Received: 15 July 2021 | Revised: 26 January 2022 | Accepted: 31 January 2022

DOI: 10.1111/1755-0998.13594

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

# A high-quality functional genome assembly of *Delia radicum* L. (Diptera: Anthomyiidae) annotated from egg to adult

Rebekka Sontowski[1,2] | Yvonne Poeschl[2,3,4] | Yu Okamura[5] | Heiko Vogel[5] | Cervin Guyomar[3,6] | Anne-Marie Cortesero[7] | Nicole M. van Dam[1,2]

[1]Molecular Interaction Ecology, German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

[2]Institute of Biodiversity, Friedrich Schiller University Jena, Jena, Germany

[3]Bioinformatics Unit, German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

[4]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle, Germany

[5]Department of Insect Symbiosis, Max Planck Institute for Chemical Ecology, Jena, Germany

[6]GenPhySE, Université de Toulouse, INRAE, ENVT, Castanet Tolosan, France

[7]IGEPP, INRAE, Institut Agro, Univ Rennes, Rennes, France

**Correspondence**
Rebekka Sontowski and Nicole M. van Dam, Molecular Interaction Ecology, German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany.
Email: rebekka.sontowski@idiv.de; nicole.vandam@idiv.de

**Funding information**
German Research Foundation (DFG) Collaborative Research Center 1127 ChemBioSys, Grant/Award Number: 09161509; Deutsches Zentrum für integrative Biodiversitätsforschung Halle-Jena-Leipzig, Grant/Award Number: FZT 118 and 202548816

**Handling Editor:** Joanna Kelley

## Abstract

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

**KEYWORDS**
belowground pest, chromosome-scale genome, de novo genome assembly, functional gene annotation, herbivory, insects

Rebekka Sontowski and Yvonne Poeschl are equal first authors.

## 1 | INTRODUCTION

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

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

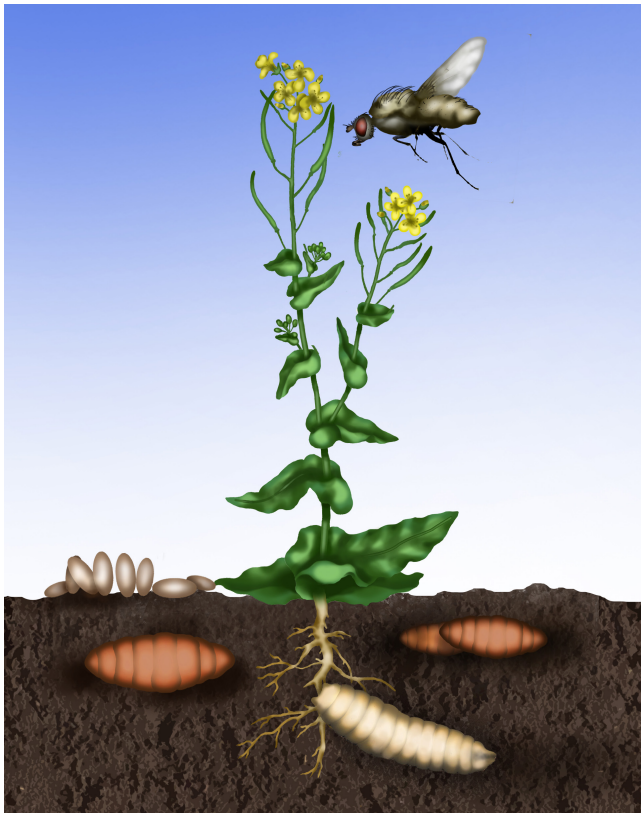

**FIGURE 1** Schematic illustration of the life stages and their habitats of the cabbage root fly *Delia radicum*. Adult flies are attracted by their host plants for feeding and oviposition. Eggs are deposited on the soil where the larvae hatch. Larvae dig into the soil to feed on the roots until they pupate. After completing metamorphosis, adult flies make their way above ground to feed on pollen and nectar and to reproduce. Illustration: Jennifer Gabriel

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

Controlling *D. radicum* in agriculture is a major challenge. Natural resistance to this specialist herbivore has not been identified in currently used cultivars yet (Ekuere et al., 2005) and several effective synthetic insecticides, such as neonicotinoids, have been banned from use due to environmental concerns (Allema et al., 2017). Moreover, pesticide resistance has already developed in this species, for example against chlorpyrifos (van Herk et al., 2016). Alternative and more sustainable pest management strategies are urgently needed. Heritable natural resistance to *D. radicum* is present in wild brassicaceous species, but introgression of these traits may be hampered by crossing barriers and linkage of resistance with undesired traits (Ekuere et al., 2005; Wang et al., 2016). Several studies examined the application of entomopathogenic fungi, natural predators or parasitoids, mixed cropping and soil microbes to better control *D. radicum* (Bruck et al., 2005; Dixon et al., 2004; Fournet et al., 2000; Kapranas et al., 2020; Lachaise et al., 2017; Neveu et al., 2000). Even though each of these measures may reduce *D. radicum* infestations, they cannot prevent yield loss as effectively as synthetic pesticides.

To understand the interaction of *D. radicum* with its host plants, the chemical ecology of this plant-herbivore interaction has been intensively studied over the last decades. These studies analysed aspects ranging from the chemosensory mechanisms of host plant attraction and oviposition choice to herbivore-induced plant responses and interactions with predators and parasitoids (Ferry et al., 2007; Gouinguene & Städler, 2005; Hopkins et al., 2009; Kergunteuil et al., 2015; Roessingh et al., 1992). However, the genetic mechanisms underpinning host-plant adaptation of *D. radicum*

are unknown. An accurate and well-annotated genome can reveal genetic mechanisms underpinning the adaptation of *D. radicum* to its host's chemical defences. In particular, understanding the preference of the different agriculturally relevant life stages (adults and larvae) which occur in separate habitats (above- and belowground) on the genetic level expands our understanding of herbivore-plant interactions. These mechanisms can also be an important starting point to develop novel approaches, such as species-specific dsRNA-based pest control strategies. So far, a genome of this species has not been published.

Here, we assembled and annotated a de novo, chromosome-level scaffolded genome of *D. radicum* using PacBio and Hi-C Illumina sequencing. We used three different approaches to annotate the genome; Cufflinks, which uses transcript assembly; GeMoMa, which is homology-based, and BRAKER, for additional prediction of genes not covered by the first two methods. Generated RNASeq data of all four life stages (eggs, larvae, pupae, adults) and two relevant stress factors (heat stress in adults, plant toxin stress in larvae), allowed us to validate predicted genes and to identify specific gene families which were expressed in each of the life stages.

## 2 | MATERIALS AND METHODS

### 2.1 | Sample material

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

Species identification was performed using a 699 bp fragment of the cytochrome oxidase I (COI) gene as a molecular marker generated with the universal COI primer pair HCO and LCO (Folmer et al., 1994). The sequence was submitted to BLAST online using the BLASTn algorithm (retrieved from https://blast.ncbi.nlm.nih.gov/Blast.cg). The top three hits matched with *D. radicum* COI accessions (MG115888.1, HQ581775.1, GU806605.1) with an identity of more than 98.45%.

### 2.2 | Genome sequencing

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

For PacBio sequencing, 18 randomly collected, fully matured *D. radicum* adults were frozen and stored at −80°C. To sterilize the surface, the flies were incubated for 2 min in bleach (2%), transferred to sodium thiosulphate (0.1 N) for neutralization, washed three times in 70% ethanol and once in autoclaved double distilled water. To reduce contamination by microorganisms from the gut, we extracted total DNA from the head and the thorax of the adults, using a phenol-chloroform extraction method according to the protocol of the sequencing facility (Figure S1). We pooled three individuals per extraction and checked the DNA quality using gel electrophoresis (0.7% agarose gel). DNA purity was assessed using a NanoPhotometer P330 (Implen) and DNA quantity using a Qubit dsDNA BR assay kit in combination with a Qubit 2.0 Fluorometer (Invitrogen). The DNA of all samples was pooled for the sequencing library. Library preparation and sequencing were provided by the facility of the Max Planck Institute of Molecular Cell Biology and Genetics, Dresden/Germany on a PacBio Sequel. A total of 16 SMRT cells were processed and 6,539,960 reads (76.2 Gbp) were generated. Due to the pooling of several females and males, we expected the reads to be highly heterozygous, which we considered during the assembly process.

