



OPEN

DATA DESCRIPTOR

# An annotated haplotype-resolved genome sequence assembly of diploid German chamomile, *Matricaria chamomilla*

Woohyeon Cho<sup>1,2</sup>, Jiawu Feng<sup>1</sup>, Manuela Knauff<sup>1</sup>, Sebastian Albrecht<sup>3</sup>, Axel Himmelbach<sup>1</sup>, Lars-Gernot Otto<sup>1</sup> & Martin Mascher<sup>1</sup>✉

*Matricaria chamomilla* L. (chamomile) is a medicinal plant that is widely used for treating skin infections and respiratory ailments. Chamomile belongs to the Asteraceae family of flowering plants and is a primarily outcrossing species with a heterozygous genome. Despite its extensive use, no reference genome has been available for chamomile until now. We present a chromosome-level genome sequence for chamomile which was assembled with TRITEX pipeline from PacBio accurate long reads and chromosome conformation capture sequencing data. The assembled pseudo-haploid genome has a total size of 2.75 Gb, organized into 9 chromosomes with a scaffold N50 of 285 Mb. This high-quality reference genome has a BUSCO value of 98.8% and includes 47,820 functional genes. Additionally, we assembled a haplotype-resolved genome, taking advantage of the high heterozygosity of chamomile. The haplotype assemblies have total sizes of 2.28 Gb and 2.34 Gb and cover 87.6% and 89.8% of the pseudo-haploid reference genome, respectively. Our assemblies provide a valuable resource for genetics and genomics works for chamomile and related members of the Asteraceae.

## Background & Summary

*Matricaria chamomilla* L. (syn. *Matricaria recutita* L., True or German chamomile, hereafter chamomile) is one of the most important medicinal plants worldwide, with a wide range of medical applications like infections and inflammatory diseases of the skin, gastrointestinal diseases and respiratory problems, and a long history of use<sup>1</sup>. *Matricaria chamomilla* belongs to the Asteraceae (syn. Compositae), one of the largest plant families consist of approximately 23,600 species. Chamomile is mainly an outcrossing species<sup>2,3</sup>. This is why individual chamomile genomes are heterozygous and chamomile varieties are heterogeneous mixtures of genotypes. To complicate matters further, chamomile genotypes can be either diploid or tetraploid<sup>4</sup>, with a chromosome number of 9 for the haploid genome<sup>5</sup>.

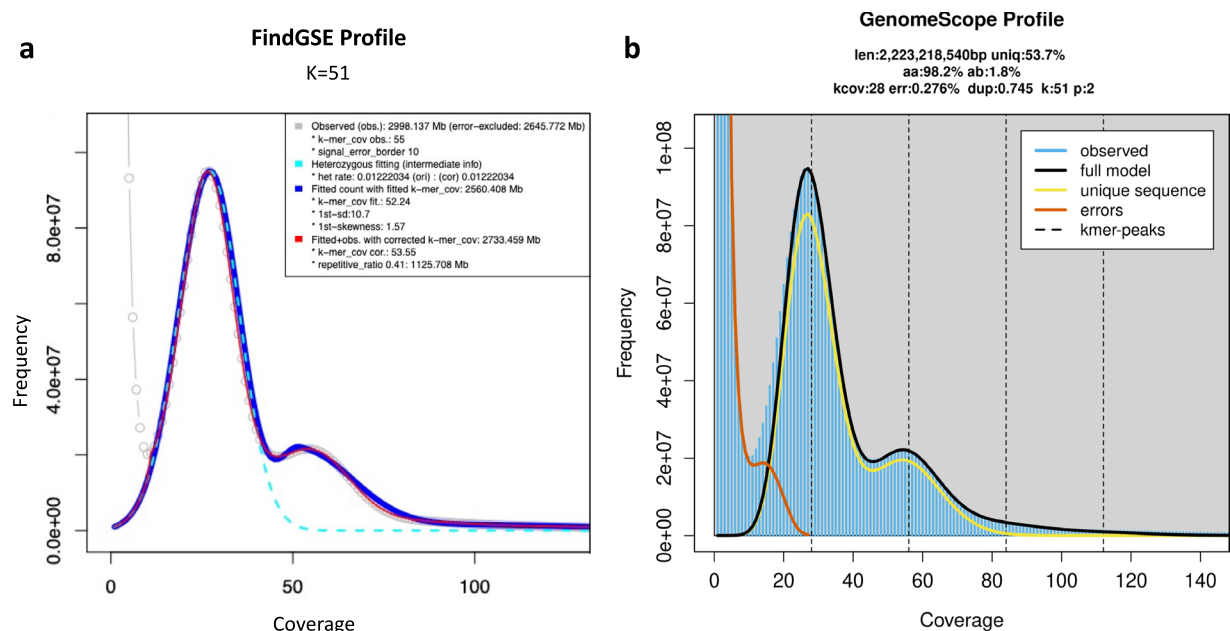
Despite rapid progress in the cost-efficiency and quality of the sequencing technologies, high quality whole-genome sequence assembly have been published hitherto for only a few medicinal and aromatic plants from the Asteraceae family such as sunflower (*Helianthus annuus*, approx. 3.6 Gb genome size<sup>6</sup>), *Artemisia annua* (approx. 1.7 Gb genome size<sup>7</sup>), or *Chrysanthemum makinoi* (approx. 3.2 Gb genome size<sup>8</sup>). No genome assembly is available for chamomile, yet.

