bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 A high-quality functional genome assembly of *Delia radicum* L. (Diptera: Anthomiidae) annotated from
2 egg to adult
3
4 Short running title (max. 45 characters): A high-quality genome of the cabbage root fly
5
6 Rebekka Sontowski[1,2, #], Yvonne Poeschl[2,3,4, #], Yu Okamura[5], Heiko Vogel[5], Cervin Guyomar[3,6], Anne-
7 Marie Cortesero[7], Nicole M. van Dam[1,2] *
8
9 1. Molecular Interaction Ecology, German Centre for Integrative Biodiversity Research (iDiv) Halle-
10 Jena-Leipzig, Puschstraße 4, 04103 Leipzig, Germany
11 2. Institute of Biodiversity, Friedrich Schiller University Jena, Dornburger-Str. 159, 07743 Jena,
12 Germany
13 3. Bioinformatics Unit, German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-
14 Leipzig, Puschstraße 4, 04103 Leipzig, Germany
15 4. Institute of Computer Science, Martin Luther University Halle-Wittenberg, Von-Seckendorff-Platz
16 1, 06120 Halle, Germany
17 5. Department of Insect Symbiosis, Max Planck Institute for Chemical Ecology, Hans-Knöll-Str. 8,
18 07745 Jena, Germany
19 6. GenPhySE, Université de Toulouse, INRAE, ENVT, 31326, Castanet Tolosan, France
20 7 IGEPP, INRAE, Institut Agro, Univ Rennes, 35000, Rennes, France
21
22 # equal first authors
23 *corresponding author: Nicole M. van Dam, nicole.vandam@idiv.de
24
25
26
27
28

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

29 Abstract

30 Belowground herbivores are overseen and underestimated, even though they can cause significant
31 economic losses in agriculture. The cabbage root fly *Delia radicum* (Anthomyiidae) is a common pest in
32 *Brassica* species, including agriculturally important crops, such as oil seed rape. The damage is caused
33 by the larvae, which feed specifically on the taproots of *Brassica* plants until they pupate. The adults are
34 aboveground-living generalists feeding on pollen and nectar. Female flies are attracted by chemical cues
35 in *Brassica* plants for oviposition. An assembled and annotated genome can elucidate which genetic
36 mechanisms underlie the adaptation of *D. radicum* to its host plants and their specific chemical defenses,
37 in particular isothiocyanates. Therefore, we assembled, annotated and analyzed the *D. radicum* genome
38 using a combination of different Next Generation Sequencing and bioinformatic approaches. We
39 assembled a chromosome-level *D. radicum* genome using PacBio and Hi-C Illumina sequence data.
40 Combining Canu and 3D-DNA genome assembler, we constructed a 1.3 Gbp genome with an N50 of
41 242 Mbp and 6 pseudo-chromosomes. To annotate the assembled *D. radicum* genome, we combined
42 homology-, transcriptome- and *ab initio*-prediction approaches. In total, we annotated 13,618 genes that
43 were predicted by at least two approaches. We analyzed egg, larval, pupal and adult transcriptomes in
44 relation to life-stage specific molecular functions. This high-quality annotated genome of *D. radicum* is a
45 first step to understanding the genetic mechanisms underlying host plant adaptation. As such, it will be
46 an important resource to find novel and sustainable approaches to reduce crop losses to these pests.

47

48 Keywords: belowground pest, herbivory, insects, chromosome-scale genome, de novo genome
49 assembly, functional gene annotation

50
51
52
53
54
55
56

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

57 ## 1. INTRODUCTION

58 The cabbage root fly, *Delia radicum* L. (Diptera; Anthomyiidae), is a severe pest in agriculture. The family
59 Anthomyiidae, or flower flies, is a large family mainly occurring in the northern hemisphere. Adult *D.*
60 *radicum* flies live aboveground and feed on nectar (Figure 1, (Gouinguene & Städler, 2005; Peter
61 Roessingh & Städler, 1990)). The females oviposit next to or on the root crown of brassicaceous plants.
62 After the eggs have hatched, the larvae occupy a new habitat and move into the soil to mine into the tap
63 roots (Figure 1). After passing through three instars in about 20 days, the larvae move back to the soil
64 to pupate (Capinera, 2008).

65 As its common name "cabbage root fly" already indicates, *D. radicum* is a specialized herbivore on
66 Brassicaceae, the cabbage and mustard family. This plant family contains several agriculturally important
67 crops, such as broccoli, turnip, Pak Choi and rapeseed. Although they are specialists on Brassicaceae,
68 females prefer some plant species of this family more for oviposition than others (Lamy et al., 2018). The
69 female flies are attracted to the plant by specific volatile organic compounds, such as sulfides and
70 terpenes (Ferry et al., 2007; Kergunteuil, Dugravot, Danner, van Dam, & Cortesero, 2015). Upon
71 contacting the plants, the females decide to oviposit based on chemical cues, in particular the presence
72 of glucosinolates (Gouinguené & Städler, 2006). The larvae are well adapted to deal with the
73 glucosinolate-myrosinase defense system that is specific to the Brassicaceae (Hopkins, van Dam, & van
74 Loon, 2009). Glucosinolates are sulfur-containing glycosylated compounds, which are stored in the
75 vacuoles of cells localized between the endodermis and phloem cells (Kissen, Rossiter, & Bones, 2009).
76 The roots of Brassica species contain high levels of glucosinolates, in particular 2-
77 phenylethylglucosinolate (van Dam, Tytgat, & Kirkegaard, 2009). Glucosinolates can be converted by
78 the enzyme myrosinase into pungent and toxic products, such as isothiocyanates (ITCs) and nitriles
79 which deter generalist herbivores (Kissen et al., 2009). The myrosinase enzymes are stored in so called
80 myrosin cells (Kissen et al., 2009). Upon tissue damage, either by mechanical damage or by herbivores,
81 such as *D. radicum* larvae, the glucosinolates and myrosinases mix. This results in the formation of
82 various conversion products, including ITCs, nitriles and sulfides (Crespo et al., 2012; Danner et al.,
83 2015; Wittstock & Gershenzon, 2002).

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

84 Indeed, *D. radicum* larvae can successfully infest the roots of a wide range of Brassicaceae (Finch &
85 Ackley, 1977; Tsunoda, Krosse, & van Dam, 2017). The damage the feeding larvae cause leads to
86 substantial fitness loss in wild plants and yield reduction in crops. In rapeseed, *D. radicum* infestation
87 reduces seed numbers and seed weight (Griffiths, 1991; McDonald & Sears, 1992). The annual
88 economic losses due to *D. radium* infestation in Western Europe and Northern America are estimated to
89 be 100 million $ (Wang, Voorrips, Steenhuis-Broers, Vosman, & van Loon, 2016).

90 Controlling *D. radicum* in agriculture is a major challenge. Natural resistance to this specialist herbivore
91 has not been identified in currently used cultivars yet (Ekuere et al., 2005) and several effective synthetic
92 insecticides, such as neonicotinoids, have been banned from use due to environmental concerns (Allema
93 B, Hoogendoorn M, van Beek J, & P, 2017). Moreover, pesticide resistance has already developed in
94 this species, for example against chlorpyrifos (van Herk et al., 2016). Alternative and more sustainable
95 pest management strategies are urgently needed. Heritable natural resistance to *D. radicum* is present
96 in wild brassicaceous species, but introgression of these traits may be hampered by crossing barriers
97 and linkage of resistance with undesired traits (Ekuere et al., 2005; Wang et al., 2016). Several studies
98 examined the application of entomopathogenic fungi, natural predators or parasitoids, mixed cropping
99 and soil microbes to better control *D. radicum* (Bruck, Snelling, Dreves, & Jaronski, 2005; Dixon, Coady,
100 Larson, & Spaner, 2004; Fournet, Stapel, Kacem, Nenon, & Brunel, 2000; Kapranas, Sbaiti, Degen, &
101 Turlings, 2020; Lachaise et al., 2017; Neveu, Krespi, Kacem, & Nénon, 2000). Even though each of
102 these measures may reduce *D. radicum* infestations, they cannot prevent yield loss as effectively as
103 synthetic pesticides.

104 To understand the interaction of *D. radicum* with its host plants, the chemical ecology of this plant-
105 herbivore interaction has been intensively studied over the last decades. These studies analyzed aspects
106 ranging from the chemosensory mechanisms of host plant attraction and oviposition choice to herbivore-
107 induced plant responses and interactions with predators and parasitoids (Ferry et al., 2007; Gouinguene
108 & Städler, 2005; Hopkins et al., 2009; Kergunteuil et al., 2015; P Roessingh et al., 1992). However, the
109 genetic mechanisms underpinning host-plant adaptation of *D. radicum* are unknown. An accurate and
110 well-annotated genome can reveal genetic mechanisms underpinning adaptation of *D. radicum* to its

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

111 host's chemical defenses. Especially, understanding the preference of the different agricultural relevant
112 life stages (adults and larvae) which occur in separate habitats (above- and belowground) on the genetic
113 level expands our understanding of herbivore-plant interactions. These mechanisms can also be an
114 important starting point to develop novel approaches, such as species-specific dsRNA-based pest
115 control strategies. So far, a genome of this species has not been published.
116 Here, we assembled and annotated a *de novo*, chromosome-level scaffolded genome of *D. radicum*
117 using PacBio and Hi-C Illumina sequencing. We used three different approaches to annotate the
118 genome; Cufflinks, which uses transcript assembly; GeMoMa, which is homology-based, and BRAKER,
119 for additional prediction of genes not covered by the first two methods. Generated RNASeq data of all
120 four life stages (eggs, larvae, pupae, adults) and two relevant stress factors (heat stress in adults, plant
121 toxin stress in larvae), allowed us to validate predicted genes and to identify specific gene families which
122 were expressed in each of the life stages.
123

124 ## 2. MATERIALS AND METHODS

125 ### 2.1 Sample material

126 A starting culture of *D. radicum* was provided by Anne-Marie Cortesero, University of Rennes, France in
127 2014. It originated from pupae collected in a cabbage field in Brittany, France, (48°6′31″ N, 1°47′1″ W)
128 the same year. More than five thousand pupae were collected to start the original culture and ~50
129 individuals from this culture were sent to iDiv. A permanent culture was established in our lab under
130 constant conditions (20 ± 2°C, 85 ± 10 % RH, 16L:8D) in a controlled environment cabinet (Percival
131 Scientific, Perry, Iowa, USA) resulting in an inbreeding line of over 60 generations. Adult flies were reared
132 in a net cage and fed with a 1:1 milk powder-yeast mixture and a water-honey solution, which was
133 changed three times a week. Water was provided ad libitum. Eggs were placed in a 10x10x10 cm plastic
134 box filled with 2 cm moistened, autoclaved sand and a piece of turnip. Once the larvae hatched, old
135 turnip pieces were removed and exchanged with new turnip every other day and the sand was moistened
136 when necessary. After the third instar, the larvae crawled into the sand, where they pupated. The pupae
137 were collected by flooding the box with water, collecting the floating pupae and placing them into the
138 adult fly cage until eclosion.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

139 Species identification was performed using a fragment of the cytochrome oxidase I (COI) gene as a
140 molecular marker generated with the universal COI primer pair HCO and LCO (Folmer et al. 1994). The
141 sequence was submitted to BLAST online using the BLASTn algorithm (Retrieved from
142 https://blast.ncbi.nlm.nih.gov/Blast.cg). The top three hits matched with *D. radicum* COI accessions
143 (MG115888.1, HQ581775.1, GU806605.1) with an identity of more than 98.45 %.
144

145 ## 2.2 Genome sequencing

146 ### 2.2.1 Sampling, DNA extraction and PacBio sequencing

147 For PacBio sequencing, 18 randomly collected, fully matured *D. radicum* adults were frozen and stored
148 at -80°C. To sterilize the surface, the flies were incubated for 2 min in bleach (2 %), transferred to sodium
149 thiosulfate (0.1 N) for neutralization, washed three times in 70 % ethanol and once in autoclaved dd
150 water. To reduce contamination by microorganisms from the gut, we extracted total DNA from the head
151 and the thorax of the adults, using a phenol-chloroform extraction method according to the protocol of
152 the sequencing facility (Figure S1). We pooled three individuals per extraction and checked the DNA
153 quality using gel electrophoresis (0.7 % agarose gel). DNA purity was assessed using a
154 NanoPhotometer® P330 (Implen, Munich/Germany) and DNA quantity using a Qubit dsDNA BR assay
155 kit in combination with a Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA/USA). The DNA of all samples
156 was pooled for the sequencing library. Library preparation and sequencing was provided by the facility
157 of the Max Planck Institute of Molecular Cell Biology and Genetics, Dresden/Germany on a PacBio
158 Sequel. A total of 16 SMRT cells were processed and 6,539,960 reads generated. Due to the pooling of
159 several females and males, we expected the reads to be highly heterozygous, which we considered for
160 the assembly process.