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

For Hi-C Illumina sequencing fresh *D. radicum* pupae from the above culture were randomly selected. A total of 10 pupae were chopped into small pieces with a razor and resuspended in 3 ml of PBS with 1% formaldehyde. The homogenized sample was incubated at room temperature for 20 min with periodic mixing. Glycine was added to the sample buffer to 125 mM final concentration and incubated at room temperature for ~15 min with periodic mixing. The homogenized tissue was spun down (1000*g* for 1 min), rinsed twice with PBS, and pelleted (1000*g* for 2 min). After removal of the supernatant, the tissue was homogenized to a fine powder in a liquid nitrogen-chilled mortar with a chilled pestle. Further sample processing and sequencing were performed by Phase Genomics on an Illumina HiSeq 4000, generating a total of 181,752,938 paired-end reads (2 × 150 bp).

### 2.3 | Genome size estimation

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

## 2.4 | Genome assembly and completeness

### 2.4.1 | PacBio data processing

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

### 2.4.2 | De novo genome assembly

The long-read assembler Canu (version 1.9) (Koren et al., 2017) was used to generate a de novo genome assembly from filtered PacBio reads. The Canu pipeline, including read error correction and assembly, was started with setting parameters based on the estimated genome size (genomeSize = 1200 m), the use of not too short (minReadLength = 5000) and high quality (stopOnReadQuality = true) PacBio reads, addressing the overlapping of sequences (minOverlapLength = 1000 corOutCoverage = 200), and accounting for the expected high heterozygosity rate of the *D. radicum* genome (batOptions = -dg 3 -db 3 -dr 1 -ca 500 -cp 50). The latter parameters were selected to prevent the haplotypes from being collapsed during the assembly process.

### 2.4.3 | Polishing and purging

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

### 2.4.4 | Chromosome-scale scaffolding

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

### 2.4.5 | Evaluating genome completeness

We used BUSCO v4 (4.0.5) (Seppey et al., 2019) to analyse the completeness of the final and intermediate genome assemblies. Three different gene sets, insecta_odb10.2019-11-20, endopterygota_odb10.2019-11-20, and diptera_odb10.2019-11-20, representing different levels in evolutionary relatedness were considered in the evaluation process. These three gene sets comprise 1367, 2124, or 3285 orthologous genes, respectively.

### 2.4.6 | Exclusion of non-*D. radicum* scaffolds and coassembly of the endosymbiont *Wolbachia*

While assembling the *D. radicum* genome we coassembled the complete genome of *Wolbachia* (Hi-C scaffold 7) a common endosymbiont in arthropods. To obtain a final assembly of *D. radicum* sequences, we excluded Hi-C scaffolds 7, 146 and 370 and trimmed Hi-C scaffold 6 after position 12,881,041 that were annotated to be contaminated with *Wolbachia* sequences during the NCBI validation process. We published Hi-C scaffold 7 as the draft genome of *Wolbachia* separately at the NCBI GenBank (accession CP091195.1). To classify the *Wolbachia* strain associated with *D. radicum*, we located the *ftsZ* nucleotide sequence in Hi-C scaffold 7 using the ftsZf1 (3′-GTTGTCGCAAATACCGATGC-5′) and ftsZr1 (3′-CTTAAGTAAGCTGGTATATC-5′) primer sequences (Werren et al., 1995). The *ftsZ* is one of the five multilocus sequence typing loci (MLST) developed for *Wolbachia* genotyping in arthropods (Baldo et al., 2006). We conducted a phylogenetic analysis including our obtained *ftsZ* sequence, 26 *Wolbachia* sequences representing 13 *Wolbachia* groups and 2 *Ehrlichia* species as an outgroup, equivalent to Konecka et al. (2019). All *ftsZ* sequences were aligned with T-Coffee version 11.00 online (Di Tommaso et al., 2011; Moretti et al., 2007; Notredame et al., 2000; Wallace et al., 2006) by invoking the M-Coffee mode with default settings. Based on this alignment, a phylogenetic tree was reconstructed with the maximum likelihood method using RAxML (version 8.2.12) (Stamatakis, 2014) and "-# 1,000" bootstrap steps and "-o AF221944.2,DQ647000.1" to set both *Ehrlichia* species as outgroup. The resulting phylogenetic tree was visualized using Phylo.io (Robinson et al., 2016).

## 2.5 | Phylogeny – comparative genomics based on BUSCOs

Phylogenetic analyses were done with BuscoOrthoPhylo (https://github.com/PlantDr430/BuscoOrthoPhylo) which is a wrapper

script to concatenate and align protein sequences and to construct a phylogenetic tree based on single-copy BUSCO genes. BUSCOs of the endopterygota_odb10.2019-11-20 gene set, consisting of 2124 genes, were used as the basis for the analysis. In the initial phase complete single-copy BUSCO genes which were shared by 10 selected species (Table 1), were computed. Protein sequences of the shared genes were extracted and concatenated for each species. MAFFT aligner (version 7.475) (Katoh et al., 2002) was run on concatenated FASTA file(s) and finally, RAxML (version 8.2.12) (Stamatakis, 2014) with "-rx_p_sub PROTGAMMAWAG" as model and "-b 100" bootstrap steps was used to reconstruct the phylogenetic tree. The resulting findings were visualized in a phylogenetic tree using Phylo.io (Robinson et al., 2016).

## 2.6 | Sampling, RNA extraction and transcriptome sequencing

All life stage samples were collected from the laboratory culture (section 2.1). We used three replicates per life stage and condition. For the egg stage, we collected 25 mg eggs (laid within 24 h) per replicate. For the larval stage, we collected 18 randomly selected second instar larvae. Nine of the selected larvae were fed on a semi-artificial diet, containing yeast, milk powder, freeze-dried turnip, agar (2:2:2:1) and 90% water. The other nine larvae were reared on the same diet containing 0.4 mg phenylethyl isothiocyanate/g diet. All larvae received freshly prepared diet every other day. After 7 days, larvae were shock-frozen at −80°C and pooled into batches containing three larvae forming three biological replicates per treatment. For the pupal stage, we randomly selected nine freshly formed pupae and pooled them into three biological replicates of three pupae each. For the adult stage, we collected 18 fully developed random adults. Nine individuals were exposed to 35°C (Michaud et al., 1997) for 2 h, whereas the control adults were kept under normal conditions. We pooled three adults for one replicate, resulting in three replicates for control and elevated temperature treatment. We applied relevant and natural stress factors in the two active life stages, host plant defence compounds in the larval stage and heat stress in the adult stage. Exposing individuals to these stress factors, activates a large number of general and specific genes and allowed us to cover a large set of expressed genes for annotation purposes.

All samples were surface sterilized using the same procedure as described for the adult flies. We extracted the total RNA of the larval stage using the ReliaPrep RNA Tissue Miniprep kit (Promega) according to the manufacturer's recommended protocol. Total RNA of all further samples was extracted using TRIzol (Life Technologies) according to the manufacturer's recommended protocol. Qualitative and quantitative RNA assessment of all samples was done by gel electrophoresis (1% agarose), NanoPhotometer P330 (Implen, Munich/Germany) and Qubit 2.0 (Invitrogen).