In this study, we obtained a high-quality pseudo-haploid sequence assembly as well as haplotype-resolved genome sequences of a diploid *Matricaria chamomilla* individual. To do so, we used PacBio HiFi and Hi-C data and assembled the data with the TRITEX pipeline<sup>9</sup>. The generated draft genome assembly provides a very valuable resource for genetics and genomics works at chamomile.

<sup>1</sup>Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, 06466, Seeland, Germany.

<sup>2</sup>Department of Agriculture, Forestry and Bioresources, Plant Genomics and Breeding Institute, College of Agriculture and Life Sciences, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, South Korea.

<sup>3</sup>PHARMAPLANT GmbH, Arznei- und Gewürzpflanzen Forschungs- und Saatzucht GmbH, Am Westbahnhof 4, D-06556, Artern, Germany. ✉e-mail: [mascher@ipk-gatersleben.de](mailto:mascher@ipk-gatersleben.de)



**Fig. 1** Genome size estimation by K-mer frequency calculation. K-mer profiles were evaluated with FindGSE (a) and GenomeScope (b).

## Methods

**Materials and sequencing.** One single diploid chamomile genotype from the variety ‘Bona’ was chosen for the genome assembly. ‘Bona’ is an agriculturally well performing variety, from which an MS-line was generated, which could be very valuable for further chamomile breeding<sup>10</sup>. Clonal plants of this genotype were preserved in *in vitro* culture at 10°C under short-day conditions (8 h light), until raised for DNA isolation in the greenhouse under long-day conditions (16 h/8 h light/dark period with 22°C/16°C). The ploidy for each clone was analysed by flow-cytometry according to Otto *et al.* 2015 and confirmed as diploid<sup>11</sup>. For the isolation of high molecular weight (HMW) DNA, 7 g leaves were collected from three clonal plants, ground into a fine powder using liquid nitrogen, and stored at -80°C. Approximately 350 µg HMW DNA was purified from the powder using a method as described<sup>12</sup>. For quantification, the Qubit dsDNA High Sensitivity assay kit (Thermo Fisher Scientific, MA, USA) was used. The DNA size profile was analyzed using the Femto Pulse system and the Genomic DNA 165 kb kit (Agilent Technologies Inc, CA, USA). For fragmentation of the HMW DNA (10 µg) into 20 kb fragments, a Megaruptor 3 device (speed: 30) was used (Diagenode, NJ, USA). HiFi SMRTbell libraries were prepared according to the manufacturer’s instructions using the SMRTbell Express Template Prep Kit (Pacific Biosciences, CA, USA). The final HiFi libraries were size-selected (narrow size range: 18–21 kb) using the SageELF system with a 0.75% Agarose Gel Cassette (Sage Sciences, MA, USA) according to standard manufacturer’s protocols. HiFi CCS reads were produced operating the PacBio Sequel IIe instrument (Pacific Biosciences, CA, USA) following the manufacturer’s instructions. In total, 7,281,234 reads corresponding to 163 Gb (HiFi CCS) were generated. The *in situ* Hi-C library preparation was based on the previously published protocol<sup>13</sup>. Sequencing and Hi-C raw data processing was performed as described before<sup>14,15</sup>. A total of 3,710,914,879 reads were generated (about 408 Gb).