161 ### 2.2.2 Sampling, DNA extraction and Hi-C Illumina sequencing

162 For Hi-C Illumina sequencing fresh *D. radicum* pupae from the above culture were randomly selected. A
163 total of 10 pupae were chopped into small pieces with a razor and resuspended in 3 ml of PBS with 1%
164 formaldehyde. The homogenized sample was incubated at RT for 20 min with periodic mixing. Glycine

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

165 was added to the sample buffer to 125 mM final concentration and incubated at RT for ~15 min with
166 periodic mixing. The homogenized tissue was spun down (1000 x g for 1 min), rinsed twice with PBS,
167 and pelleted (1000 x g for 2 min). After removal of the supernatant, the tissue was homogenized to a fine
168 powder in a liquid nitrogen-chilled mortar with a chilled pestle. Further sample processing and
169 sequencing was performed by Phase Genomics (Seattle, WA, USA) on an Illumina HiSeq 4000,
170 generating a total of 181,752,938 paired end reads (2 x 150 bp).

171 ## 2.3 Genome size estimation

172 A karyotyping study determined that *D. radicum* is a diploid organism with 2n = 12 chromosomes
173 (Hartman & Southern, 1995). To obtain a reliable estimate of the *D. radicum* genome size, we used flow
174 cytometry based on a method using propidium iodide-stained nuclei (Spencer Johnston Lab, Texas
175 A&M, USA; (Hare & Johnston, 2011). The haploid genome sizes were estimated to be 1239.0 ± 27.5
176 Mbp for females (N = 4) and 1218.0 ± 4.0 Mbp for males (N = 50).

177 ## 2.4 Genome assembly and completeness

178 ### 2.4.1 PacBio data processing

179 Raw PacBio reads in bam file format were converted into fasta files by using samtools (version 1.3.1) (Li
180 et al., 2009) as part of the SMRT link software (version 5.1.0, https://www.pacb.com/support/software-
181 downloads/). Extracted raw PacBio reads (6,539,960 reads) were checked for potential contaminations
182 with prokaryotic DNA by applying EukRep (version 0.6.2) (West, Probst, Grigoriev, Thomas, & Banfield,
183 2018) with default parameter settings. Only reads classified as eukaryotic (4,454,601 reads) were used
184 for the *de novo* genome assembly.

185 ### 2.4.2 *De novo* genome assembly

186 The long-read assembler Canu (version 1.9) (Koren et al., 2017) was used to generate a *de novo*
187 genome assembly from contamination-free PacBio reads. The Canu pipeline, including read error
188 correction and assembly, was started with setting parameters based on the estimated genome size
189 (genomeSize=1200m) and the use of not too short (minReadLength=5000) and high quality

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

190 (stopOnReadQuality=true) PacBio reads, addressing the overlapping of sequences
191 (minOverlapLength=1000 corOutCoverage=200), and accounting for the expected high heterozygosity
192 rate of the *D. radicum* genome (batOptions=-dg 3 -db 3 -dr 1 -ca 500 -cp 50). The latter parameters were
193 selected to prevent the haplotypes from being collapsed during the assembly process.

194 ### 2.4.3 Polishing and purging

195 To improve the sequence quality of the raw genome assembly, we performed two rounds of polishing.
196 All eukaryotic raw PacBio reads were aligned with pbalign (version 0.3.1 and default parameter settings)
197 and these results were used for sequence polishing with Arrow (version 2.2.2 and default parameter
198 settings). Both programs are part of the SMRT link software (version 5.1.0,
199 https://www.pacb.com/support/software-downloads/). To detect and remove duplications in the
200 assembled contigs, we applied purge_dups (version 1.2.3) (Guan et al., 2020) on the polished assembly.
201 We ran the first three steps of the purge_dups pipeline with default parameters and the last step with
202 additional setting "-e -c" to allow only clipping at the end of contigs and retaining high coverage contigs.

203 ### 2.4.4 Chromosome-scale scaffolding

204 Hi-C Illumina reads were aligned to the purged assembly with the Juicer pipeline incorporating
205 juicer_tools (version 1.22.01) (Durand et al., 2016), "-s DpnII" and a restriction site file (generated with
206 the generate_site_positions.py script contained in juicer) provided by "-z" option. The sequences of the
207 purged assembly were scaffolded with the Juicer output on Hi-C read alignments into chromosome-scale
208 super-scaffolds applying the 3D-DNA genome assembler (version 18011) (Dudchenko et al., 2017) with
209 the additional setting of "--splitter-coarse-stringency 30 --gap-size 100". This resulted in the final genome
210 assembly from *D. radicum*.

211 ### 2.4.5 Evaluating genome completeness

212 We used BUSCO v4 (4.0.5) (Seppey, Manni, & Zdobnov, 2019) to analyze the completeness of the final
213 and intermediate genome assemblies. Three different gene sets, insecta_odb10.2019-11-20,
214 endopterygota_odb10.2019-11-20, and diptera_odb10.2019-11-20, representing different levels in

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

215 evolutionary relatedness were considered in the evaluation process. These three gene sets comprise
216 1367, 2124, or 3285 orthologous genes, respectively.

217 #### 2.4.6 Exclusion of Hi-C scaffolds

218 While assembling the *D. radicum* genome we co-assembled the complete genome of *Wolbachia* (Hi-C
219 scaffold 7) a common endosymbiont in arthropods. To obtain a contamination-free final assembly, we
220 excluded Hi-C scaffolds 7, 146 and 370 and trimmed Hi-C scaffold 6 after position 12,881,041 that were
221 annotated to be contaminated with *Wolbachia* sequences during the NCBI validation process.

222 ### 2.5 Phylogeny - Comparative genomics based on BUSCOs

223 Phylogenetic analyses were done with BuscoOrthoPhylo
224 (https://github.com/PlantDr430/BuscoOrthoPhylo) which is a wrapper script to concatenate and align
225 protein sequences, and to construct a phylogenetic tree based on single-copy BUSCO genes. BUSCOs
226 of the endopterygota_odb10.2019-11-20 gene set, consisting of 2124 genes, were used as the basis for
227 the analysis. In the initial phase complete single-copy BUSCO genes which were shared by 10 selected
228 species (Table 1), were computed. Protein sequences of the shared genes were extracted and
229 concatenated for each species. MAFFT aligner (version 7.475) (Katoh, Misawa, Kuma, & Miyata, 2002)
230 was run on concatenated FASTA file(s) and finally RAxML (version 8.2.12) (Stamatakis, 2014) with "-
231 rx_p_sub PROTGAMMAWAG" as model and "-b 100" bootstrap steps was used to reconstruct the
232 phylogenetic tree. The resulting findings were visualized in a phylogenetic tree using Phylo.io (Robinson,
233 Dylus, & Dessimoz, 2016).

234 ### 2.6 Sampling, RNA extraction and Transcriptome sequencing

235 All life stage samples were collected from the laboratory culture (section 2.1). We used three replicates
236 per life stage and condition. For the egg stage, we collected 25 mg eggs (laid within 24 h) per replicate.
237 For the larval stage, we collected 18 randomly selected second instar larvae. Nine of the selected larvae
238 were fed on a semi-artificial diet, containing yeast, milk powder, freeze dried turnip, agar (2:2:2:1) and
239 90% water. The other nine larvae were reared on the same diet containing 0.4 mg phenylethyl

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

240 isothiocyanate/g diet. All larvae received freshly prepared diet every other day. After 7 days, larvae were
241 shock-frozen at -80°C and pooled into batches containing 3 larvae forming three biological replicates per
242 treatment. For the pupal stage, we randomly selected nine freshly formed pupae and pooled them into
243 three biological replicates of three pupae each. For the adult stage, we collected 18 fully developed
244 random adults. Nine individuals were exposed to 35°C (Michaud, Marin, Westwood, & Tanguay, 1997)
245 for 2 hours, whereas the control adults were kept under normal conditions. We pooled three adults for
246 one replicate, resulting in three replicates for control and elevated temperature treatment.
247 All samples were surface sterilized using the same procedure as described for the adult flies. We
248 extracted the total RNA of the larval stage using ReliaPrep RNA Tissue Miniprep kit (Promega, Madison,
249 WI/USA) according to the supplier's recommended protocol. Total RNA of all further samples was
250 extracted using TRIzol (Life technologies, Carlsbad, CA/USA) according to the supplier's recommended
251 protocol. Qualitative and quantitative RNA assessment of all samples was done by gel electrophoresis
252 (1% agarose), a NanoPhotometer® P330 (Implen, Munich/Germany) and a Qubit 2.0 (Invitrogen,
253 Carlsbad, CA/USA).
254 Library preparation and sequencing of the larval samples (control and stressed larvae) were performed
255 by the Deep Sequencing group of Biotech TU Dresden/Germany on an Illumina NextSeq next generation
256 sequencer. The poly(A) enriched strand specific libraries generated for all samples ran on one lane
257 generating in total 400 Mio of 75 bp paired end reads. Egg, pupal and adult (control and stressed)
258 samples were sequenced by Novogene (Hong Kong/China) with strand specific library preparations and
259 sequencing on an Illumina NovaSeq 6000 next generation sequencer, generating 20 Mio paired-end (2
260 x 150 bp) reads per sample.
261

262 2.7 Genome annotation – prediction of protein-coding genes

263 2.7.1 Mapping of Transcriptome data

264 Including RNASeq data can improve the quality of gene predictions as optional input by several gene
265 prediction algorithms. We mapped the *D. radicum* RNASeq data of the 18 samples, consisting of six

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

conditions (4 life stages and 2 stress treatments) with three replicates each to the *D. radicum* genome with STAR (version 020201) (Dobin et al., 2012) and store mapping results in bam files.

2.7.2 Homology-based gene prediction

Homology-based GeMoMa (version 1.6.4 und 1.7.2) (Keilwagen, Hartung, Paulini, Twardziok, & Grau, 2018; Keilwagen et al., 2016) gene predictions the *D. radicum* genome were performed based on the annotated genomes of four Diptera species (*Anopheles gambiae*, *Drosophila melanogaster*, *Lucilia cuprina*, and *Musca domestica*), four Lepidoptera species (*Manduca sexta*, *Pieris rapae*, *Plutella xylostella*, and *Spodoptera litura*), and one Coleoptera species (*Tribolium castaneum*) obtained from NCBI (Table 1). For each of these nine species, extracted CDS were aligned with MMseqs2 (version 11.e1a1c) (Steinegger & Söding, 2017) to the *D. radicum* genome sequence with parameter values suggested by GeMoMa. Alignments and RNASeq mappings were used for predictions of gene models in the genome with GeMoMa and default parameters, separately for each species and by incorporating mapped RNASeq data for refining intron boundaries. The resulting nine gene annotation sets were filtered and merged using the GeMoMaAnnotationFilter (GAF) with "f="start=='M' and stop=='*'" atf=""". Only transcripts of genes starting with the start codon “M(ethionine)” and ending with a stop codon “*” were considered and all isoforms were retained. We finally predicted and added UTR annotations to the resulting filtered set of transcripts by using the GeneAnnotationFinalizer with "u=YES rename=NO", which is also part of the GeMoMa suite.