Library preparation and sequencing of the larval samples (control and stressed larvae) were performed by the Deep Sequencing group of Biotech TU Dresden/Germany on an Illumina NextSeq next-generation sequencer. The poly(A) enriched strand-specific libraries generated for all samples ran on one flow cell generating approximately 50 million paired-end reads of length 75 bp per sample. Egg, pupal and adult (control and stressed) samples were sequenced by Novogene (Hong Kong/China) with strand-specific library preparations and sequencing on an Illumina NovaSeq 6000 next-generation sequencer, generating 20 million paired-end (2 × 150 bp) reads per sample.

## 2.7 | Genome annotation – prediction of protein-coding genes

### 2.7.1 | Mapping of transcriptome data

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

**TABLE 1** Nine insect species (four Diptera, four Lepidoptera, and one Coleoptera species) selected for comparative genomics and phylogenetic analyses. Insect species were chosen according to their phylogenetic relatedness to *D. radicum*, or because they share their host plant range with *D. radicum* or because they are also common pests in agriculture. All nine species are fully sequenced and annotated, and information can be obtained from National Center for Biotechnology (https://www.ncbi.nlm.nih.gov) ([data set] Assembly, 2012)

| Order | Species | NCBI taxid | Common name | 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 |

### 2.7.2 | Homology-based gene prediction

Homology-based GeMoMa (version 1.6.4 and 1.7.2) (Keilwagen et al., 2016, 2018) gene predictions on 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 AnnotationFinalizer 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, https://github.com/TransDecoder/TransDecoder). Predicted transcripts were filtered for a proper start and end of protein-coding transcripts and retaining the UTR annotations by applying the GAF with parameters "f="start=='M' and stop=='*'" atf='' aat=true tf=true". Finally, RNA-Seq-based annotations were formatted with AnnotationFinalizer ("tf=true rename=NO").

### 2.7.4 | Ab initio gene prediction

Additionally, we aimed to predict genes not covered by the homology-based and the transcriptome-based approach, due to a lack of homology or because of no or low expression under the specific conditions of the sampled life stages. To obtain such ab initio gene predictions, we ran RepeatMasker (version 4.1.0, http://www.repeatmasker.org) with RMBlast (version 2.10.0, http://www.repeatmasker.org/RMBlast.html) and "-species insecta -gff -xsmall" to find and mask repetitive sequences annotated for insects in the RepeatMasker repeat database. For ab initio prediction of protein-coding genes on the masked genome sequences, we ran BRAKER (version 1.9) (Brůna et al., 2021; Hoff et al., 2015, 2019), which combines GeneMark (version 4.59_lic) (Lomsadze et al., 2014) and AUGUSTUS (version 3.4.0) (Stanke et al., 2006) with "--gff3 --softmasking" and provided the mapped RNASeq data as hints for initial training of gene models and gene predictions. Predicted transcripts were filtered for proper start and end of protein-coding transcripts by applying the GAF with "f="start=='M' and stop=='*'" atf='"'. Finally, UTR annotations were predicted and added using AnnotationFinalizer with "u=YES rename=NO".

### 2.7.5 | Final genome annotation, supportive filtering, and completeness evaluation

We ran GeMoMa`s GeMoMaAnnotationFilter (GAF) with "f='' atf='' tf=true aat=true" to integrate the predicted gene models from all three applied approaches, the homology-based, the RNASeq-based and ab initio gene prediction approach, and yield a master gene annotation file for the *D. radicum* genome. As gene-related features, we include mRNA, CDS, five_prime_UTR, and three_prime_UTR specificities in the annotation file and several attributes that give additional information on the predicted transcripts and can be used for user-specific filtering. And, finally, we applied AnnotationFinalizer with "tf=true rename=NO" on the integrated gene annotations.

We refined the set of genes predicted by the transcriptome-based and the ab initio approach to the most reliable prediction, by including external evidence based on GO annotation (section 2.8), putative homology to annotated Dipteran proteins (section 2.10), and gene expression levels (section 2.11). In other words, we removed genes that were predicted only by the transcriptome-based or ab initio approach from the final set of genes, if these genes had no GO annotation, no match to an annotated Dipteran protein, and were not (or very lowly) expressed (TPM value <1). We retained only gene predictions of both approaches if the genes were supported by external evidence.

We evaluated the completeness of this final set of protein-coding genes with BUSCO v4 similar to the evaluation of the genome completeness (section 2.4.5), but this time applying the protein mode by setting "-m proteins".

## 2.8 | Functional annotation

Predicted *D. radicum* protein sequences were subjected to PANNZER2 (Protein ANNotation with Z-scoRE) (Törönen et al., 2018), which predicts functional descriptions and GO classes.

Additionally, extracted protein sequences were subjected to InterProScan (version 5.45-80.0.) (Blum et al., 2020; Jones et al., 2014) and scanned for information on protein family and domains in all member data bases (-appl CDD, HAMAP, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITEPATTERNS, SMART, TIGRFAM, Gene3D, SFLD, SUPERFAMILY, MobiDBLite) and for GO- or pathway annotation ("-goterms -iprlookup -pa").

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

### 2.9 | Synteny analysis

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

### 2.10 | Support by homology to dipteran species

We obtained protein sequences of 64 dipteran species from NCBI (Table S1) to create a local Blast database (Altschul et al., 1990). We ran blastp (2.11.0) to align the sequences of proteins predicted for *D. radicum* to that database by setting "evalue=1E-30" to ensure high-quality matches. We interpreted a hit as putative homology and therefore as support for the corresponding coding gene.

### 2.11 | Analysis of life cycle data

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

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

We performed gene ontology (GO) analyses of predefined gene sets using R (version 4.0.4) with the latest version of the R package topGO (2.42.0) (Alexa & Rahnenfuhrer, 2020) with GO.db (3.12.1) (Carlson et al., 2020). We used Fisher's exact test to identify over-represented GO terms. Raw *p*-values were corrected for multiple testing using the method proposed by Benjamini and Yekutieli (2001) implemented in p.adjust contained in the basic R package stats. To get an indication of which processes were active, we aggregated single significant GO-terms (adjusted *p*-value <.05) into self-assigned generic categories. Results were visualized using the R package pheatmap (1.0.12) (Kolde, 2019). For visualization of the results for generic categories, we computed the relative frequency of GO terms determined in a predefined gene set for a generic category. The relative frequency was calculated by the number of significant GO terms in a gene set divided by the total number of GO terms that were sorted into the appropriate generic category.

We defined six gene sets for life stage (eggs, larva, pupa, adults) and condition-specific GO analysis (ITC, heat stress). For the analysis of the whole life cycle, we determined genes that were exclusively expressed in one of the four life stages under control conditions. As we have additional stress conditions in the larval and the adult life stage, we extended the defined gene sets for these two life stages by genes contained in the intersection of both conditions (control and stress) within these stages. For condition-specific GO analysis, we additionally determined the genes exclusively expressed in the stressed condition of the larval and adult stage, respectively. Again, we also extended the stress-specific genes sets by the respective intersection gene set.

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

## 3 | RESULTS AND DISCUSSION

### 3.1 | Genome assembly

PacBio reads classified as eukaryotic (4,454,601 reads) were used for a contamination-free assembly of the *D. radicum* genome with Canu. We expected a high heterozygosity rate due to the pooling of multiple *D. radicum* individuals. Setting Canu parameters accordingly to prevent haplotypes from being collapsed during the assembly process resulted in a raw assembly with the length of approximately

2.538 Gbp, which was almost twice the size of the expected genome, an N50 contig of approximately 205.3 Kbp and in total 29,244 contigs (Table 2). By evaluating the completeness of the raw genome assembly with BUSCO (using three different sets of orthologous genes at different levels of evolutionary relatedness), the raw assembly revealed a completeness of at least 95.7% for the Diptera (Figure 2b, Table S2) and for more than 98% for the Endopterygota gene set (Table S3). These results showed a high completeness of the raw assembly, but also the existence of a reasonable percentage of duplicated sequences.