**Genome size estimation.** The genome size was estimated by counting k-mer of HiFi data with word size 51 using Jellyfish<sup>16</sup>. A k-mer frequency histogram was also generated with Jellyfish. Genome size and heterozygosity was estimated using findGSE (<https://github.com/schnebergerlab/findGSE>). The estimated genome size is 2.73 Gb with heterozygosity rate of 1.2% and repetitive rate of 41% (Fig. 1a). To determine its ploidy, we analyzed the data using GenomeScope<sup>17</sup> and confirmed it to be diploid (Fig. 1b).

**Pseudo-haploid genome assembly.** The TRITEX long-read assembly pipeline<sup>9</sup> was used for chromosome-scale pseudo-haploid assembly. Contig assembly was performed using hifiasm with Hi-C and HiFi reads, resulting in 2,662 contigs with an N50 of 50 Mb and a total length of 2.75 Gb (Table 1). To exclude bacterial contamination, we filtered 38 contigs using FCS-GX<sup>18</sup>, leaving 2,624 contigs in the resulting dataset. Hi-C reads were mapped to this contig assembly, with 77,998,277 read pairs mapped. After manual curation as described by Marone *et al.*<sup>9</sup>, the final assembly comprised 2,541 scaffolds, with an N50 of 285 Mb. Out of the 2,624 contigs, 92 contigs were anchored to 9 chromosomes with total length of 2.60 Gb, which is 94.6% of total contig assembly (Table 2, Fig. 2). The nine chromosomes are arranged in order from largest to smallest.

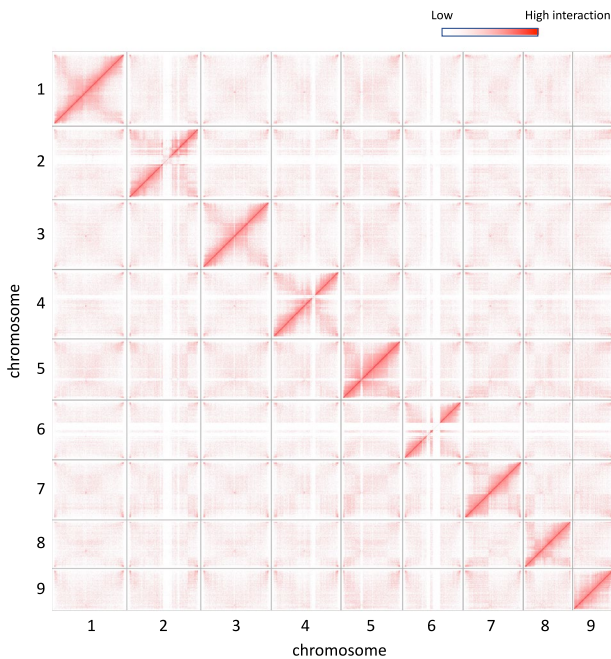
**Haplotype-resolved assembly.** Haplotype-resolved assembly was conducted with the unitig assembly obtained through hifiasm as described by Mascher *et al.*<sup>19</sup>. The unitig assembly contained 28,889 unitigs with an N50 of 2.15 Mb and a total length of 5.76 Gb (Table 1). Total of 4.57 Gb among 5.76 Gb, which is 79% of unitigs were positioned to 9 chromosomes and used for haplotype phasing. Unitigs were phased using PCA with Hi-C

|                     | After primary contig assembly |               | After scaffold assembly     |
|---------------------|-------------------------------|---------------|-----------------------------|
|                     | Unitig                        | Contig        | Pseudomolecule + unanchored |
| No. of assembly     | 28,889                        | 2,662         | 9 + 2,532                   |
| Total length (bp)   | 5,769,542,610                 | 2,757,670,726 | 2,752,017,381               |
| N50 (bp)            | 2,156,804                     | 50,350,694    | 285,044,652                 |
| N90 (bp)            | 46,254                        | 8,462,871     | 195,322,941                 |
| Maximum length (bp) | 22,543,845                    | 141,852,184   | 348,051,304                 |