2.7.3 Transcriptome assembly - RNA-Seq-based gene predictions

To assemble one transcriptome per life stage and condition, we merged the read mappings (bam files) of the three replicates per condition and life stage. For the transcriptome assembly of the mapped RNASeq data, we used Cufflinks (version 2.2.1) (Trapnell et al., 2010). Initially, soft-clipped read mappings were clipped, and assembled to six transcriptomes using Cufflinks with default parameters and "-fr-firstrand". The resulting six transcriptomes were subjected to Cuffmerge, which is part of the Cufflinks toolbox, to generate a single master transcriptome. While Cufflinks assembled transcripts with exon annotation, missing coding regions and UTRs were identified with TransDecoder (version 5.5.0,

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

292 https://github.com/TransDecoder/TransDecoder). Finally, predicted transcripts were filtered for a proper
293 start and end of protein coding transcripts and retaining the UTR annotations by applying the GAF with
294 parameters “f="start=='M' and stop=='*'" atf='' aat=true tf=true”.

295 #### 2.7.4 *Ab initio* gene prediction

296 Additionally, we aimed to predict genes not covered by the homology-based and the transcriptome-
297 based approach, due to a lack of homology or because of no or low expression under the specific
298 conditions of the sampled life stages. To obtain such *ab initio* gene predictions, we ran RepeatMasker
299 (version 4.1.0, http://www.repeatmasker.org) with RMBlast (version 2.10.0,
300 http://www.repeatmasker.org/RMBlast.html) and "-species insecta -gff -xsmall" to find and mask
301 repetitive sequences annotated for insects in the RepeatMasker repeat database. For *ab initio* prediction
302 of protein-coding genes on the masked genome sequences, we ran BRAKER (version 1.9) (Brůna, Hoff,
303 Lomsadze, Stanke, & Borodovsky, 2021; Katharina J. Hoff, Lange, Lomsadze, Borodovsky, & Stanke,
304 2015; Katharina J Hoff, Lomsadze, Borodovsky, & Stanke, 2019), which combines GeneMark (version
305 4.59_lic) (Lomsadze, Burns, & Borodovsky, 2014) and AUGUSTUS (version 3.4.0) (Stanke, Schöffmann,
306 Morgenstern, & Waack, 2006) with "--gff3 --softmasking" and provided the mapped RNASeq data as
307 hints for initial training of gene models and gene predictions.

308 Predicted transcripts were filtered for proper start and end of protein coding transcripts by applying the
309 GAF with "f="start=='M' and stop=='*'" atf=''". Finally, UTR annotations were predicted and added using
310 GeneAnnotationFinalizer with "u=YES rename=NO”.

311 #### 2.7.5 Final genome annotation and completeness evaluation

312 We ran GeMoMa`s GeMoMaAnnotationFilter (GAF) with "f='' atf='' tr=true aat=true" to integrate the
313 predicted gene models from all three applied approaches, the homology-based, the RNASeq-based and
314 *ab initio* gene prediction approach, and yield a master gene annotation file for the *D. radicum* genome.
315 As gene-related features we include mRNA, CDS, five_prime_UTR, and three_prime_UTR specificities

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

316 in the annotation file and several attributes that gave additional information on the predicted transcripts
317 and can be used for user-specific filtering.

318 We evaluated the completeness of the final set of protein-coding genes with BUSCO v4 as we did for
319 the evaluation of the completeness of the genome (section 2.4.5), but set "-m proteins".

320 2.8 Functional annotation

321 Predicted *D. radicum* protein sequences were subjected to PANNZER2 (Protein ANNotation with Z-
322 scoRE) (Törönen, Medlar, & Holm, 2018), which predicts functional descriptions and GO classes.
323 Additionally, extracted protein sequences were subjected to InterProScan (version 5.45-80.0.) (Blum et
324 al., 2020; Jones et al., 2014) and scanned for information on protein family and domains in all member
325 data bases (-appl CDD, HAMAP, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITEPATTERNS,
326 SMART, TIGRFAM, Gene3D, SFLD, SUPERFAMILY, MobiDBLite) and for GO- or pathway annotation
327 ("-goterms -iprlookup -pa").

328 GO terms annotated for transcripts with PANNZER2 and InterProScan were merged. Additionally, to get
329 a functional annotation per gene, we merged the annotations of all respective transcripts.

330 2.9 Synteny analysis

331 Annotated CDSs of *D. melanogaster* (Table 1) were extracted and aligned to the *D. radicum* genome
332 with MMseqs2 (version 11.e1a1c) (Steinegger & Söding, 2017). Alignments were used for homology-
333 based predictions of gene models in the *D. radicum* genome with GeMoMa (version 1.6.4) (Keilwagen
334 et al., 2018) with default parameters. Predicted gene models were filtered using the
335 GeMoMaAnnotationFilter (GAF) with "f="start=='M' and stop=='*'" atf=""". Finally, a table containing the
336 relation and positions of the gene models was generated with SyntenyChecker, which is part of the
337 GeMoMa toolbox. Syntenic relationships of *D. radicum* to *D. melanogester* were visualized using Circos
338 (version 0.69-9) (Krzywinski et al., 2009).

339 2.10 Analysis of life cycle data

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

340 We extracted the sequences of all annotated transcripts and quantified their abundances with kallisto
341 (version 0.46.1, (Bray, Pimentel, Melsted, & Pachter, 2016)) with "-b 100" bootstraps and "—rf-stranded".
342 The abundances were imported into the statistical framework R (version 3.6.2) (R Core Team, 2020) for
343 further analyses using the R package tximport (1.14.2) (Soneson, Love, & Robinson, 2015). Using
344 tximport transcript-level, estimates for abundances were summarized for further gene-level analyses.

345 We denoted a gene as *expressed* if it had a TPM (transcript per million) value ≥ 1 in at least one of the
346 18 transcriptome samples. We refer to this set of genes as the "data set of expressed genes". We called
347 a gene present in a life stage or condition if it occurred in at least one replicate. This aggregation resulted
348 in a matrix with six columns (four life stages, two conditions). These six sets were analyzed for life stage
349 and condition-specific gene expression and also for intersections with the R package UpSetR (1.4.0)
350 (Gehlenborg, 2019).

351 We performed Gene Ontology (GO) analyses of pre-defined gene sets using R (version 4.0.4) with the
352 latest version of the R package topGO (2.42.0) (Alexa & Rahnenfuhrer, 2020) with GO.db (3.12.1)
353 (Carlson, Falcon, Pages, & Li, 2020). We used Fisher's exact test to identify over-represented GO terms.
354 Raw p-values were corrected for multiple testing using the method proposed by Benjamini and Yekutieli
355 (Benjamini & Yekutieli, 2001) implemented in p.adjust contained in the basic R package stats. To get an
356 indication of which processes were active, we aggregated single significant GO-terms (adjusted p value
357 < 0.05) into self-assigned generic categories. Results were visualized using the R package pheatmap
358 (1.0.12) (Kolde, 2019). For visualization of the results for generic categories, we computed the relative
359 frequency of GO terms determined in a pre-defined gene set for a generic category. The relative
360 frequency was calculated by the number of significant GO terms in a gene set divided by the total number
361 of GO terms that were sorted into the appropriate generic category.

362 We defined six gene sets for life stage (eggs, larva, pupa, adults) and condition-specific GO analysis
363 (ITC, heat stress). For the analysis of the whole life cycle we determined genes that were exclusively
364 expressed in one of the four life stages under control conditions. As we have additional stress conditions
365 in the larval and the adult life stage, we extended the defined gene sets for these two life stages by

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

366 genes contained in the intersection of both conditions (control and stress) within these stages. For
367 condition-specific GO analysis, we additionally determined the genes exclusively expressed in the
368 stressed condition of the larval and adult stage, respectively. Again, we also extended the stress-specific
369 genes sets by the respective intersection gene set.

370 We clustered samples and genes contained in the data of expressed genes using the R package umap
371 (0.2.7.0) (Konopka, 2020). UMAP (uniform manifold approximation and projection) is a technique to
372 reduce dimensions and bring similar data vectors, samples (columns) or genes (rows) in close proximity.
373 In our analyses we projected the data vectors in both cases in a two-dimensional space and tested
374 different values for the size of the neighborhood (n_neighbors) and the minimal distance (min_dist)
375 between data points (either samples or genes).

376

377 # 3 RESULTS AND DISCUSSION

378 ## 3.1 Genome assembly

379 PacBio reads classified as eukaryotic (4,454,601 reads) were used for a contamination-free assembly
380 of the *D. radicum* genome with Canu. We expected a high heterozygosity rate due to the pooling of
381 multiple *D. radicum* individuals. Setting Canu parameters accordingly to prevent haplotypes from being
382 collapsed during the assembly process, resulted in a raw assembly with the length of approximately
383 2.538 Gbp, which was nearly twice the size of the expected genome, an N50 contig of approximately
384 205.3 Kbp and in total 29,244 contigs (Table 2). By evaluating the completeness of the raw genome
385 assembly with BUSCO (using three different sets of orthologous genes at different levels of evolutionary
386 relatedness), the *raw* assembly revealed a completeness of at least 95.7 % for the Diptera (Figure 2b,
387 Table S1) and for more than 98 % for the Endopterygota gene set (Table S2). These results showed a
388 high completeness of the raw assembly, but also the existence of a reasonable percentage of duplicated
389 sequences.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

390 Improving the sequence quality of the raw genome assembly by performing two rounds of polishing with
391 Arrow, increased not only the size of the assembly to approximately 2.544 Gbp (Table 2), but also the
392 completeness of the *polished* assembly. Especially the percentage of complete genes in the BUSCO
393 Diptera gene set increased to more than 97 %. Simultaneously, the number of duplicated genes
394 increased (Figure 2b, Table S1).

395 Next, removing duplicated sequences in the polished assembly with purge_dups successfully reduced
396 the size of the assembly to approximately 1.326 Gbp and a total of 7,014 contigs with an N50 of nearly
397 656.5 Kbp. The size of the *purged* assembly was already close to the experimentally determined genome
398 size. By evaluating the completeness of the purged assembly, we observed a strong reduction in the
399 percentage of duplicated genes, for the Diptera gene set to 6.1 % (Figure 2b, Table S1). As a side effect
400 of removing sequences, the completeness of the gene sets dropped slightly to 93.5 % (Figure 2b, Table
401 S1).

402 For the final *chromosome-scale* assembly, we scaffolded the contigs of the purged assembly with Hi-C
403 data using Juicer and the 3D-DNA genome assembler. The resulting assembly comprised six
404 chromosome-scale contigs (Figure 2a, Table S3), which was consistent with the number of
405 chromosomes determined by karyotyping (Hartman & Southern, 1995), and 2,981 smaller, not-
406 assembled contigs. The final assembly of the *D. radicum* genome yielded approximately 1.326 Gbp,
407 where 96.67 % of the bases (nearly 1.281 Gbp) were anchored to the six pseudo-chromosomes. The
408 size of the six pseudo-chromosomes ranged from one small chromosome with 13 Mbp to five larger
409 chromosomes between 209 and 328.5 Mbp (Table S3). This is in line with the karyotype of *D. radicum*,
410 which comprises five large and one much smaller chromosome (3.3 % of the large chromosomes' size)
411 (Hartman & Southern, 1995).

412 Validation of the final assembly with BUSCO (Table S1, S2) showed no considerable change in the
413 number of complete genes, but the number of single-copy genes increased to 92.2 % (3,030 genes)
414 while the number of duplicated genes decreased to 1.2 % (40 genes) for the Diptera gene set. The six
415 pseudo-chromosomes along with the small contigs were used for all further analyses and are referred to

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

416 as the *D. radicum* genome hereafter. The number of single-copy BUSCOs of the Diptera gene set in the
417 *D. radicum* genome, was similar to those of other Diptera genomes (Figure 2c, Table S4), indicating that
418 the chromosome-scale genome assembly of *D. radicum* was of comparable quality. Based on our
419 findings, we can conclude that the final *D. radicum* chromosome-scale assembly was accurate, complete
420 and without prokaryotic contamination.