Improving the sequence quality of the raw genome assembly by performing two rounds of polishing with Arrow increased not only the size of the assembly to approximately 2.544 Gbp (Table 2) but also the completeness of the polished assembly. Especially the percentage of complete genes in the BUSCO Diptera gene set increased to more than 97%. Simultaneously, the number of duplicated genes increased as well (Figure 2b, Table S2).

Next, removing duplicated sequences in the polished assembly with purge_dups successfully reduced the size of the assembly to approximately 1.326 Gbp and a total of 7014 contigs with an N50 of nearly 656.5 Kbp. The size of the purged assembly was already close to the genome size determined with flow cytometry. By evaluating the completeness of the purged assembly, we observed a strong reduction in the percentage of duplicated genes in the Diptera gene set to 6.1% (Figure 2b, Table S2). As a side effect of removing sequences, the completeness of the gene sets dropped slightly to 93.5% (Figure 2b, Table S2).

For the final chromosome-scale assembly, we scaffolded the contigs of the purged assembly with Hi-C data using Juicer and the 3D-DNA genome assembler. The resulting assembly comprised six chromosome-scale contigs (Figure 2a, Table S4), which was consistent with the number of chromosomes determined by karyotyping (Hartman & Southern, 1995), and 2981 smaller, not-assembled contigs. The final assembly of the *D. radicum* genome yielded approximately 1.326 Gbp, where 96.67% of the bases (nearly 1.281 Gbp) were anchored to the six pseudochromosomes. The size of the six pseudochromosomes ranged from one small chromosome with 13 Mbp to five larger chromosomes between 209 and 328.5 Mbp (Table S4). This is in line with the karyotype of *D. radicum*, which comprises five large and one much smaller chromosome (3.3% of the large chromosomes' size) (Hartman & Southern, 1995).

Validation of the final assembly with BUSCO (Tables S2 and S3) showed no considerable change in the number of complete genes, but the number of single-copy genes increased to 92.2% (3030 genes) while the number of duplicated genes decreased to 1.2% (40 genes) for the Diptera gene set. The six pseudochromosomes along with the small contigs were used for all further analyses and are referred to as the *D. radicum* genome hereafter. The number of single-copy BUSCOs of the Diptera gene set in the *D. radicum* genome, was similar to those of other Diptera genomes (Figure 2c, Table S5), indicating that the chromosome-scale genome assembly of *D. radicum* was of comparable quality. Based on our findings, we can conclude that the final *D. radicum* chromosome-scale assembly was accurate, complete and without prokaryotic contamination.

## 3.2 | Phylogeny and synteny

To examine the phylogenetic relationship of *D. radicum* to other insects, we compared complete single-copy BUSCOs of the Endopterygota gene set (comprising a total of 2124 genes) shared by the selected nine insect species belonging to Diptera (4), Coleoptera (1) and Lepidoptera (4, Table 1). We identified 1217 (Table S6, Table S7) shared, and therefore conserved, single-copy genes (Figure 3a, Table S6). Reconstruction of the evolutionary relationships among these ten species based on the shared gene sets revealed that the root fly *D. radicum* was most closely related to the blowfly, *L. cuprina*, followed by the house fly, *M. domestica*, and the fruit fly, *D. melanogaster* (Figure 3a). These relations were consistent with their taxonomic position (Wiegmann et al., 2011).

In our synteny analysis, we successfully mapped the six pseudochromosomes of *D. radicum* to the six Muller elements of *D. melanogaster* (Figure 3b). This was achieved by predicting gene models

**TABLE 2** Summary of assembly statistics. The raw, polished, and purged assemblies are intermediate assemblies after PacBio read assembly with Canu, two rounds of polishing with arrow, and purging with purge_dups. The final, chromosome-scale assembly, generated with the 3D-DNA genome assembly pipeline that assembled contigs of the purged assembly by integration of Hi-C Illumina reads into (chromosome-scale) scaffolds. The final chromosome-scale assembly contains 6190 gaps of length 100 bp, whereby 6188 gaps are located on the six pseudochromosomes

| Assembly | Number of bases | Number of contigs[a] or scaffolds[b] | N50 | L50 | N90 | L90 | Longest contig[a] or scaffold[b] |
|---|---|---|---|---|---|---|---|
| Raw assembly | 2,538,077,247 | 29,244[a] | 205,306 | 2197 | 32,594 | 16,335 | 6,127,675 |
| Polished assembly | 2,544,504,558 | 29,244[a] | 205,665 | 2201 | 32,715 | 16,338 | 6,133,028 |
| Purged assembly | 1,325,508,377 | 7014[a] | 656,541 | 485 | 74,470 | 2765 | 6,133,028 |
| Final chromosome-scale assembly | 1,326,127,377 | 2987[b] | 242,504,274 | 3 | 208,954,159 | 5 | 328,483,116 |
| 6 pseudochromosomes only | 1,281,926,506 | 6[b] | 242,504,274 | 3 | 208,954,149 | 5 | 328,483,116 |

[a]Numbers given for the raw, polished and purged assembly refer to contigs.

[b]Numbers given for the chromosome-scale assembly and the six pseudochromosomes refer to scaffolds.