**Table 1.** Pseudo-haploid assembly statistics after primary contig assembly and scaffold assembly.

| Chromosome | Length (Mb)    |             |             |
|------------|----------------|-------------|-------------|
|            | Pseudo-haploid | Haplotype 1 | Haplotype 2 |
| chr1       | 348.1          | 374.8       | 339.7       |
| chr2       | 340.4          | 290.4       | 308.0       |
| chr3       | 325.8          | 318.7       | 316.7       |
| chr4       | 322.3          | 318.2       | 305.2       |
| chr5       | 285.0          | 98.9        | 105.9       |
| chr6       | 279.4          | 232.7       | 277.3       |
| chr7       | 278.3          | 258.2       | 257.7       |
| chr8       | 227.7          | 210.4       | 231.9       |
| chr9       | 195.3          | 181.0       | 197.3       |
| Total      | 2,602.3        | 2,283.5     | 2,340.0     |

**Table 2.** Chromosome length information of reference and haplotype assembly.



**Fig. 2** Inter-chromosomal Hi-C heatmap of pseudo-haploid assembly.

linkage information. We found unitigs from homozygous region (i.e. those where both haplotypes are identical in sequence) by read coverage and assigned them to both haplotypes. The final, manually curated pseudomolecules of each haplotype have total lengths of 2.28 Gb (87.6%) and 2.34 Gb (89.8%), respectively (Table 2, Fig. 3). Each chromosome is named according to its corresponding chromosome in the pseudo-haploid genome. We only assembled unitigs larger than 300 Kb, and most of the unitigs assigned to chromosome 5 were smaller than the threshold. As a result, chromosome 5 assembled to 98.9 Mb (34.7% of pseudo-haploid chromosome 5) in haplotype 1 and 105.9 Mb (37.2% of pseudo-haploid chromosome 5) in haplotype 2.

**Pseudo-haploid assembly gene annotation.** The EDTA v2.0.0 program was used to annotate the repetitive DNA in the pseudo-haploid sequence assembly<sup>20</sup>. Of a total of 2.75 Gb of sequence, 73.17%, corresponding

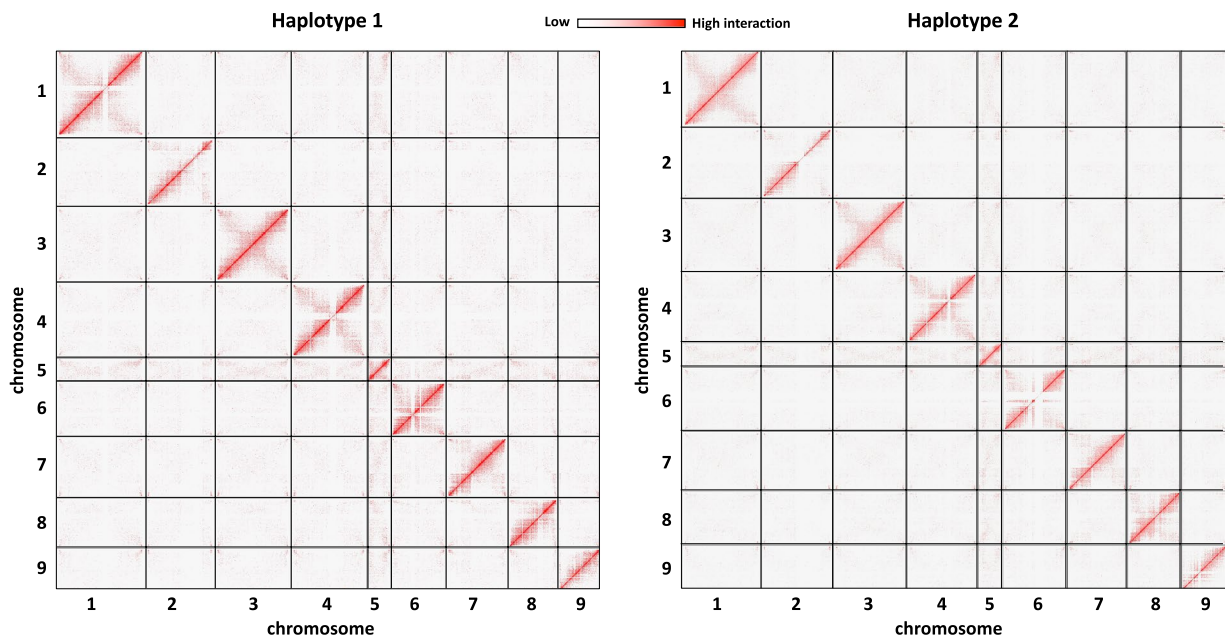


Fig. 3 Inter-chromosomal Hi-C heatmap of two haplotypes.

