencing result overview. 405 Supporting Information
406 AdditionalFile 1. Sequen 406 AdditionalFile 1. Sequencing result overview.

-
- 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**
-
-
-
-
-

are listed for various k-mer sizes. Two different reductors of M. australis were used for the genome
size estimation (1=ERR2401802, 2=ERR2401614) to check the reproducibility.
AdditionalFile 8. Evaluation of assembly attem Additional File 3. Evaluation of assembly attempts of *K. caespitosa.*

Additional File 4. Evaluation of assembly attempts of *M. australis.*

Additional File 5. Evaluation of assembly attempts of *P. exiguum.*

Additional

410 AdditionalFile 4. Evaluation of assembly attempts of *N. cusspitosa.*
411 AdditionalFile 5. Evaluation of assembly attempts of *P. exiguum.*
413 AdditionalFile 6. Detailed list of assembly parameters.
414 AdditionalFil AdditionalFile 5. 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 predicti AdditionalFile 6. Detailed list of assembly parameters.

AdditionalFile 7. Gene prediction hint sources. These RNA-Seq

assemblies were incorporated in the gene annotation process as hi

AdditionalFile 8. Assembly attempt AdditionalFile 7. Gene prediction hint sources. These
assemblies were incorporated in the gene annotation pr
AdditionalFile 8. Assembly attempt evaluation results
to identify the best parameter settings. Since k-mer
import 414 AdditionalFile 8. Assembly attempt evaluation process as hints.

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

417 to identify the best parameter settings. Sin AdditionalFile 8. Assembly attempt evaluation results. Statistics of rational and the best parameter settings. Since k-mer size was presimportant parameter, extensive optimization was performed. In additional sizes were ev 417 distribution and 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 wer

Entertion, the best parameter settings finds the protecting protecting for insettings for insettings for insett

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

42 419 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.** Diffe Example 1219 sizes were evaluated for P. Exigual pheodo-pheodo). Parameter optimization for M. australis was
approaches were evaluated for P. exiguity M. and the overlapple. The analysis was performed on a subset of all re 421 **AdditionalFile 9. Gene prediction statistics.** Differen
422 during the optimization process. Results of these pr
⁴²³ hint-based approaches. RNA-Seq reads and coding s
424 assemblies are two incorporated hint types. Additional File 9. Gene prediction statistics. Different gene prediction approaches were performed
A22 during the optimization process. Results of these predictions include *ab initio* gene prediction and
A23 hint-based ap 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 t assemblies are two incorporated hint types. In addition, we assessed the impact of repeat masking
prior to gene prediction.
426 425 prior to gene prediction.
426 426 $\frac{1}{2}$