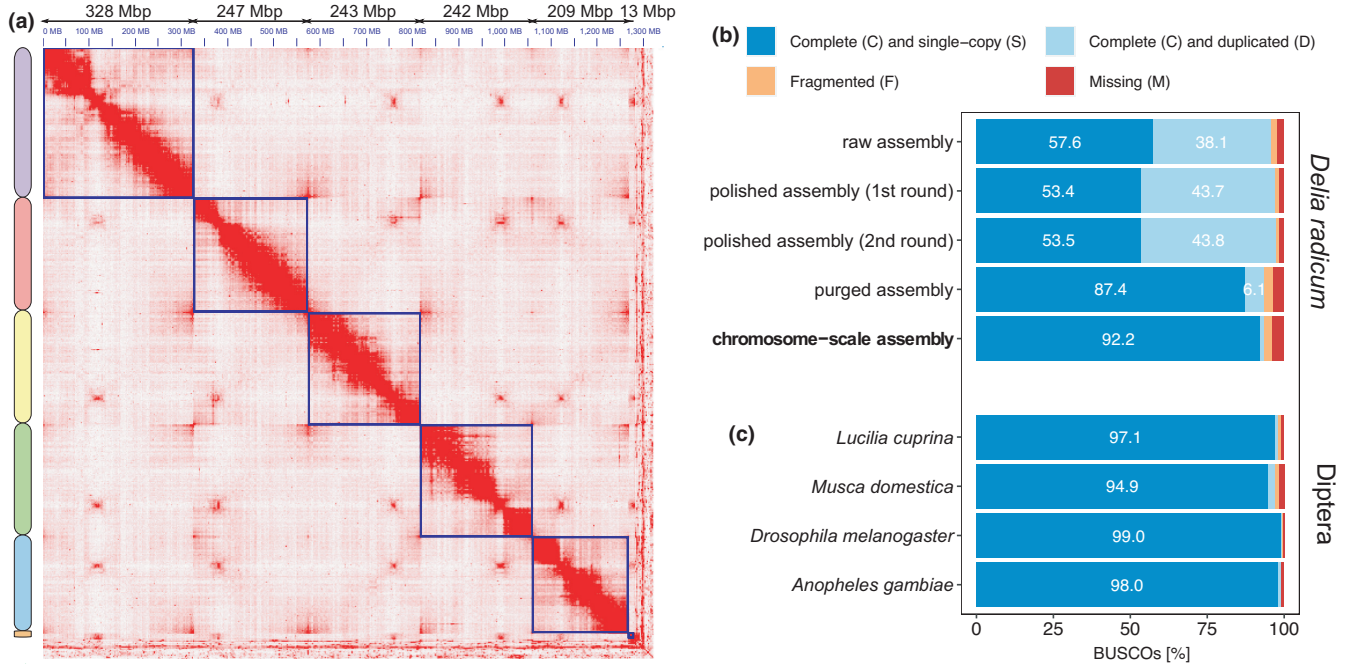


**FIGURE 2** Chromosome-scale assembly of the *Delia radicum* genome. (a) Heat map showing the Hi-C contacts map of the final chromosome-scale assembly, where the six chromosomes (six super-scaffolds) are indicated by the blue boxes. The chromosomes are ordered from largest to smallest; their concrete lengths are given in Mbp above the Hi-C map. (b) Bar plot showing the result of BUSCO analyses of the intermediate and final assemblies using the "Diptera" gene set containing 3285 genes. Numbers in the bars give the percentage of genes found for the category indicated by the colour of the bar. (c) Bar plot showing the result of BUSCO analyses using the "Diptera" gene set of four other dipteran species with published genomes. Numbers in the bars give the percentage of genes found for the category indicated by the colour of the bar

(Table 1) in the *D. radicum* genome based on the annotated *D. melanogaster* genome using GeMoMa (Table S8). Genes annotated on the Muller element A (X chromosome) of *D. melanogaster* mapped successfully on the second-largest chromosome (HiC_scaffold_2) in the *D. radicum* assembly. Genes annotated for the other *D. melanogaster* Muller elements were mainly localized on the remaining four larger *D. radicum* chromosomes (Figure 3b). For the smallest chromosome (HiC_scaffold_6) we found indications that this might be related to the Muller element F (chromosome 4) (NC_004353.4) of *D. melanogaster* (Table S8).

## 3.3 | Genome annotation and functional gene annotation

### 3.3.1 | Process of genome annotation and evaluation

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

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

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

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

The integration of the predictions of all three approaches into a comprehensive annotation led to 81,000 putative genes covering 121,731 transcripts (Table 3), where a relatively high number of putative genes was predicted specifically by the ab initio approach (Figure 4a). To have a more reliable set of genes, we excluded 18,578 putative genes predicted by the transcriptome-based or the ab initio approach, which are not supported by external evidence (Table S9). The final set comprised 62,422 supported genes covering 103,130

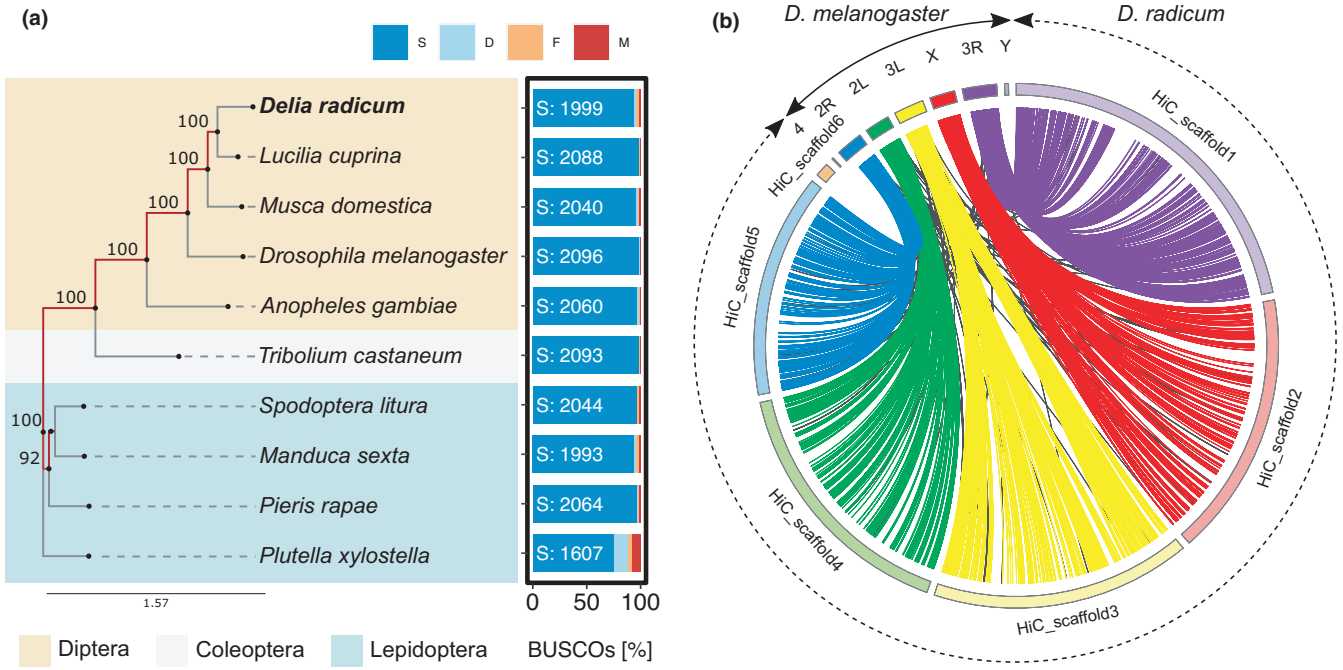


**FIGURE 3** Phylogenetic analyses. (a) A phylogenetic tree reconstructed with RAxML on concatenated alignments of proteins of 1271 genes of BUSCOs' Endopterygota gene set ($n = 2124$) shared by all 10 insect species. Tree reconstruction was done including 100 bootstrapping steps. The level of bootstrapping support is given at the edges. The bar plot to the right of the phylogenetic tree shows BUSCO results of each species on the Endopterygota gene set, where S, number of complete single-copy BUSCO genes (dark blue bar); D, duplicated complete copy genes (light blue); F, fragmented genes (orange); M, missing genes (red). (b) A Circos plot linking genes on the assembled scaffolds of *Delia radicum* (HiC_scaffold 1–6) to homologues on the *Drosophila melanogaster* chromosomes (2R/2L [Muller elements C and B], 3R/3L [Muller elements E and D], 4 [Muller element F], X and Y [Muller element A]). Each line connects homologous regions of at least two consecutive genes. Coloured lines indicate that homologous regions of a *D. melanogaster* chromosome are connected to those of the syntenic chromosome of *D. radicum*. Otherwise, they are coloured in black

**TABLE 3** Summary of gene prediction statistics. Number of gene predictions made on the chromosome-scale genome assembly of *D. radicum* by the three different approaches: GeMoMa a sequence homology-based approach, Cufflinks an RNASeq data-based approach to assemble transcriptomes, and BRAKER an approach for ab initio predictions of genes. The final comprehensive gene annotation for the *D. radicum* genome contains 62,422 putative genes that are supported by external evidence

| Approach | Description | Number of transcripts | Number of genes |
|---|---|---|---|
| Cufflinks | Transcriptome-based | 23,729 | 16,188 |
| GeMoMa | Homology-based | 46,286 | 19,343 |
| BRAKER | Ab initio | 82,473 | 72,613 |
| Final | Raw | 121,731 | 81,000 |
| Final | Supported[a] | 103,130 | 62,422 |

[a]Supported by any external evidence or predicted by the homology-based approach.

transcripts (Table 3). Nearly 95.68% of the genes were located on the six chromosomes (Table S4).

Evaluation with BUSCO showed that our genome annotation covered 93.6% complete-copy genes of the Diptera gene set, and 95.4% of the Endopterygota gene set (Table S10). By determining the overlap of the predictions, we found 7,129 genes that were predicted by all three approaches and a total of 13,153 genes by at least two approaches (Figure 4a). The annotation of the latter set of genes covered 87.5% of complete-copy genes of the Diptera gene set and 89.5% of the Endopterygota gene set (Table S10). Only the combination of all three approaches led to a complete annotation of the *D. radicum* genome. BUSCO completeness was not affected by our applied filtering setup of the annotated genes.