| Class         | Number of elements | Sequence length (bp) | Percentage of genome |
|---------------|--------------------|----------------------|----------------------|
| LTR           |                    |                      |                      |
| Copia         | 566,348            | 748,049,544          | 27.18%               |
| Gypsy         | 498,678            | 600,515,119          | 21.82%               |
| unknown       | 846,479            | 646,444,484          | 23.49%               |
| TIR           |                    |                      |                      |
| CACTA         | 2,482              | 3,123,318            | 0.11%                |
| Mutator       | 4,129              | 4,120,180            | 0.15%                |
| PIF_Harbinger | 532                | 397,841              | 0.01%                |
| Tc1_Mariner   | 600                | 282,726              | 0.01%                |
| hAT           | 7,229              | 4,337,844            | 0.16%                |
| nonTIR        |                    |                      |                      |
| helitron      | 4,344              | 6,276,603            | 0.23%                |
| Total         | 1,930,821          | 2,013,547,659        | 73.17%               |

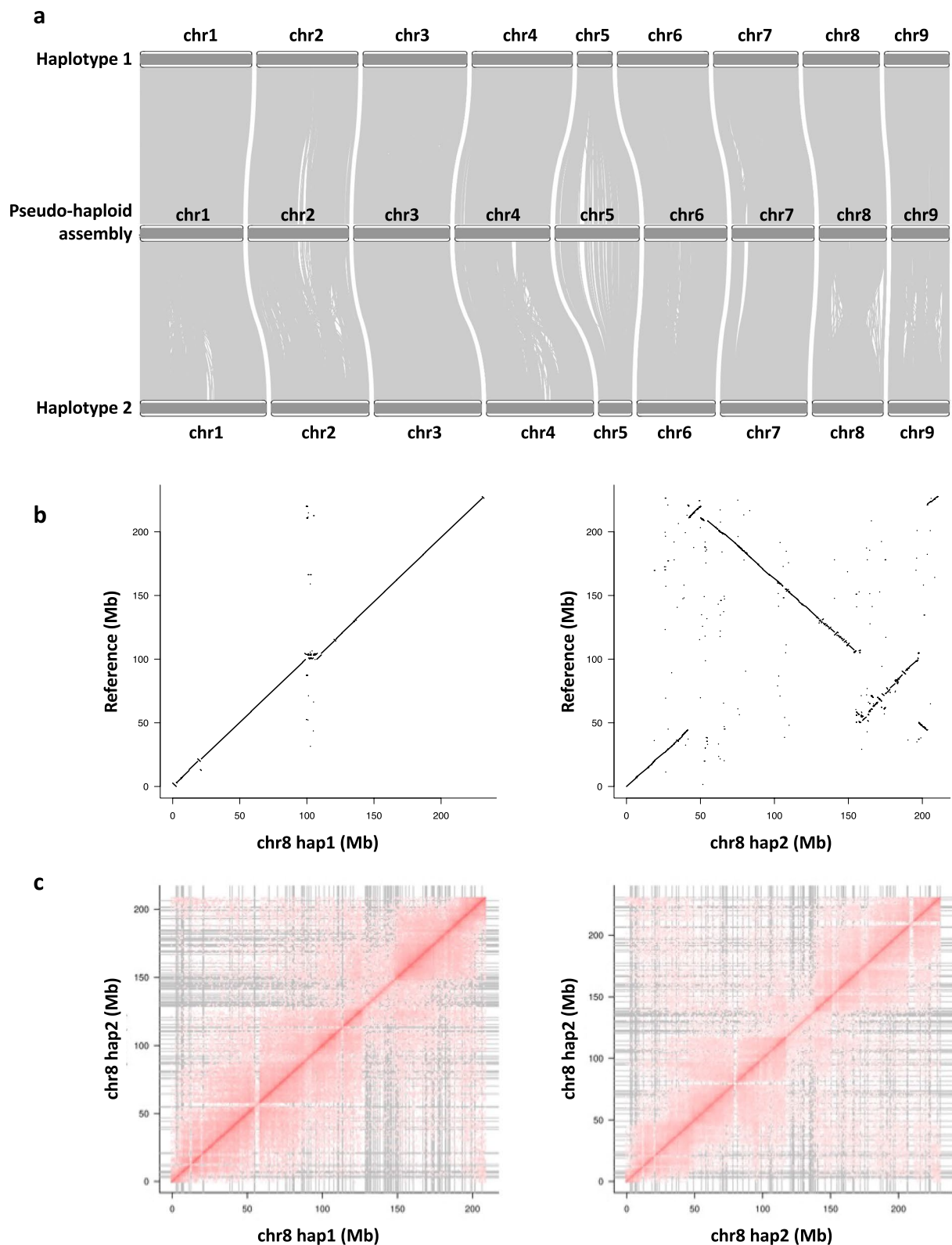
Table 3. Repetitive elements annotation in pseudo-haploid assembly.