421 ## 3.2 Phylogeny and synteny

422 To examine the phylogenetic relationship of *D. radicum* to other insects, we compared complete single-
423 copy BUSCOs of the Endopterygota gene set (comprising totally 2,124 genes) shared by the selected
424 nine insect species belonging to Diptera (4), Coleoptera (1) and Lepidoptera (4, Table 1). We identified
425 1,217 (Table S5, Table S6) shared, and therefore conserved, single-copy genes (Figure 3a, Table S5).
426 Reconstruction of the evolutionary relationships among these ten species based on the shared gene
427 sets revealed that the root fly *D. radicum* was most closely related to the blow fly, *L. cuprina*, followed by
428 the house fly, *M. domestica,* and the fruit fly, *D. melanogaster* (Figure 3a). These relations were
429 consistent with their taxonomic position (Wiegmann et al., 2011).

430 In our synteny analysis, we successfully mapped the six pseudo-chromosomes of *D. radicum* to the six
431 chromosomes of *D. melanogaster* (Figure 3b). This was achieved by predicting gene models (Table 1)
432 in the *D. radicum* genome based on the annotated *D. melanogaster* genome using GeMoMa (Table S7).
433 Genes annotated on the X chromosome of *D. melanogaster* mapped successfully on the second largest
434 chromosome (HiC_scaffold_2) in the *D. radicum* assembly. Genes annotated for the other *D.*
435 *melanogaster* chromosomes were mainly localized on the remaining four larger *D. radicum*
436 chromosomes (Figure 3b). For the smallest chromosome (HiC_scaffold_6) we found indications that this
437 might be related to chromosome 4 (NC_004353.4) of *D. melanogaster* (Table S7).

438 ## 3.3 Genome annotation and functional gene annotation

439 ### 3.3.1 Process of genome annotation and evaluation

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

440 We sequenced the transcriptomes of all four life stages (eggs, larvae, pupae, and adults) of *D. radicum,*
441 and included two stress factors (heat stress on adults and plant toxin on larvae) that are relevant for the
442 survival of *D. radicum* to support the prediction of a comprehensive set of protein-coding genes in the *D.*
443 *radicum* genome.

444 Our homology-based protein-coding gene prediction with GeMoMa relied on nine already sequenced
445 and annotated genomes of phylogenetically related species, herbivore species sharing the same host
446 plant range or common pests on crop plants or stored grains (Table 1). We predicted 19,343 protein-
447 coding genes comprising 46,286 transcripts (Table 3) having a homologue in at least one of the nine
448 selected species.

449 As a complementary approach, we assembled the transcriptomes of all life stages from egg to adult,
450 plus adults and larvae subjected to two stage-related stress factors using Cufflinks. From the pure
451 RNASeq-based transcriptome data, we were able to predict 16,188 protein-coding genes covering
452 23,729 transcripts (Table 3) that were expressed at the sampled time points of the different life stages.

453 To cover the hitherto non-annotated and not or low expressed *D. radicum*-specific genes under our
454 conditions, we performed *ab initio* gene prediction. A total of 81,150 genes yielding 82,473 transcripts
455 were predicted (Table 3). Similarly, as before, we retained all predicted genes, to allow future users the
456 option to choose their own filtering criteria in later studies.

457 The integration of the predictions of all three approaches into a comprehensive annotation led to 81,000
458 putative genes covering 121,731 transcripts (Table 3), where a relatively high number of putative genes
459 was predicted specifically by the *ab initio* approach (Figure 4a). Nearly 95.5 % of the genes were located
460 on the six chromosomes (Table S3).

461 Evaluation with BUSCO showed that our genome annotation covered 93.6 % complete-copy genes of
462 the Diptera gene set, and 95.4 % of the Endopterygota gene set (Table S8). By determining the overlap
463 of the predictions, we found 7,129 genes that were predicted by all three approaches and a total of
464 13,264 genes by at least two approaches (Figure 4a). The annotation of the latter set of genes covered

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

465 87.5 % of complete-copy genes of the Diptera gene set and 89.5 % of the Endopterygota gene set (Table
466 S8). Only the combination of all three approaches led to a complete annotation of the *D. radicum*
467 genome.

468 ### 3.3.1 Functional annotation

469 Overall, 77.1 % (62,418) of the genes were functionally annotated with at least one GO-term and/or
470 protein family or domain information (Figure 4b, Figure S2, Table S9), including 71.15 % (42,244) of the
471 only *ab initio* predicted genes.

472 Focusing on the expressed genes by using our in-house whole life stage RNASeq data, we found a
473 reasonable number of 30,492 genes (37.64 %) having an estimated expression of ≥ 1 transcript per
474 million (TPM) (Figure 4c). A high number of genes was predicted by BRAKER only, but most of these
475 genes were not expressed under our conditions, although the number of expressed genes is higher than
476 in the other sets (Figure 4c). From the set of expressed genes, 50 % (15,270) were functionally annotated
477 with at least one GO-term (Figure 4d).

478 Taken together, these findings indicate that our gene annotation is complete and accurate. We will
479 demonstrate its applicability to generate biologically relevant information in the following section, where
480 we analyzed the transcriptomes of all life stages of *D.* radicum to identify life stage-specific functional
481 gene expression underlying adaptations to their stage-specific life styles, especially to their host plants.

482 ## 3.4 The *D. radicum* life cycle

483 An unsupervised clustering analysis of the expressed gene set with UMAP showed a high similarity of
484 samples belonging to the same life stage (larva or adult, Figure 5a), even if they were subjected to
485 different conditions (control and stressed). We also found that all samples of the egg and pupal stage
486 clustered together. This seems logical, considering that the egg and pupal life stages both undergo
487 considerable morphological and physiological transformation processes, and, in contrast to larvae or
488 adults, are less involved with digestive, locomotory, gustatory and olfactory processes.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

489 We also found that the total number of expressed genes differed among life stages (Figure 5b). The
490 lowest total number of expressed genes was detected in the egg stage and the highest in the larval and
491 pupal stages (Figure 5b, horizontal bar plot). When looking at the overlap among the life stages, we
492 found 31.6 % of the 30,492 genes to be expressed across all life stages (Figure 5b, vertical bar plot).
493 Another 36 % of the genes were exclusively expressed in either a single life stage or condition, in the
494 intersection of both conditions of the larval and the adult stage, respectively or specific in egg and pupal
495 stage (Figure 5b, Figure S3). In the UMAP plot (Figure 5c), genes expressed in single life stages were
496 located at the top and formed life stage-specific spots, whereas genes expressed in all life stages also
497 clustered but were located on the opposite side. The remaining one-third of the genes (not shown in
498 Figure 5b, Figure S3) clustered in between. For larval or adult stages, we observed that genes expressed
499 under different conditions clustered closely together and formed life stage-specific clusters (Figure 5c).

500 An ontology-based gene expression analysis revealed life-stage specific groups related to biological
501 processes (BP), molecular functions (MF) and cellular components (CC, Figure 6, Figure S4, Table
502 S10). In the egg stage, mainly genes involved in the embryonic development (BP), transcriptomic
503 activity (MF) and genetic material (CC, Figure 6) were expressed. Especially genes belonging to the
504 GO biosynthetic processes DNA biosynthesis, metabolic processes, egg shell layer formation
505 (amnioserosa formation) and organ development (muscle and organ formation) were expressed
506 (Figure S4a). These processes are involved in the transition from embryo to larva, which requires
507 active cell division and involves a broad range of metabolic processes to synthesize cell components,
508 membranes and organs (Beutel, Friedrich, Yang, & Ge, 2013). These structures require different
509 macromolecules; indeed we found several expressed genes related to molecular biosynthesis
510 processes in the eggs (Figure S4a). Cell differentiation and organ formation require regulation,
511 coordination and binding activation (Izumi, Yano, Yamamoto, & Takahashi, 1994) which was reflected
512 in our BP expression data (Figure 6, Figure S4a).
513 Genes involved in the body development (BP), structural and transposase (MF), and extracellular matrix
514 (CC) were more frequently expressed in pupae (Figure 6). These genes belong to GOs comprising
515 regulators, binding activity, biosynthesis, metabolism and DNA amplification (Figure S4). During the

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

516 pupal stage, metamorphosis results in the 'disassembly' of larval structures to form adult wings,
517 compound eyes and legs (Buszczak & Segraves, 2000; Chapman & Chapman, 1998). This requires the
518 expression of genes involved in catabolic processes, as well as in organ and cuticle formation. Indeed,
519 we found an increased expression of genes responsible for nuclease and peptidase activity (MF) and
520 chitin-based cuticle structures (CC, Figure 6, S4). This is in line with the gene expression profiles in *D.*
521 *melanogaster* pupae (Arbeitman et al., 2002).

522 Genes connected to the metabolic processes (BP) were highly expressed in the larval stage (Figure 6).
523 We found genes coding for peptidases and polymerases (MF), involved in DNA processes or functions
524 and biosynthetic processes (BF) to be highly expressed (Figure 6, Figure S4). These genes are likely
525 related to feeding and digestion as well as to growth and molting, which are the main processes in the
526 larval stage (Chapman & Chapman, 1998; Chen, 1966). In larvae exposed to the plant toxin ITC, we
527 found that peptidase genes (MF) and genes involved in metabolic and biosynthetic processes (BP)
528 were activated (Figure 6, Figure S4). These enzymes may be involved in catabolizing plant toxins as
529 has been described for other herbivores feeding on *Brassica* plants (Schramm, Vassão, Reichelt,
530 Gershenzon, & Wittstock, 2012).
531 Genes coding for the detection of visible and UV-light, optomotor capability, detection of chemical
532 stimuli (taste, smell) and temperature (BP) were exclusively expressed in adults (Figure 6, Figure S4a).
533 The expression of these gene sets, which are involved in the sensory, optomotor and nervous systems
534 (BP), are important to localize food sources and suitable hosts for oviposition (Gouinguene & Städler,
535 2005; Gouinguené & Städler, 2006; Peter Roessingh & Städler, 1990). In addition, several genes
536 coding for receptors and ion channels were expressed (Figure 6, MF). These genes are involved in the
537 detection of environmental stimuli and signal transmission via the nervous cells to the brain (Sato &
538 Touhara, 2008). Specific for the adult life stage were also the expression of adult behavior linked
539 genes (Figure S4a).
540 Exposing adult flies to a higher temperature resulted in the enhanced expression of peptidases, ion
541 binding (MF), sensory system, especially smell and egg formation (BP) related genes compared to
542 control adults (Figure 6, Figure S4). High temperatures alter protein stability, structures and folding,

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

543 followed by functional changes (Jaenicke et al., 1990). The activation of peptidases might avoid
544 malfunction of proteins under heat stress. Temperature changes affect also the volatility of volatile
545 organic compounds (VOCs) as well as the emission rates of plants (Copolovici & Niinemets, 2016).
546 Since adults of *D. radicum* are attracted by VOCs to localize host plants (Finch, 1978), the enhanced
547 expression of genes related to VOC perception (smell) in flies might indicate towards an adaptive
548 response. Investing in offspring, under these circumstances might ensure the survival for the fly
549 population, and to localize a possible host plant for their oviposition, *D. radicum* females utilize odor
550 signals (Nottingham, 1988).

551

552 ## 4 Conclusion

553 An increasing number of assembled and annotated insect genomes have been published over the last
554 decade. However, genomes of belowground insects and especially root-feeding herbivores are
555 underrepresented. We sequenced the genome of a belowground-feeding agricultural pest, the cabbage
556 root fly *Delia radicum,* whose larvae are also used as a 'model' belowground herbivore in studies on
557 optimal defense allocation and systemic induced responses in plants. Using PacBio and Hi-C
558 sequencing, we generated a 1.3 Gbp assembly with an N50 of 242 Mbp, 6 pseudo-chromosomes and
559 13,618 annotated genes using homology-, transcriptome- and model-predicted approaches, predicted
560 by at least two approaches. During the assembly process, we identified a co-assembled *Wolbachia*
561 species, a very common endosymbiont in insects (Werren & Windsor, 2000). The *Wolbachia* genome
562 consisted of a single contig of 1.59 Mbp matching to the size of the *Wolbachia* supergroups A and B
563 (~1.4 - 1.6 Mbp) which are typical for arthropods (Lo, Casiraghi, Salati, Bazzocchi, & Bandi, 2002). Such
564 co-assembled endosymbiont genomes can be valuable to understand host-symbiont interactions and
565 their roles in other interactions such as host-plant adaptations.