### 3.3.2 | Functional annotation

Overall, 84.4% (52,689) of the genes were functionally annotated with at least one GO term and/or protein family or domain information (Figure 4b, Figure S2, Table S11), including 80.06% (32,648) of the only ab initio predicted genes.

Focusing on the expressed genes by using our in-house whole life stage RNASeq data, we found that 30,492 genes (48.85%) had an estimated expression of ≥1 transcript per million (TPM) (Figure 4c). A high number of genes was predicted by BRAKER only; however, most of these genes were not expressed under our conditions, although the absolute number of expressed genes in the ab initio set is higher than in the other sets (Figure 4c). From the set of expressed genes, 50.08% (15,270) were functionally annotated with at least one GO term (Figure 4d).

Taken together, these findings indicate that our gene annotation is complete and accurate. We will demonstrate its applicability to generate biologically relevant information in the following section. We will do so by analysing the transcriptomes of all life stages of *D. radicum* to identify life stage-specific functional gene expression underlying adaptations to their stage-specific demands, especially adaptations to their host plant defences and global warming.

### 3.4 | Expression analysis over the *D. radicum* life cycle

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

We also found that the total number of expressed genes differed among life stages (Table 4, Figure 5b). The lowest total number of expressed genes was detected in the egg stage and the highest in the larval and pupal stages (Table 4, Figure 5b, horizontal bar plot). When looking at the overlap among the life stages, we found 31.6% of the 30,492 genes to be expressed across all life stages (Figure 5b, vertical bar plot). Another 36% of the genes were exclusively expressed in either a single life stage or condition, in the intersection of both conditions of the larval and the adult stage, respectively or specifically in the egg and pupal stages (Figure 5b, Figure S3). In the UMAP plot (Figure 5c), genes expressed in single life stages were located at the top and formed life stage-specific clusters, whereas genes expressed in all life stages also clustered, but were located on the opposite side. The remaining one-third of the genes (not shown in Figure 5b, included in Figure S3) clustered in between. For larval and adult stages, we observed again that genes expressed under different conditions clustered closely together and formed life stage-specific clusters (Figure 5c).

An ontology-based gene expression analysis revealed life-stage specific groups related to biological processes (BP), molecular functions (MF) and cellular components (CC, Figure 6, Figure S4, Table S12). In the egg stage, mainly genes involved in embryonic development (BP), transcriptomic activity (MF) and genetic material (CC, Figure 6) were expressed. In particular, genes belonging to the GO biosynthetic processes DNA biosynthesis, metabolic processes, eggshell layer formation (amnioserosa formation) and organ development (muscle and organ formation) were activated (Figure S4a). These processes are involved in the transition from embryo to larva, which requires active cell division and involves a broad range of metabolic processes to synthesize cell components, membranes and organs (Beutel et al., 2013). These structures require different macromolecules; indeed, we found several expressed genes related to molecular biosynthesis processes in the eggs (Figure S4a). Cell

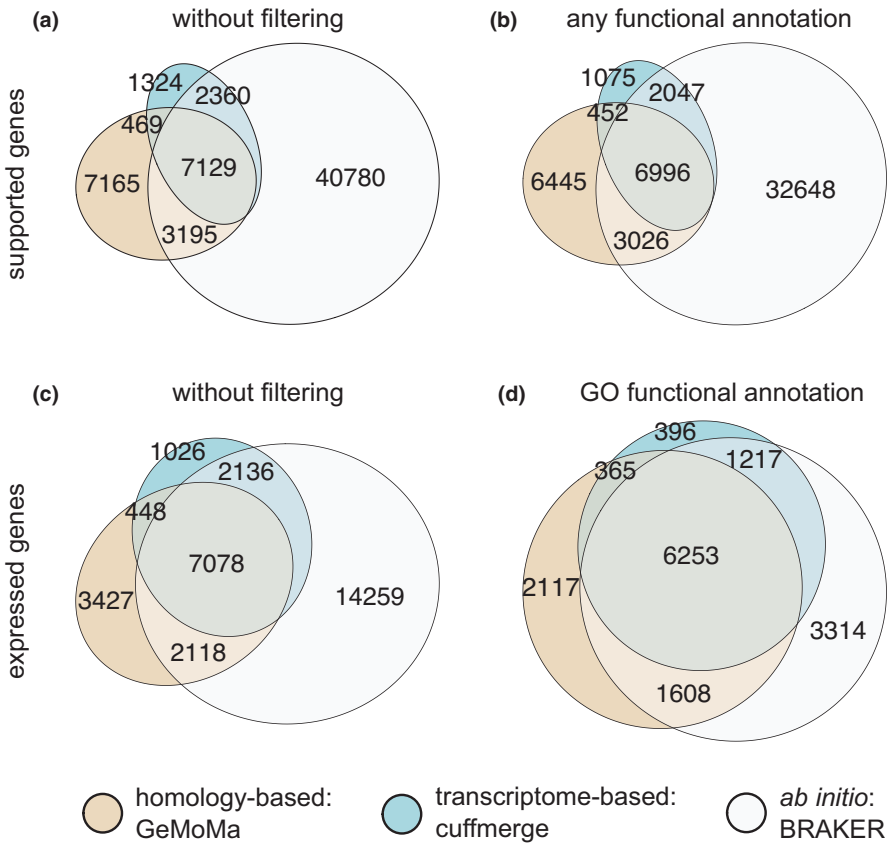


**FIGURE 4** Venn diagrams containing the numbers of genes in the *Delia radicum* genome predicted by homology-based, transcriptome-based or ab initio approaches, or a combination thereof. The numbers of genes in the diagrams are based on all predicted genes that are supported by external evidence; all predicted genes (that are supported by external evidence) with any functional annotation, which includes GO annotation and/or protein family or domain annotation; predicted genes that were expressed with a transcript per million (TPM) value ≥1. TPM values result from analyses of our in-house life cycle RNASeq data; expressed genes with a functional annotation based on GO annotation