|                             |            |
|-----------------------------|------------|
| Number of annotated genes   | 47,820     |
| Total length of genes (bp)  | 80,605,119 |
| Smallest gene length (bp)   | 150        |
| Largest gene length (bp)    | 57,108     |
| Average gene length (bp)    | 1,685      |
| Number of unannotated genes | 167,619    |
| Total number of genes       | 215,439    |

Table 4. Gene annotation in pseudo-haploid assembly.

to 2.01 Gb, were identified as repetitive sequence. The 72.49% of 2.75 Gb which were long terminal repeat (LTR) elements consisted of 27.18% *Copia* elements and 21.82% *Gypsy* elements (Table 3).

To annotate protein-coding genes, we used BRAKER2 with RNA-Seq evidence mode<sup>21</sup>. We downloaded published RNA sequencing data from closely related genera from NCBI. RNA-seq reads of *Chrysanthemum lavandulifolium* (SRR14723013) and *Chrysanthemum nankingense* (SRR18155269, SRR18155270) were adapter trimmed using Trim Galore! (<https://github.com/FelixKrueger/TrimGalore>), aligned to the genome with HISAT2 (v.2.2.1)<sup>22</sup>, and SAM files were converted to BAM using Samtools (v.1.15.1)<sup>23</sup>. Using all the data, BRAKER2 predicted 236,848 genes. Among them, we obtained only the longest isoform with gtfutils (0.5-r234)<sup>24</sup>, resulting in 215,439 genes.



**Fig. 4** Assembly validation. (a) Collinearity between haplotypes and pseudo-haploid genome. (b) Collinearity between chromosome 8 of both haplotypes and the pseudo-haploid assembly. (c) Hi-C heatmap of chromosome 8 for both haplotypes.

We carried out the functional annotation of genes using a homology-based approach with Blast+ (v.2.13.0)<sup>25</sup>. Sequences were compared against the reference proteins (*Arabidopsis thaliana* TAIR10,

TAIR10\_pep\_20101214\_updated.fasta.gz - <https://www.araport.org>) using blastp with an e-value threshold of 1e-10. Among the 215,439 transcripts, 47,820 genes were functionally annotated, having significant homology to proteins in the TAIR database (Table 4).

## Data Records

PacBio and Hi-C raw reads can be accessed in the NCBI Sequence Read Archive under project accession number PRJEB64706. PacBio reads are available under accession numbers ERR11788247<sup>26</sup>, ERR11788248<sup>27</sup>, ERR11788249<sup>28</sup>, and ERR11788250<sup>29</sup>. Hi-C reads have been deposited under accession numbers ERR11788245<sup>30</sup> and ERR11788246<sup>31</sup>. The pseudo-haploid genome assembly sequence has been deposited in the NCBI under the accession number PRJNA1055910 and GenBank accession GCA\_963920955.1<sup>32</sup>. The haplotype-resolved genome assembly sequences have been deposited under accession numbers PRJNA1044733 and PRJNA1044734, with GenBank accessions GCA\_964234995.1<sup>33</sup> and GCA\_964235185.1<sup>34</sup>, respectively. The gene annotation files are available in Figshare (<https://doi.org/10.6084/m9.figshare.27037327>)<sup>35</sup>.