566 Our accurate and well-annotated genome can reveal genetic mechanisms underpinning adaptation of
567 *D. radicum* to its host plants and their specific chemical defenses, the glucosinolate-isothiocyanate
568 system. With our work we provide a tool to understand how the different life stages of this herbivore have
569 adapted to their host plants by identifying adult-specific genes involved in olfactory orientation or the

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

570 detoxification of plant defense compounds in larvae. The genome and the transcriptomes can further be
571 used to understand adaptation to specific conditions, i.e. the evolution of pesticide resistance and
572 adaptive responses to environmental stress factors, such as temperature increase or soil pollution. This
573 high-quality genome is also an important tool to develop novel strategies to combat this pest, for example
574 highly specific dsRNA-based pesticides, which can discriminate between target and non-target species.
575 Moreover, the genus *Delia* contains several other pest species, such as the turnip root fly *D. floralis*, the
576 onion fly *D. antiqua* and the seed bulb maggot, *D. platura.* As their common names indicate, they attack
577 a range of agricultural crops. The genome of *D. radicum* is an excellent foundation to further explore the
578 genetic mechanisms underlying adaptation to chemical host-plant defenses among member of the genus
579 *Delia*.
580

581 Acknowledgements

582 We thank the Long Read Team of the DRESDEN-concept Genome Center, DFG NGS Competence
583 Center, part of the Center for Molecular and Cellular Bioengineering (CMCB), Technische Universität
584 Dresden and the MPI-CBG, especially Sylke Winkler for their great support and the collaboration.
585 Dominik Jakob is acknowledged for his assistance with the insect culture. Great thanks to Denis Tagu
586 and Fabrice Legeai (INRAE, Rennes, France), Denis Poinsot (University of Rennes, France) and
587 Ekaterina Shelest (BIU, iDiv, Leipzig, Germany) for their helpful advice and encouragement in earlier
588 stages of this project. We also greatly thank Jens Keilwagen (JKI, Quedlinburg) for valuable discussions
589 on gene prediction and support on GeMoMa.
590

591 Funding

592 This research was funded by the German Research Foundation (DFG) Collaborative Research Center
593 1127 ChemBioSys (project number 09161509) to RS and NvD, and the German Centre for Integrative
594 Biodiversity Research (iDiv) funded by DFG, grant number- FZT 118, 202548816) to RS, YP, CG, and
595 NvD. HV and YO thank the Max-Planck-Gesellschaft for funding.
596

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

597 ## Author Contributions

598 YP, RS, ND designed the project, AC provided the starting culture of the insects, RS, HV performed the
599 laboratory work, YP, CG, YO, HV performed the data processing and analysis, YP created the figures,
600 YP, RS, ND, HV wrote a first version. All authors contributed to the writing process.

601

602 ## Data Availability Statement

603 Genome sequences and raw data used for genome assembly (PacBio sequences and Illumina Hi-C
604 sequences) and annotation (Illlumina RNASeq sequences) will be available at NCBI. Final genome
605 annotation, respective annotations by GeMoMa, Cufflinks and BRAKER, and functional transcript
606 annotations made by InterProScan and PANNZER2 will be available via Zenodo.

607

608 ## ORCID

609 Rebekka Sontowski: 0000-0001-5791-8814
610 Yvonne Poeschl: 0000-0002-6727-6891
611 Yu Okamura: 0000-0001-6765-4998
612 Heiko Vogel: 0000-0001-9821-7731
613 Cervin Guyomar: 0000-0003-2707-2541
614 Nicole M. van Dam : 0000-0003-2622-5446

615

616 ## References

617 Alexa, A., & Rahnenfuhrer, J. (2020). topGO: enrichment analysis for gene ontology. *R package*
618 *version, 2.42.0*(0).
619 Allema B, Hoogendoorn M, van Beek J, & P, L. (2017). *Neonicotinoids in European agriculture. Main*
620 *applications, main crops and scope for alternatives.* . Retrieved from Culemborg, The
621 Netherlands:
622 Arbeitman, M. N., Furlong, E. E. M., Imam, F., Johnson, E., Null, B. H., Baker, B. S., . . . White, K. P.
623 (2002). Gene Expression During the Life Cycle of Drosophila melanogaster. *Science,*
624 *297*(5590), 2270-2275. doi:10.1126/science.1072152
625 Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under
626 dependency. *Annals of statistics*, 1165-1188.
627 Beutel, R. G., Friedrich, F., Yang, X.-K., & Ge, S.-Q. (2013). *Insect morphology and phylogeny: a*
628 *textbook for students of entomology*: Walter de Gruyter Berlin/Boston.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., . . . Finn, R. D. (2020). The InterPro protein families and domains database: 20 years on. *Nucleic acids research, 49*(D1), D344-D354. doi:10.1093/nar/gkaa977

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology, 34*(5), 525-527. doi:10.1038/nbt.3519

Bruck, D. J., Snelling, J. E., Dreves, A. J., & Jaronski, S. T. (2005). Laboratory bioassays of entomopathogenic fungi for control of Delia radicum (L.) larvae. *Journal of invertebrate pathology, 89*(2), 179-183. doi:https://doi.org/10.1016/j.jip.2005.02.007

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics, 3*(1). doi:10.1093/nargab/lqaa108

Buszczak, M., & Segraves, W. A. (2000). Insect metamorphosis: Out with the old, in with the new. *Current Biology, 10*(22), R830-R833. doi:https://doi.org/10.1016/S0960-9822(00)00792-2

Capinera, J. L. (2008). *Encyclopedia of entomology* (2nd edition ed.): Springer Science & Business Media.

Carlson, M., Falcon, S., Pages, H., & Li, N. (2020). GO. db: A set of annotation maps describing the entire Gene Ontology. *R package version, 3.12.1*(0).

Chapman, R. F., & Chapman, R. F. (1998). *The insects: structure and function*: Cambridge university press.

Chen, P. S. (1966). Amino Acid and Protein Metabolism in Insect Development. In J. W. L. Beament, J. E. Treherne, & V. B. Wigglesworth (Eds.), *Advances in Insect Physiology* (Vol. 3, pp. 53-132): Academic Press.

Copolovici, L., & Niinemets, Ü. (2016). Environmental Impacts on Plant Volatile Emission. In J. D. Blande & R. Glinwood (Eds.), *Deciphering Chemical Language of Plant Communication* (pp. 35-59). Cham: Springer International Publishing.

Crespo, E., Hordijk, C. A., de Graaf, R. M., Samudrala, D., Cristescu, S. M., Harren, F. J., & van Dam, N. M. (2012). On-line detection of root-induced volatiles in Brassica nigra plants infested with Delia radicum L. root fly larvae. *Phytochemistry, 84*, 68-77.

Danner, H., Brown, P., Cator, E. A., Harren, F. J. M., van Dam, N. M., & Cristescu, S. M. (2015). Aboveground and Belowground Herbivores Synergistically Induce Volatile Organic Sulfur Compound Emissions from Shoots but Not from Roots. *Journal of chemical ecology, 41*(7), 631-640. doi:10.1007/s10886-015-0601-y

Dixon, P. L., Coady, J. R., Larson, D. J., & Spaner, D. (2004). Undersowing rutabaga with white clover: impact on Delia radicum (Diptera: Anthomyiidae) and its natural enemies. *The Canadian Entomologist, 136*(3), 427-442.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15-21. doi:10.1093/bioinformatics/bts635

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., . . . Aiden, E. L. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science, 356*(6333), 92-95. doi:10.1126/science.aal3327

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems, 3*(1), 95-98. doi:https://doi.org/10.1016/j.cels.2016.07.002

Ekuere, U. U., Dosdall, L. M., Hills, M., Keddie, A. B., Kott, L., & Good, A. (2005). Identification, Mapping, and Economic Evaluation of QTLs Encoding Root Maggot Resistance in Brassica. *Crop Science, 45*(1), cropsci2005.0371. doi:https://doi.org/10.2135/cropsci2005.0371

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

676 Ferry, A., Dugravot, S., Delattre, T., Christides, J.-P., Auger, J., Bagnères, A.-G., . . . Cortesero, A.-M.
677 (2007). Identification of a Widespread Monomolecular Odor Differentially Attractive to
678 Several Delia Radicum Ground-dwelling Predators in the Field. *Journal of chemical ecology,*
679 *33*(11), 2064-2077. doi:10.1007/s10886-007-9373-3
680 Finch, S. (1978). Volatile plant chemicals and their effect on host plant finding by the cabbage root fly
681 (Delia Brassicae). *Entomologia Experimentalis et Applicata, 24*(3), 350-359.
682 doi:https://doi.org/10.1111/j.1570-7458.1978.tb02793.x
683 Finch, S., & Ackley, C. M. (1977). Cultivated and wild host plants supporting populations of the
684 cabbage root fly. *Annals of Applied Biology, 85*(1), 13-22. doi:https://doi.org/10.1111/j.1744-
685 7348.1977.tb00626.x
686 Fournet, S., Stapel, J., Kacem, N., Nenon, J., & Brunel, E. (2000). Life history comparison between two
687 competitive Aleochara species in the cabbage root fly, Delia radicum: implications for their
688 use in biological control. *Entomologia Experimentalis et Applicata, 96*(3), 205-211.
689 Gehlenborg, N. (2019). UpSetR: a more scalable alternative to Venn and Euler diagrams for visualizing
690 intersecting sets [internet].
691 Gouinguene, S. P. D., & Städler, E. (2005). Comparison of the sensitivity of four Delia species to host
692 and non-host plant compounds. *Physiological Entomology, 30*(1), 62-74.
693 doi:https://doi.org/10.1111/j.0307-6962.2005.00432.x
694 Gouinguené, S. P. D., & Städler, E. (2006). Comparison of the egg-laying behaviour and
695 electrophysiological responses of Delia radicum and Delia floralis to cabbage leaf compounds.
696 *Physiological Entomology, 31*(4), 382-389. doi:https://doi.org/10.1111/j.1365-
697 3032.2006.00532.x
698 Griffiths, G. (1991). *Economic assessment of cabbage maggot damage in canola in Alberta.* Paper
699 presented at the Proceedings of the GCIRC Eighth International Rapeseed Congress.
700 Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing
701 haplotypic duplication in primary genome assemblies. *Bioinformatics, 36*(9), 2896-2898.
702 doi:10.1093/bioinformatics/btaa025
703 Hare, E. E., & Johnston, J. S. (2011). Genome Size Determination Using Flow Cytometry of Propidium
704 Iodide-Stained Nuclei. In V. Orgogozo & M. V. Rockman (Eds.), *Molecular Methods for*
705 *Evolutionary Genetics* (pp. 3-12). Totowa, NJ: Humana Press.
706 Hartman, T. P. V., & Southern, D. I. (1995). Genome reorganization from polyteny to polyploidy in the
707 nurse cells found in onion fly (Delia antiqua) and cabbage root fly (Delia radicum) ovaries
708 (Diptera, Anthomyiidae). *Chromosome Research, 3*(5), 271-280. doi:10.1007/BF00713064
709 Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2015). BRAKER1: Unsupervised
710 RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics,*
711 *32*(5), 767-769. doi:10.1093/bioinformatics/btv661
712 Hoff, K. J., Lomsadze, A., Borodovsky, M., & Stanke, M. (2019). Whole-genome annotation with
713 BRAKER *Gene prediction* (Vol. 1962, pp. 65-95). New York: Springer.
714 Hopkins, R. J., van Dam, N. M., & van Loon, J. J. (2009). Role of glucosinolates in insect-plant
715 relationships and multitrophic interactions. *Annual review of entomology, 54*, 57-83.
716 Izumi, S., Yano, K., Yamamoto, Y., & Takahashi, S. Y. (1994). Yolk proteins from insect eggs: structure,
717 biosynthesis and programmed degradation during embryogenesis. *Journal of Insect*
718 *Physiology, 40*(9), 735-746.
719 Jaenicke, R., Heber, U., Franks, F., Chapman, D., Griffin, M. C. A., Hvidt, A., . . . Franks, F. (1990).
720 Protein structure and function at low temperatures. *Philosophical Transactions of the Royal*
721 *Society of London. B, Biological Sciences, 326*(1237), 535-553. doi:doi:10.1098/rstb.1990.0030