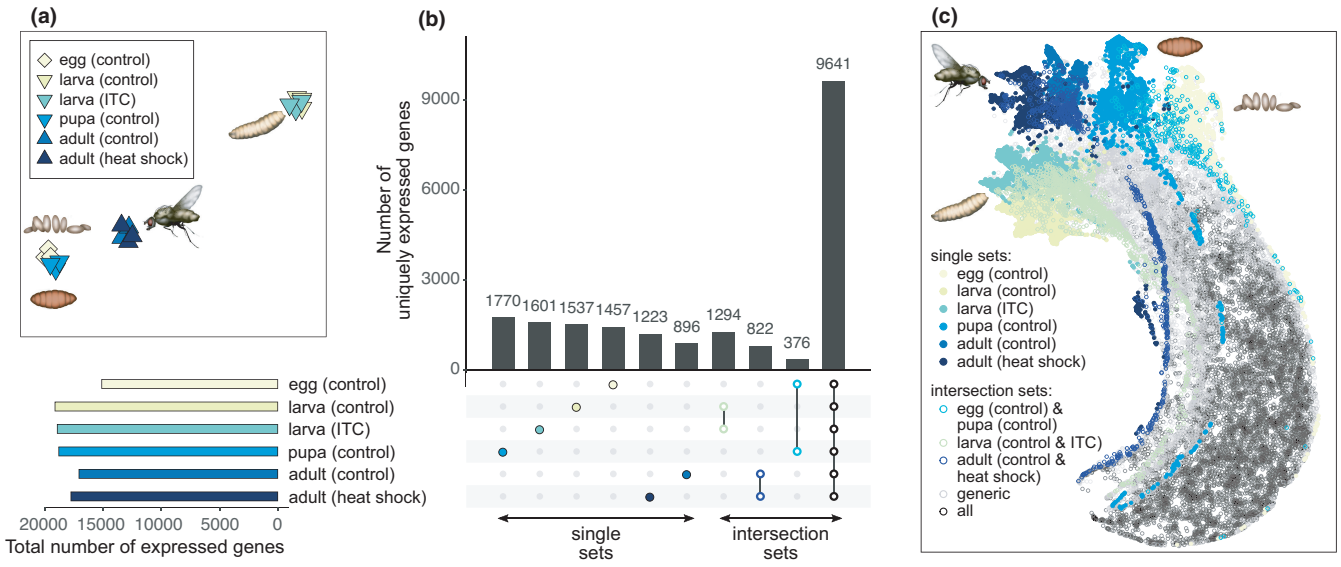


**FIGURE 5** Differences in gene expression profiles among *Delia radicum* life stages and stress conditions. (a) Uniform manifold approximation and projection (UMAP) plot showing differences among the life stages based on differing gene expression. ITC, larvae fed on diets with 2 μM phenylethyl isothiocyanate in their diet. (b) UpSet plot showing the number of genes that are exclusively expressed (Transcripts Per Million (TPM) value ≥1) in at least one replicate of a life stage or a stress condition (first 6 bars, filled circles); expressed in both conditions within larva and adult life stages (bar 7 and 8, green and dark blue open circles), or both in eggs and pupae (bar 9, cyan open circles), and those expressed in all 18 samples (last bar, open black circles). A selection of intersection sets is shown, whereas the full set is presented in Figure S3. To the left, the total number of expressed genes per life stage and stress condition is shown (coloured horizontal bar plot below subfigure a). The remaining genes referred to as generic, are not shown and sum up to 9875 genes. (c) UMAP of expressed genes. Genes are coloured according to the sets in (b) and are plotted with filled circles when they belong to single sets and with open circles when they belong to intersection sets. Genes expressed in all 18 samples are labelled as "all" (black open circles). The remaining genes are labeled as "generic" (grey open circles)

**TABLE 4** Summary of gene ontology (GO) annotations of expressed genes. In total 30,492 genes of the 62,422 genes (81,000 raw genes) were expressed (transcripts per million [TPM] value ≥1) in our in-house life stages RNASeq data set. Genes were annotated to GO classes using PANNZER2 and InterProScan. Genes exclusively expressed in one specific life stage were grouped into gene sets named according to the life stage. For the larval (isothiocyanate [ITC] in diet) and the adult (heat shock) stage where control and stress conditions are present in the data set, genes that are expressed in both conditions within one life stage were added to the life stage and condition-specific gene sets. Numbers of the row labelled with "total" are in concordance with Figure 5b. Listed gene sets were used for life stage-specific GO enrichment analyses

| Set/ontology | Complete | Egg (control) | Larva (control) | Pupa (control) | Adult (control) | Larva (ITC) | Adult (heat shock) |
|---|---|---|---|---|---|---|---|
| Total | 30,492 | 1457 | 2831 | 1770 | 1718 | 2895 | 2045 |
| noGO | 15,222 | 1029 | 2021 | 1288 | 1096 | 2079 | 1225 |
| GO | 15,270 | 428 | 810 | 482 | 622 | 816 | 820 |
| BP | 11,262 | 262 | 501 | 289 | 406 | 495 | 534 |
| MF | 13,513 | 397 | 717 | 429 | 537 | 723 | 729 |
| CC | 10,917 | 230 | 444 | 260 | 381 | 434 | 454 |

Abbreviations: BP, biological process; CC, cellular compartment; MF, molecular function.

differentiation and organ formation require regulation, coordination and binding activation (Izumi et al., 1994) which was reflected in our BP expression data (Figure 6, Figure S4a).

Genes involved in the body development (BP), structural and transposase function (MF), and extracellular matrix (CC) were more frequently expressed in pupae (Figure 6). These genes belong to GOs comprising regulators, binding activity, biosynthesis, metabolism and DNA amplification (Figure S4). During the pupal stage, metamorphosis results in the "disassembly" of larval structures to form adult wings, compound eyes and legs (Buszczak & Segraves, 2000; Chapman & Chapman, 1998). This requires the expression of genes involved in catabolic processes, as well as in organ and cuticle formation. Indeed, we found an increased expression of genes responsible for nuclease and peptidase activity (MF) and chitin-based cuticle structures (CC, Figure 6, Figure S4). This is in line with the gene expression profiles in *D. melanogaster* pupae (Arbeitman et al., 2002).

Genes connected to the metabolic processes (BP) were highly expressed in the larval stage (Figure 6). We found genes coding for peptidases and polymerases (MF), involved in DNA processes or functions and biosynthetic processes (BF) to be highly expressed (Figure 6, Figure S4). These genes are likely related to feeding and digestion as well as to growth and molting, which are the main processes in the larval stage (Chapman & Chapman, 1998; Chen, 1966). In larvae exposed to the plant toxin ITC, we found that peptidase genes (MF) and genes involved in metabolic and biosynthetic processes (BP) were activated (Figure 6, Figure S4). These enzymes may be involved in catabolizing plant toxins as has been described for other herbivores feeding on *Brassica* plants (Schramm et al., 2012).

Genes coding for the detection of visible and UV-light, optomotor capability, detection of chemical stimuli (taste, smell) and temperature (BP) were exclusively expressed in adults (Figure 6, Figure S4a). The expression of these gene sets, which are involved in the sensory, optomotor and nervous systems (BP), is important to localize food sources and suitable hosts for oviposition (Gouinguene & Städler, 2005, Gouinguene & Städler, 2006; Roessingh & Städler, 1990). In addition, several genes coding for receptors and ion channels were expressed (Figure 6, MF). These genes are involved in the detection of environmental stimuli and signal transmission via the nervous cells to the brain (Sato & Touhara, 2008). Specific for the adult life stage was also the expression of adult behaviour-linked genes (Figure S4a).

Exposing adult flies to a higher temperature resulted in the enhanced expression of peptidases, ion binding (MF), sensory system, especially smell and egg formation (BP) related genes compared to control adults (Figure 6, Figure S4). High temperatures alter protein stability, structures and folding, followed by functional changes (Jaenicke et al., 1990). The activation of peptidases might avoid the malfunction of proteins under heat stress. Temperature changes affect also the volatility of volatile organic compounds (VOCs) as well as the emission rates of plants (Copolovici & Niinemets, 2016). Since adults of *D. radicum* are attracted by VOCs to localize host plants (Finch, 1978), the enhanced expression of genes related to VOC perception (smell) in flies might indicate an adaptive response to a higher temperature. Investing in offspring, under these circumstances might ensure the survival of the fly population. To localize a possible host plant for their oviposition, *D. radicum* females utilize odor signals (Nottingham, 1988).