## Technical Validation

The gene space completeness of the genome assembly was estimated at 98.8% (94.0% single-copied genes, 4.8% duplicated genes, 0.3% fragmented, and 0.9% missing genes) as assessed using BUSCO (v.5.4.6)<sup>36</sup> with the 'embryophyta\_odb10' database (n = 1614). The completeness for functionally annotated gene set was 94.6% (88.8% single-copied genes, 5.8% duplicated genes, 2.1% fragmented, and 3.3% missing genes).

We evaluated the haplotype-phased assembly by plotting the collinearity between pseudo-haploid assembly. Both haplotypes were overall collinear with some structural variants (Fig. 4a). For example on chromosome 8, we observed a large inversion about 150 Mb in size with smaller inversion (~50 Mb in size) nested therein (Fig. 4b). We confirmed inversions by inspecting the Hi-C contact matrices (Fig. 4c), which were devoid of the strong off-diagonal signals characteristic of misassemblies.

## Code availability

Programs and options used for data processing are described in the 'Method' section. For pseudo-haploid genome assembly, we followed long-read TRITEX pipeline<sup>9</sup>. For haplotype-phased assembly, we followed the open-source pipeline ([https://github.com/jia-wu-feng/Pan\\_Bulbosum](https://github.com/jia-wu-feng/Pan_Bulbosum))<sup>19</sup>. Assembly code modified for this study is available in <https://github.com/whyonc/chamomile-genome-assembly>.

Received: 23 October 2024; Accepted: 20 February 2025;

Published online: 28 February 2025

## References

1. Franke, R. & Schilcher, H. *Chamomile: Industrial Profiles* (CRC press, 2005).
2. Letchamo, W. Ökologische, genetische und ontogenetische Einflüsse auf Wachstum, Ertrag und Wirkstoffgehalt von diploiden und tetraploiden Kamillen, *Chamomilla recutita* (L.), RAUSCHERT. *Dissertation, Justus-Liebig-Universität Gießen* (1992).
3. Massoud, H. Quantitative Vererbung einiger Ertragsmerkmale und der Hauptkomponenten des ätherischen Öles von Kamille, *Matricaria chamomilla* L. (syn. *Chamomilla recutita* L.). Munich: Dissertation, Technische Universität München (1988).
4. Seidler-Lozykowska, K. Determination of the ploidy level in chamomile (*Chamomilla recutita* (L.) Rausch.) strains rich in alpha-bisabolol. *J. Appl. Genet.* **44**, 151–155 (2003).
5. Otto, L. G., Junghanns, W. R., Plescher, A., Sonnenschein, M. & Sharbel, T. F. Towards breeding of triploid chamomile (*Matricaria recutita* L.)—Ploidy variation within German chamomile of various origins. *Plant Breed.* **134**, 485–493 (2015).
6. Badouin, H. *et al.* The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **546**, 148–152 (2017).
7. Shen, Q. *et al.* The genome of *Artemisia annua* provides insight into the evolution of Asteraceae family and artemisinin biosynthesis. *Mol. Plant* **11**, 776–788 (2018).
8. van Lieshout, N. *et al.* De novo whole-genome assembly of *Chrysanthemum makinoi*, a key wild chrysanthemum. *G3* **12**, jkab358 (2022).
9. Marone, M. P., Singh, H. C., Pozniak, C. J. & Mascher, M. A technical guide to TRITEX, a computational pipeline for chromosome-scale sequence assembly of plant genomes. *Plant Methods* **18**, 128 (2022).
10. Faehrich, B. *et al.* Toward pollen sterility in German chamomile (*Matricaria chamomilla*)—Conventional breeding approaches of cytoplasmic/genic male sterility and chemical emasculation. *Plant Breed.* **139**, 1299–1309 (2020).
11. Otto, L. G. *et al.* Towards breeding of triploid chamomile (*Matricaria recutita* L.)—Ploidy variation within German chamomile of various origins. *Plant Breed.* **134**, 485–493 (2015).
12. Jayakodi, M. *et al.* Structural variation in the pangenome of wild and domesticated barley. *Nature*. 1–9 (2024).
13. Padmarasu, S., Himmelbach, A., Mascher, M. & Stein, N. *In situ* Hi-C for plants: an improved method to detect long-range chromatin interactions. In: *Plant Long Non-Coding RNAs: Methods and Protocols* (eds. Chekanova, J. A. & Wang, H.-L. V.) vol. 1933 441–472 (Springer New York, 2019).
14. Himmelbach, A. *et al.* Discovery of multi-megabase polymorphic inversions by chromosome conformation capture sequencing in large-genome plant species. *Plant J.* **96**, 1309–1316 (2018).
15. Himmelbach, A., Walde, I., Mascher, M. & Stein, N. Tethered chromosome conformation capture sequencing in Triticeae: a valuable tool for genome assembly. *Bio Protoc.* **8**, e2955 (2018).
16. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
17. Ranallo-Benavidez, T. R. *et al.* GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature communications* **11**, 1432 (2020).
18. Astashyn, A. *et al.* Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome biology* **25**, 60 (2024).
19. Mascher, M. *et al.* A haplotype-resolved pangenome of the barley wild relative *Hordeum bulbosum*. Preprint at <https://doi.org/10.21203/rs.3.rs-3916840/v1> (2024).
20. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 1–18 (2019).
21. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom. Bioinform.* **3**, lqaa108 (2021).

22. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
23. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
24. Li, H.-D., Lin, C.-X. & Zheng, J. GTFtools: a software package for analyzing various features of gene models. *Bioinformatics* **38**, 4806–4808 (2022).
25. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
26. EMBL-EBI European Nucleotide Archive <https://www.ebi.ac.uk/ena/browser/view/ERR11788247> (2024).
27. EMBL-EBI European Nucleotide Archive <https://www.ebi.ac.uk/ena/browser/view/ERR11788248> (2024).
28. EMBL-EBI European Nucleotide Archive <https://www.ebi.ac.uk/ena/browser/view/ERR11788249> (2024).
29. EMBL-EBI European Nucleotide Archive <https://www.ebi.ac.uk/ena/browser/view/ERR11788250> (2024).
30. EMBL-EBI European Nucleotide Archive <https://www.ebi.ac.uk/ena/browser/view/ERR11788245> (2024).
31. EMBL-EBI European Nucleotide Archive <https://www.ebi.ac.uk/ena/browser/view/ERR11788246> (2024).
32. NCBI GenBank [https://identifiers.org/ncbi/insdc.gca:GCA\\_963920955.1](https://identifiers.org/ncbi/insdc.gca:GCA_963920955.1) (2024).
33. NCBI GenBank [https://identifiers.org/ncbi/insdc.gca:GCA\\_964234995.1](https://identifiers.org/ncbi/insdc.gca:GCA_964234995.1) (2024).
34. NCBI GenBank [https://identifiers.org/ncbi/insdc.gca:GCA\\_964235185.1](https://identifiers.org/ncbi/insdc.gca:GCA_964235185.1) (2024).
35. Cho, W. Gene annotation files of *Matricaria chamomilla*. *figshare* <https://doi.org/10.6084/m9.figshare.27037327> (2024).
36. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).

## Acknowledgements

Parts of this work were financed based upon a resolution of the Federal Parliament of Germany (Deutscher Bundestag) by the Bundesministerium für Ernährung und, Landwirtschaft (BMEL, support code 22008718) via the Fachagentur Nachwachsende Rohstoffe e.V. (FNR) as the project executing organisation for the funding programme “Renewable Resources”. The funding body did not exert any influence on the study nor the manuscript.

## Author contributions

Martin Mascher and Lars-Gernot Otto designed and supervised the study. Woohyeon Cho analyzed data and prepared figures. Martin Mascher and Jiawu Feng wrote code. Woohyeon Cho, Martin Mascher, and Lars-Gernot Otto wrote the paper. Manuela Knauf provided technical assistance. Sebastian Albrecht performed the *in vitro* cultivation and propagation of the sequenced chamomile genotype. Axel Himmelbach generated sequencing data.

## Funding

Open Access funding enabled and organized by Projekt DEAL.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to M.M.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025