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

722 Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., . . . Hunter, S. (2014). InterProScan 5:
723 genome-scale protein function classification. *Bioinformatics, 30*(9), 1236-1240.
724 doi:10.1093/bioinformatics/btu031
725 Kapranas, A., Sbaiti, I., Degen, T., & Turlings, T. C. J. (2020). Biological control of cabbage fly Delia
726 radicum with entomopathogenic nematodes: Selecting the most effective nematode species
727 and testing a novel application method. *Biological Control, 144*, 104212.
728 doi:https://doi.org/10.1016/j.biocontrol.2020.104212
729 Katoh, K., Misawa, K., Kuma, K. i., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple
730 sequence alignment based on fast Fourier transform. *Nucleic acids research, 30*(14), 3059-
731 3066. doi:10.1093/nar/gkf436
732 Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O., & Grau, J. (2018). Combining RNA-seq data
733 and homology-based gene prediction for plants, animals and fungi. *BMC bioinformatics,*
734 *19*(1), 189. doi:10.1186/s12859-018-2203-5
735 Keilwagen, J., Wenk, M., Erickson, J. L., Schattat, M. H., Grau, J., & Hartung, F. (2016). Using intron
736 position conservation for homology-based gene prediction. *Nucleic acids research, 44*(9), e89-
737 e89. doi:10.1093/nar/gkw092
738 Kergunteuil, A., Dugravot, S., Danner, H., van Dam, N. M., & Cortesero, A. M. (2015). Characterizing
739 Volatiles and Attractiveness of Five Brassicaceous Plants with Potential for a 'Push-Pull'
740 Strategy Toward the Cabbage Root Fly, Delia radicum. *Journal of chemical ecology, 41*(4), 330-
741 339. doi:10.1007/s10886-015-0575-9
742 Kissen, R., Rossiter, J. T., & Bones, A. M. (2009). The 'mustard oil bomb': not so easy to assemble?!
743 Localization, expression and distribution of the components of the myrosinase enzyme
744 system. *Phytochemistry Reviews, 8*(1), 69-86.
745 Kolde, R. (2019). Package 'pheatmap'. *R package, 1.0.12*.
746 Konopka, T. (2020). umap: Uniform Manifold Approximation and Projection. *R package version,*
747 *0.2.7.0*.
748 Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu:
749 scalable and accurate long-read assembly via adaptive k-mer weighting and repeat
750 separation. *Genome research*, gr. 215087.215116.
751 Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., . . . Marra, M. A. (2009).
752 Circos: An information aesthetic for comparative genomics. *Genome research, 19*(9), 1639-
753 1645. doi:10.1101/gr.092759.109
754 Lachaise, T., Ourry, M., Lebreton, L., Guillerm-Erckelboudt, A.-Y., Linglin, J., Paty, C., . . . Mougel, C.
755 (2017). Can soil microbial diversity influence plant metabolites and life history traits of a
756 rhizophagous insect? A demonstration in oilseed rape. *Insect Science, 24*(6), 1045-1056.
757 doi:10.1111/1744-7917.12478
758 Lamy, F., Dugravot, S., Cortesero, A. M., Chaminade, V., Faloya, V., & Poinsot, D. (2018). One more
759 step toward a push-pull strategy combining both a trap crop and plant volatile organic
760 compounds against the cabbage root fly Delia radicum. *Environmental Science and Pollution*
761 *Research, 25*(30), 29868-29879. doi:10.1007/s11356-017-9483-6
762 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Subgroup, G. P. D. P. (2009). The
763 Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079.
764 doi:10.1093/bioinformatics/btp352
765 Lo, N., Casiraghi, M., Salati, E., Bazzocchi, C., & Bandi, C. (2002). How Many Wolbachia Supergroups
766 Exist? *Molecular Biology and Evolution, 19*(3), 341-346.
767 doi:10.1093/oxfordjournals.molbev.a004087

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

768 Lomsadze, A., Burns, P. D., & Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into
769 automatic training of eukaryotic gene finding algorithm. *Nucleic acids research, 42*(15), e119-
770 e119. doi:10.1093/nar/gku557
771 McDonald, R., & Sears, M. (1992). Assessment of larval feeding damage of the cabbage maggot
772 (Diptera: Anthomyiidae) in relation to oviposition preference on canola. *Journal of economic*
773 *entomology, 85*(3), 957-962.
774 Michaud, S., Marin, R., Westwood, J. T., & Tanguay, R. M. (1997). Cell-specific expression and heat-
775 shock induction of Hsps during spermatogenesis in Drosophila melanogaster. *Journal of Cell*
776 *Science, 110*(17), 1989-1997. doi:10.1242/jcs.110.17.1989
777 Neveu, N., Krespi, L., Kacem, N., & Nénon, J. P. (2000). Host-stage selection by Trybliographa rapae, a
778 parasitoid of the cabbage root fly Delia radicum. *Entomologia Experimentalis et Applicata,*
779 *96*(3), 231-237.
780 Nottingham, S. (1988). Host-plant finding for oviposition by adult cabbage root fly, Delia radicum.
781 *Journal of Insect Physiology, 34*(3), 227-234.
782 R Core Team. (2020). R: A Language and Environment for Statistical Computing [R]. R Foundation for
783 Statistical Computing. Vienna, Austria.
784 Robinson, O., Dylus, D., & Dessimoz, C. (2016). Phylo.io : Interactive Viewing and Comparison of Large
785 Phylogenetic Trees on the Web. *Molecular Biology and Evolution, 33*(8), 2163-2166.
786 doi:10.1093/molbev/msw080
787 Roessingh, P., & Städler, E. (1990). Foliar form, colour and surface characteristics influence oviposition
788 behaviour in the cabbage root fly Delia radicum. *Entomologia Experimentalis et Applicata,*
789 *57*(1), 93-100. doi:10.1111/j.1570-7458.1990.tb01419.x
790 Roessingh, P., Städler, E., Fenwick, G., Lewis, J., Nielsen, J. K., Hurter, J., & Ramp, T. (1992).
791 Oviposition and tarsal chemoreceptors of the cabbage root fly are stimulated by
792 glucosinolates and host plant extracts. *Entomologia Experimentalis et Applicata, 65*(3), 267-
793 282.
794 Sato, K., & Touhara, K. (2008). Insect olfaction: receptors, signal transduction, and behavior. In S.
795 Korsching & W. Meyerhof (Eds.), *Chemosensory systems in mammals, fishes, and insects* (Vol.
796 47, pp. 203-220). Berlin, Heidelberg: Springer.
797 Schramm, K., Vassão, D. G., Reichelt, M., Gershenzon, J., & Wittstock, U. (2012). Metabolism of
798 glucosinolate-derived isothiocyanates to glutathione conjugates in generalist lepidopteran
799 herbivores. *Insect biochemistry and molecular biology, 42*(3), 174-182.
800 Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: assessing genome assembly and annotation
801 completeness. In K. M. (Ed.), *Gene prediction* (Vol. 1962, pp. 227-245). New York: Springer.
802 Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level
803 estimates improve gene-level inferences. *F1000Research, 4*.
804 Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
805 phylogenies. *Bioinformatics, 30*(9), 1312-1313. doi:10.1093/bioinformatics/btu033
806 Stanke, M., Schöffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in eukaryotes with
807 a generalized hidden Markov model that uses hints from external sources. *BMC*
808 *bioinformatics, 7*(1), 62. doi:10.1186/1471-2105-7-62
809 Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the
810 analysis of massive data sets. *Nature biotechnology, 35*(11), 1026-1028. doi:10.1038/nbt.3988
811 Törönen, P., Medlar, A., & Holm, L. (2018). PANNZER2: a rapid functional annotation web server.
812 *Nucleic acids research, 46*(W1), W84-W88. doi:10.1093/nar/gky350
813 Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., . . . Pachter, L.
814 (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

815 and isoform switching during cell differentiation. *Nature biotechnology, 28*(5), 511-515.
816 doi:10.1038/nbt.1621
817 Tsunoda, T., Krosse, S., & van Dam, N. M. (2017). Root and shoot glucosinolate allocation patterns
818 follow optimal defence allocation theory. *Journal of Ecology, 105*(5), 1256-1266.
819 doi:10.1111/1365-2745.12793
820 van Dam, N. M., Tytgat, T. O., & Kirkegaard, J. A. (2009). Root and shoot glucosinolates: a comparison
821 of their diversity, function and interactions in natural and managed ecosystems.
822 *Phytochemistry Reviews, 8*(1), 171-186.
823 van Herk, W. G., Vernon, R. S., Waterer, D. R., Tolman, J. H., Lafontaine, P. J., & Prasad, R. P. (2016).
824 Field Evaluation of Insecticides for Control of Cabbage Maggot (Diptera: Anthomyiidae) in
825 Rutabaga in Canada. *Journal of economic entomology, 110*(1), 177-185.
826 doi:10.1093/jee/tow238
827 Wang, S., Voorrips, R. E., Steenhuis-Broers, G., Vosman, B., & van Loon, J. J. (2016). Antibiosis
828 resistance against larval cabbage root fly, Delia radicum, in wild Brassica-species. *Euphytica,*
829 *211*(2), 139-155.
830 Werren, J. H., & Windsor, D. M. (2000). Wolbachia infection frequencies in insects: evidence of a
831 global equilibrium? *Proceedings of the Royal Society of London. Series B: Biological Sciences,*
832 *267*(1450), 1277-1285. doi:doi:10.1098/rspb.2000.1139
833 West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C., & Banfield, J. F. (2018). Genome-
834 reconstruction for eukaryotes from complex natural microbial communities. *Genome*
835 *research, 28*(4), 569-580. doi:10.1101/gr.228429.117
836 Wiegmann, B. M., Trautwein, M. D., Winkler, I. S., Barr, N. B., Kim, J.-W., Lambkin, C., . . . Yeates, D. K.
837 (2011). Episodic radiations in the fly tree of life. *Proceedings of the National Academy of*
838 *Sciences, 108*(14), 5690-5695. doi:10.1073/pnas.1012675108
839 Wittstock, U., & Gershenzon, J. (2002). Constitutive plant toxins and their role in defense against
840 herbivores and pathogens. *Current Opinion in Plant Biology, 5*(4), 300-307.
841 doi:https://doi.org/10.1016/S1369-5266(02)00264-9
842 (Fournet et al., 2000)

843

844

845

846

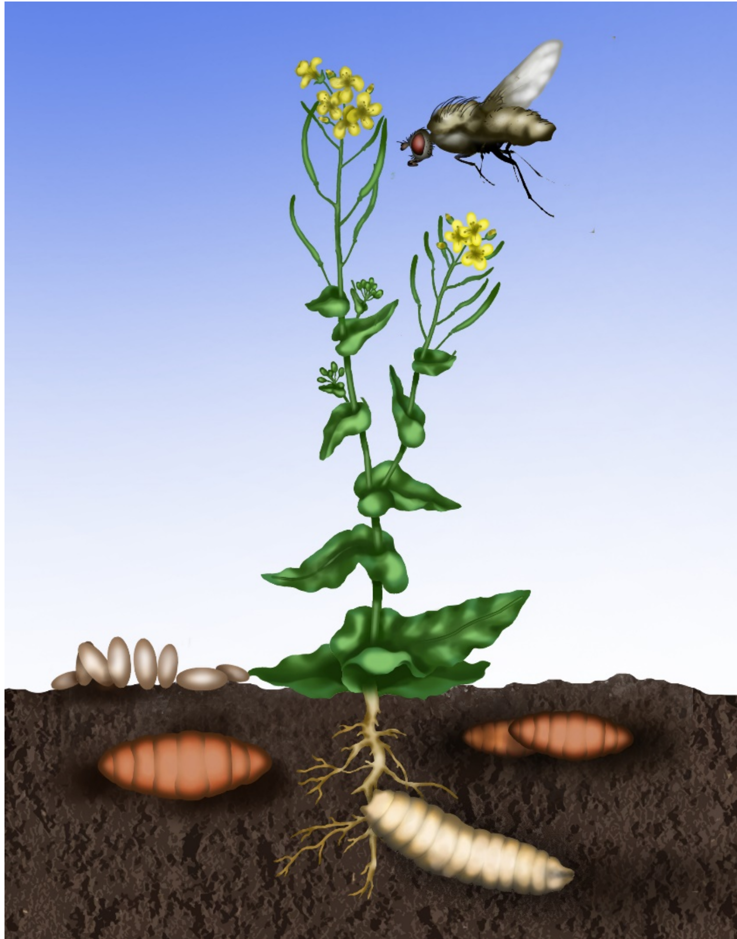
847

848

849

850 Tables and Figures

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

851 

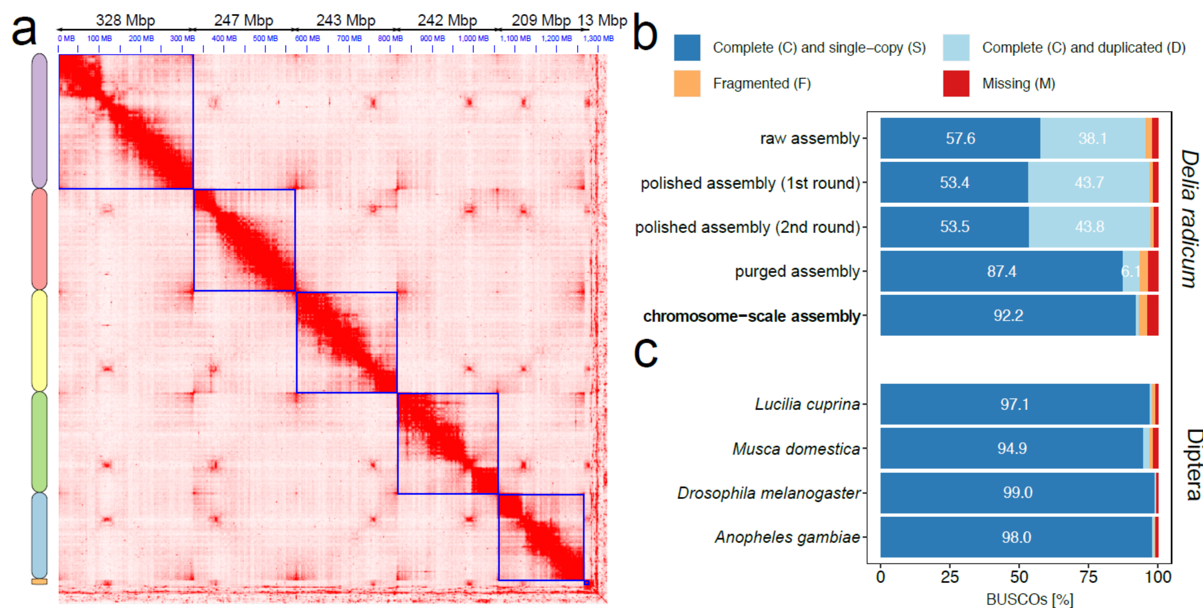
852 Figure 1

853 Schematic illustration of the life stages and their habitats of the cabbage root fly *Delia radicum*. Adult
854 flies are attracted by their host plants for food consumption and oviposition. Eggs are deposited on the
855 soil, where the larvae hatch. Larvae dig into the soil to feed on the roots until they pupate. After
856 completing metamorphosis, adult flies make their way above ground to feed on pollen and nectar, and
857 to reproduce.

858 Picture: Jennifer Gabriel

859

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



860

861 **Figure 2**

862 Chromosome-scale assembly of the *Delia radicum* genome.

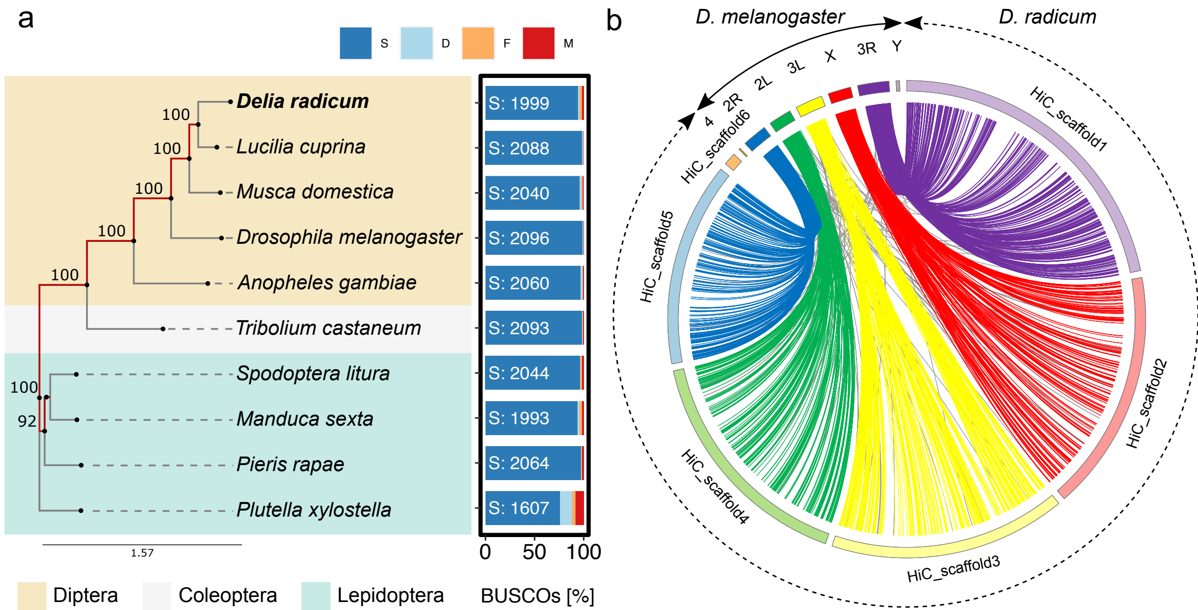
863 (a) Heatmap showing the Hi-C contacts map of the final chromosome-scale assembly, where
864 the six chromosomes (six super-scaffolds) are indicated by the blue boxes. The chromosomes
865 are ordered from largest to smallest; their concrete lengths are given in Mbp above the Hi-C
866 map.

867 (b) Barplot showing the result of BUSCO analyses of the intermediate and final assemblies using
868 the 'Diptera' gene set containing 3,285 genes. Numbers in the bars give the percentage of genes
869 found for the category indicated by the color of the bar.

870 (c) Barplot showing the result of BUSCO analyses using the 'Diptera' gene set of four other
871 dipteran species with published genomes. Numbers in the bars give the percentage of genes
872 found for the category indicated by the color of the bar.

873
874

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



875
876
877

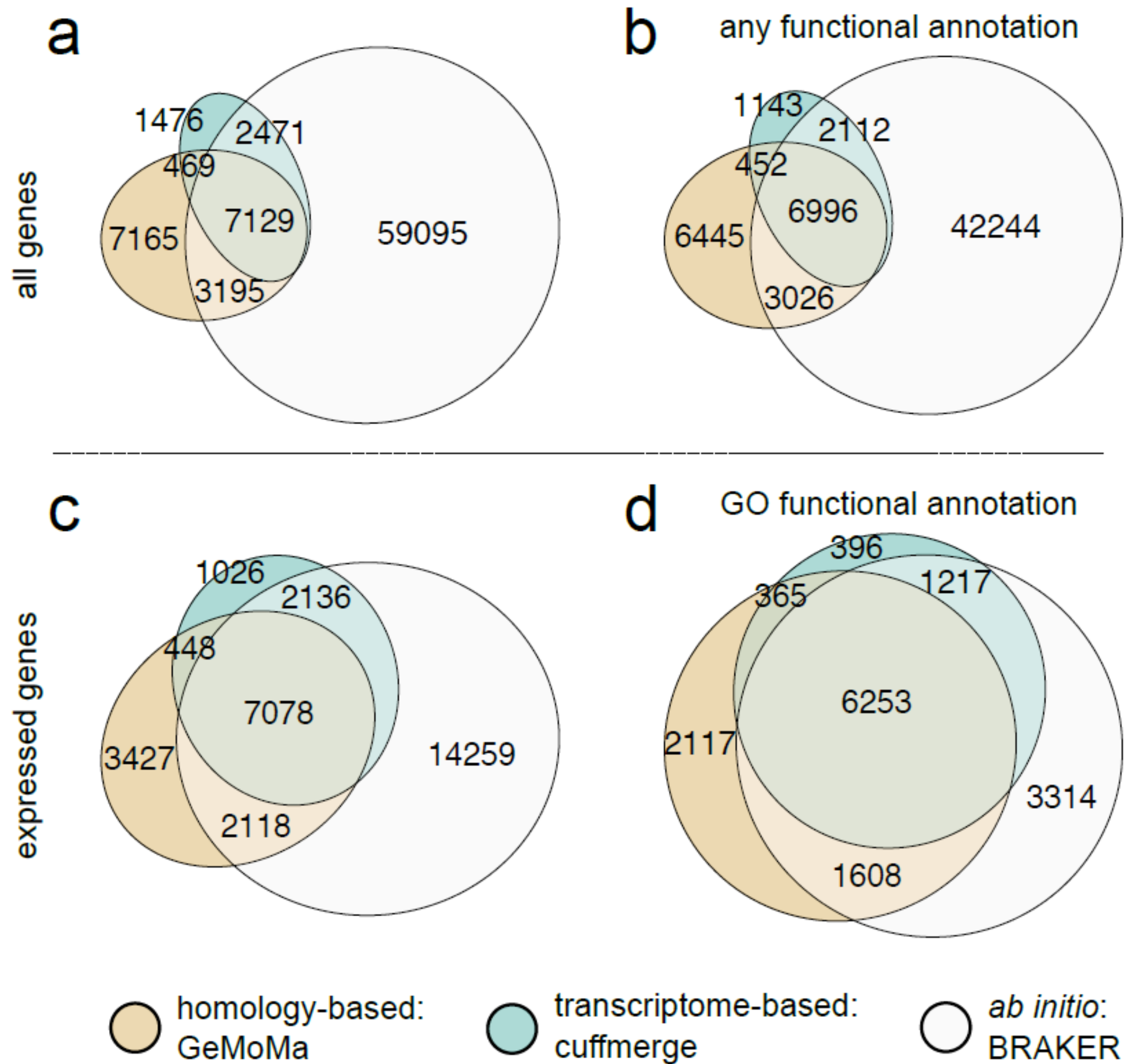
878 **Figure 3**

879 Phylogenetic analyses.

880 (a) A phylogenetic tree reconstructed with RAxML on concatenated alignments of proteins of 1,271
881 genes of BUSCOs' Endopterygota gene set (n=2,124) shared by all ten insect species. Tree
882 reconstruction was done including 100 bootstrapping steps. The level of bootstrapping support is given
883 at the edges. The barplot to the right of the phylogenetic tree shows BUSCO results of each species on
884 the Endopterygota gene set, where S: number of complete single-copy BUSCO genes (dark blue bar);
885 D: duplicated complete copy genes (light blue), F: fragmented genes (orange), M: missing genes (red).
886 (b) A Circos plot linking genes on the assembled scaffolds of *Delia radicum* (HiC_scaffold 1 - 6) to
887 homologues on the *Drosophila melanogaster* chromosomes (2R/2L, 3R/3L, 4, X and Y). Each line
888 connects homologue regions of at least two consecutive genes. Colored lines indicate that homologue
889 regions of a *D. melanogaster* chromosome are connected to those of the syntenic chromosome of *D.*
890 *radicum*. Otherwise they are colored in grey.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

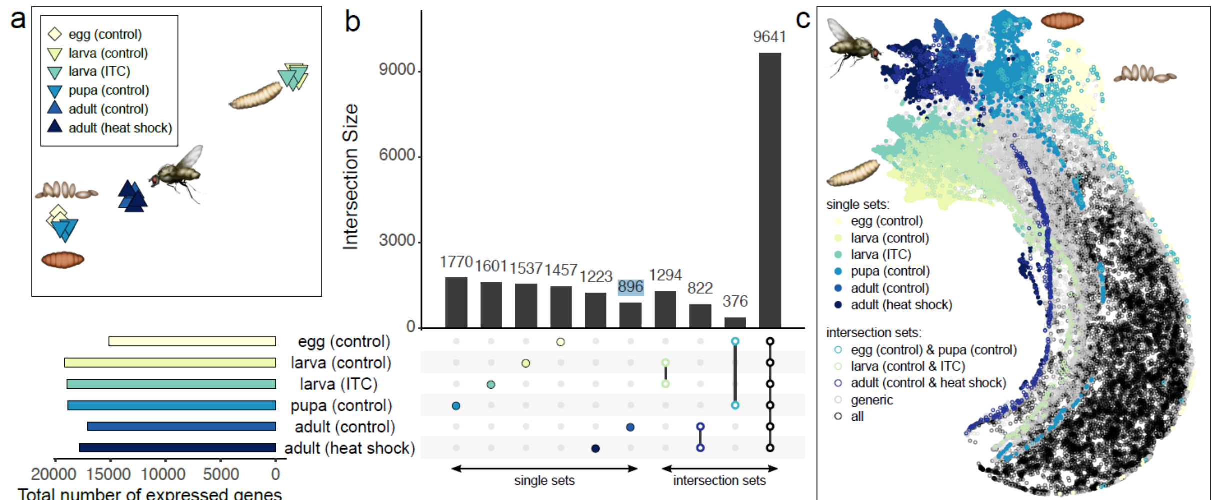
891



892
893
894 **Figure 4**

895 Venn diagrams containing the numbers of genes in the *Delia radicum* genome predicted by
896 homology-based, transcriptome-based or *ab initio* approaches, or a combination thereof. The
897 numbers of genes in the diagrams are based on
898 (a) all predicted genes;
899 (b) all predicted genes with any functional annotation, which includes GO annotation and/or
900 protein family or domain annotation;
901 (c) expressed genes which were predicted genes with a Transcript Per Million (TPM) value ≥ 1.
902 TPM values result from analyses of our in-house life cycle RNASeq data.;
903 (d) expressed genes with I think that is a great a functional annotation based on GO annotation.
904

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



905
906
907

908 **Figure 5**

909 Differences in gene expression profiles among *Delia radicum* life stages and stress conditions.

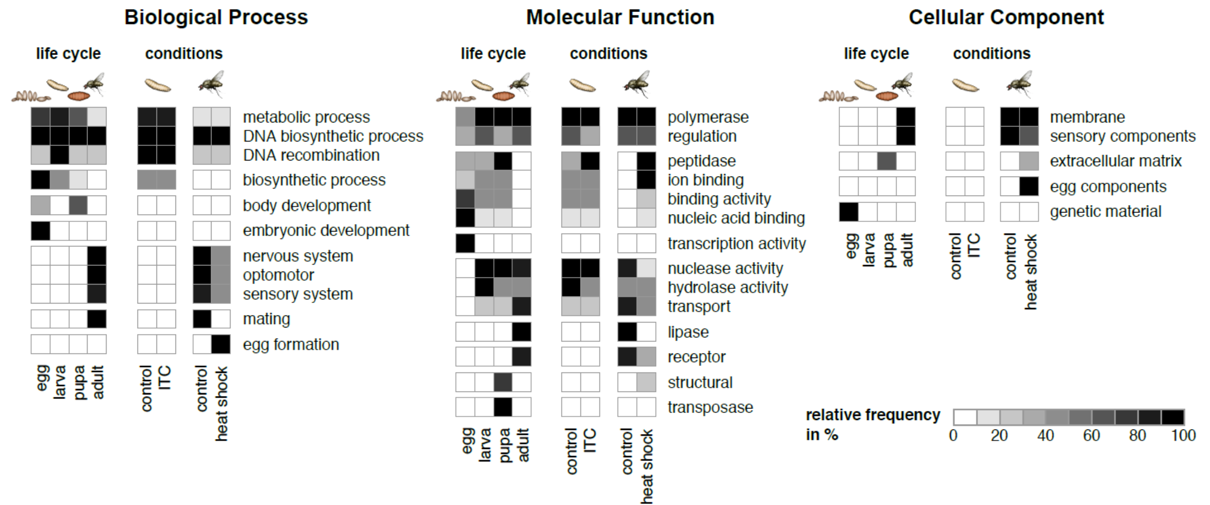
910 (a) Uniform Manifold Approximation and Projection (UMAP) plot showing differences among the
911 life stages based on differing gene expression. ITC = larvae fed on diets with 2 µM phenylethyl
912 isothiocyanate in their diet.

913 (b) UpSet plot showing the number of genes that are exclusively expressed (Transcripts Per
914 Million (TPM) value ≥ 1) in at least one replicate of a life stage or a stress condition (first 6 bars,
915 filled circles); expressed in both conditions within larva and adult life stage (bar 7 and 8, green
916 and dark blue open circles), or both in eggs and pupae (bar 9, cyan open circles), and those
917 expressed in all 18 samples (last bar, open black circles). A selection of intersection sets is
918 shown, whereas the full set is presented in Figure S4. To the left, the total number of expressed
919 genes per life stage and stress condition is shown (colored horizontal bar plot below sub figure
920 a). The remaining genes, referred to as generic, are not shown and sum up to 9,875 genes.

921 (c) UMAP of expressed genes. Genes are colored according to the sets in (b) and are plotted
922 with filled circles when they belong to single sets and with open circles when they belong to
923 intersection sets. Genes expressed in all 18 samples are labeled as "all" (black open circles).
924 Remaining genes are labeled as "generic" (grey open circles).

925
926
927

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



928

929 **Figure 6**

930 Gene ontology (GO) analyses on biological process (BP), molecular function (MF) and cellular
931 component (CC) ontologies, based on expressed genes (Transcripts Per Million (TPM) value ≥
932 1). Results are shown in three respective heatmaps, where rows are labeled by generic
933 categories and columns with life stages and/or conditions. Explicit GO annotations of expressed
934 genes are collapsed into more generic categories. Hence, each cell in a heatmap contains the
935 relative frequency of GO terms sorted into a specific generic category for a specific life stage
936 and/or condition. Only GO terms that were significantly overrepresented in a GO-enrichment
937 analysis (Fisher's exact test, $P<0.05$ after correction with Benjamini - Yekutieli) are considered.
938 Expanded versions of the heatmaps, where detailed GO annotations for each generic category
939 are listed, are provided in Figure S4.

940 In all heatmaps the block with four columns to the left shows the results of all stage of the life
941 cycle under control conditions, whereas the columns to the right show the relative frequencies
942 determined for larvae and adults under control or stress conditions; the data for the control
943 conditions in larva and adult stage are duplicated for easier comparison.

944 ITC = larvae fed on diet with 2-phenylethyl isothiocyanate.

945

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

946 **Table 1** 9 selected insect species.
947 9 insect species (4 Diptera, 4 Lepidoptera, and 1 Coleoptera species) selected for comparative
948 genomics and phylogenetic analyses. Insect species were chosen according to their phylogenetic
949 relatedness to *D. radicum*, or because they shared the same host plant range with *D. radicum* or
950 because they are also common pests in agriculture. All 9 species are fully sequenced and annotated,
951 and information can be obtained from National Center for Biotechnology
952 (https://www.ncbi.nlm.nih.gov).
953
954

Order	Species	NCBI taxid	Names	RefSeq ID	Reason for selection
Diptera	*Anopheles gambiae*	180454	African malaria mosquito	GCF_000005575.2	phylogenetically related
Diptera	*Drosophila melanogaster*	7227	fruit fly	GCF_000001215.4	phylogenetically related
Diptera	*Lucilia cuprina*	7375	Australian sheep blowfly	GCF_000699065.1	phylogenetically related
Diptera	*Musca domestica*	7370	house fly	GCF_000371365.1	phylogenetically related
Lepidoptera	*Manduca sexta*	7130	tobacco hornworm	GCF_000262585.1	common pests on crop plants
Lepidoptera	*Pieris rapae*	64459	cabbage white	GCF_001856805.1	sharing host plant
Lepidoptera	*Plutella xylostella*	51655	diamondback moth	GCF_000330985.1	sharing host plant
Lepidoptera	*Spodoptera litura*	69820	Tobacco cutworm	GCF_002706865.1	common pests on crop plants
Coleoptera	*Tribolium castaneum*	7070	red flour beetle	GCF_000002335.3	common pests on stored grains

955
956

957 **Table 2** Summary of assembly statistics.
958 The raw, polished, and purged assemblies are intermediate assemblies after PacBio read assembly with
959 Canu, two rounds of polishing with Arrow, and purging with purge_dups. The final, chromosome-scale
960 assembly, generated with the 3D-DNA genome assembly pipeline that assembled contigs of the purged
961 assembly by integration of Hi-C Illumina reads into (chromosome-scale) scaffolds. The final
962 chromosome-scale assembly contains 6,190 gaps of length 100 bp, whereby 6,188 gaps are located on
963 the 6 pseudo-chromosomes.
964

Assembly	Number of bases	Number of contigs† or scaffolds‡	N50	L50	N90	L90	Longest contig† or scaffold‡
raw assembly	2,538,077,247	29,244†	205,306	2,197	32,594	16,335	6,127,675
polished assembly	2,544,504,558	29,244†	205,665	2,201	32,715	16,338	6,133,028
Purged assembly	1,325,508,377	7,014†	656,541	485	74,470	2,765	6,133,028
final chromosome-scale assembly	1,326,127,377	2,987‡	242,504,274	3	208,954,159	5	328,483,116
6 pseudo-chromosomes only	1,281,926,506	6‡	242,504,274	3	208,954,149	5	328,483,116

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.11.447147; this version posted June 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

965 † numbers given for the raw, polished and purged assembly refer to contigs
966 ‡ numbers given for the chromosome-scale assembly and the 6 pseudo-chromosomes refer to
967 scaffolds
968
969

970 **Table 3** Summary of gene prediction statistics.
971 Number of gene predictions made on the chromosome-scale genome assembly of *D. radicum* by the
972 three different approaches: GeMoMa a sequence homology-based approach, Cufflinks a RNASeq
973 data-based approach to assemble transcriptomes, and BRAKER an approach for *ab initio* predictions
974 of genes. The final comprehensive gene annotation for the *D. radicum* genome contains 81,000
975 putative genes.
976

Approach	Description	Number of transcripts	Number of genes
Cufflinks	transcriptome-based	23729	16188
GeMoMa	homology-based	46286	19343
BRAKER	ab initio	82473	72613
final		121731	81000

977
978

979 **Table 4** Summary of Gene Ontology (GO) annotations of expressed genes.
980 30,492 genes of the 81,000 genes were expressed (TPM value ≥ 1) in our in-house life stage RNASeq
981 data set. Genes were annotated with GO classes using PANNZER2 and InterProScan. Genes
982 exclusively expressed in one specific life stage were grouped into gene sets named according to the
983 life stage. For the larvae and the adult stage were control and stress conditions are present in the data
984 set, genes that are expressed in both conditions within one life stage were added to the life stage and
985 condition specific gene sets. Numbers of the row labeled with "total" are in concordance with Figure 5b.
986 Listed gene sets were used for life stage specific GO enrichment analyses.
987

set/ontology	complete	egg (control)	larva (control)	pupa (control)	adult (control)	larva (ITC)	adult (heatshock)
total	30492	1457	2831	1770	1718	2895	2045
noGO	15222	1029	2021	1288	1096	2079	1225
GO	15270	428	810	482	622	816	820
BP†	11262	262	501	289	406	495	534
MF‡	13513	397	717	429	537	723	729
CC§	10917	230	444	260	381	434	454

988 † BP: Biological Process, ‡ MF: Molecular Function, § CC: Cellular Compartment
989