## 3.5 | Coassembly of the *D. radicum* associated endosymbiont *Wolbachia*

Together with the genome of *D. radicum*, we assembled the genome of a *Wolbachia* species, a very common endosymbiont in insects (Werren & Windsor, 2000). The coassembled *Wolbachia* genome consisted of a single contig with 1.59 Mbp matching to the size of *Wolbachia* genomes discovered in other arthropods (~1.4–1.6 Mbp)

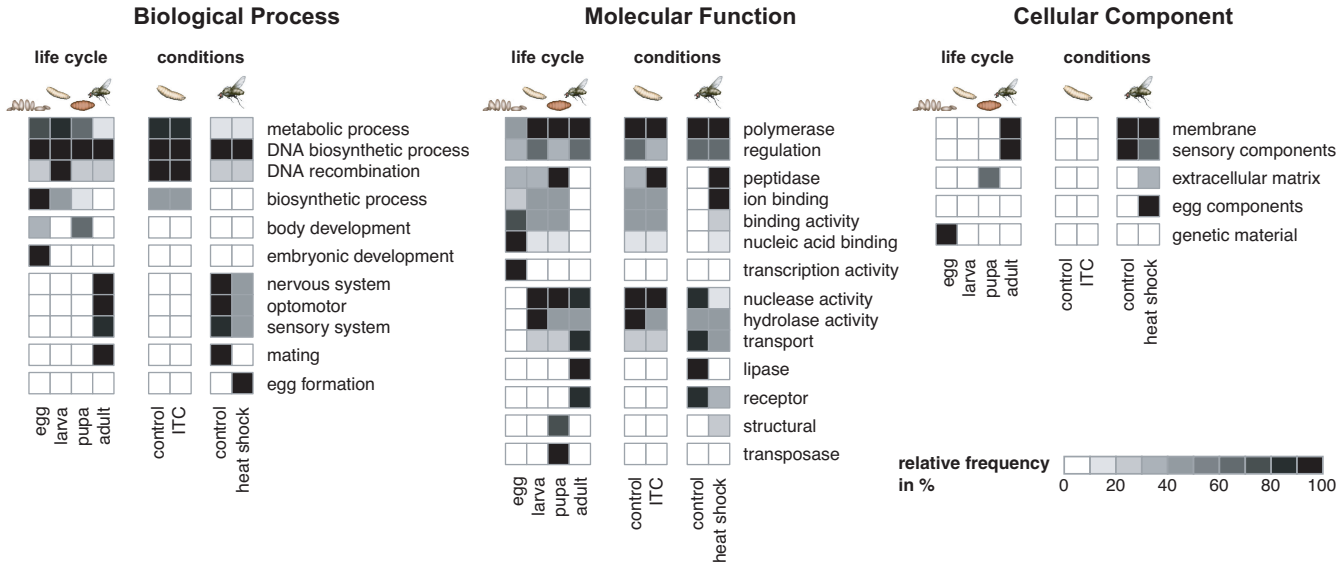


FIGURE 6 Gene ontology (GO) analyses on the biological process (BP), molecular function (MF) and cellular component (CC) ontologies, based on expressed genes (transcripts per million [TPM] value ≥1). Results are shown in three respective heat maps, where rows are labelled by generic categories and columns with life stages and/or conditions. Explicit GO annotations of expressed genes are collapsed into more generic categories. Hence, each cell in a heatmap contains the relative frequency of GO terms sorted into a specific generic category for a specific life stage and/or condition. Only GO terms that were significantly overrepresented in a GO-enrichment analysis (Fisher's exact test, $p < .05$ after correction with Benjamini - Yekutieli) are considered. Expanded versions of the heat maps, where detailed GO annotations for each generic category are listed, are provided in Figure S4. In all heat maps, the block with four columns to the left shows the results of all stages of the life cycle under control conditions, whereas the columns to the right show the relative frequencies determined for larvae and adults under control or stress conditions; the data for the control conditions in larval and adult stage are duplicated for easier comparison. ITC = larvae fed on diet with 2-phenylethyl isothiocyanate

(Lo et al., 2002). Based on the *ftsZ* sequence, the *Wolbachia* strain of *D. radicum* clustered in the supergroup A (Figure S5) close to *Wolbachia* strains occurring in *D. melanogaster*, *Telema cucurbitina* (Araneae) and *Phyllonorycter blancardella* (Lepidoptera). Infections with *Wolbachia* endosymbionts can affect their host's fecundity, body mass and sex ratio either positively or negatively (Werren et al., 2008). For *D. radicum* the effect of *Wolbachia* symbionts on its host's performance was experimentally assessed using genetically similar *Wolbachia*-free (W−) and *Wolbachia*-infested (W+) lines (Lopez et al., 2018). *Wolbachia* infection reduced egg hatching rate but increased larval survival. This had the consequence that the overall performance of the W− and W+ lines were similar (Lopez et al., 2018). Using the same two lines, it was shown that *Wolbachia* also affects the bacterial community of *D. radicum* (Ourry et al., 2021). The frequency of *Erwinia* bacteria strongly decreased, whereas *Providencia* and *Serratia* bacteria increased in W+ lines (Ourry et al., 2021). Bacteria, in particular those in the gut (Sontowski & van Dam, 2020), may play an important role in digestive and detoxification processes. *Wolbachia*-induced changes in the bacterial community thus may cause indirect effects on host performance. However, the experimental results obtained by Lopez et al. (2018), do not point to such indirect effects.

## 4 | CONCLUSION

An increasing number of assembled and annotated insect genomes have been published over the last decade. However, genomes of belowground insects and especially root-feeding herbivores are underrepresented. We sequenced the genome of a belowground-feeding agricultural pest, the cabbage root fly *Delia radicum*, whose larvae are also used as a "model" belowground herbivore in studies on optimal defence allocation and systemic induced responses in plants. Using PacBio and Hi-C sequencing, we generated a 1.3 Gbp assembly with an N50 of 242 Mbp, six pseudochromosomes and 13,153 annotated genes using homology-, transcriptome- and model-predicted approaches, predicted by at least two approaches. During the assembly process, we identified one Hi-C scaffold as the genome of a *Wolbachia* species (1.59 Mbp), a very common endosymbiont in insects (Werren & Windsor, 2000). Such coassembled endosymbiont genomes can be valuable to understanding host-symbiont interactions and their roles in other interactions such as host-plant adaptations.

Our accurate and well-annotated genome can reveal genetic mechanisms underpinning the adaptation of *D. radicum* to its host plants and their specific chemical defences, the glucosinolate-isothiocyanate system. With our work, we provide a tool to understand how the different life stages of this herbivore have adapted to their host plants by identifying adult-specific genes involved in olfactory orientation or the detoxification of plant defence compounds in larvae. The genome and the transcriptomes can further be used to understand adaptation to specific conditions, *i.e.* the evolution of pesticide resistance and adaptive responses to environmental stress factors, such as temperature increase or soil pollution. This high-quality genome is also an important tool to develop novel strategies to combat this pest, for example highly specific dsRNA-based pesticides, which can discriminate between target and non-target species. Moreover, the genus *Delia* contains several other pest species, such as the turnip root fly *D. floralis*, the onion fly *D. antiqua* and the seed bulb maggot, *D. platura*. As their common names indicate, they attack a range of crops. The genome of *D. radicum* is an excellent foundation to further explore the genetic mechanisms underlying adaptation to chemical host-plant defences among members of the genus *Delia*.

### ACKNOWLEDGEMENTS

We thank the Long Read Team of the DRESDEN-concept Genome Center, DFG NGS Competence Center, part of the Center for Molecular and Cellular Bioengineering (CMCB), Technische Universität Dresden and the MPI-CBG, especially Sylke Winkler for their great support and kind collaboration. Dominik Jakob is acknowledged for his assistance with the insect culture. Great thanks to Denis Tagu and Fabrice Legeai (INRAE, Rennes, France), Denis Poinsot (University of Rennes, France) and Ekaterina Shelest (BIU, iDiv, Leipzig, Germany) for their helpful advice and encouragement in the earlier stages of this project. We also greatly thank Jens Keilwagen (JKI, Quedlinburg) for valuable discussions on gene prediction and support with GeMoMa. We thank two anonymous reviewers and the editor Benjamin Sibbett for their helpful comments on previous versions, which improved our manuscript. This study was funded by the German Research Foundation (DFG) Collaborative Research Center 1127 ChemBioSys (project number 09161509) to RS and NvD, and the German Centre for Integrative Biodiversity Research (iDiv) funded by DFG, grant no. FZT 118, 202548816) to RS, YP, CG, and NvD. HV and YO thank the Max-Planck-Gesellschaft for funding. Open access funding enabled and organized by ProjektDEAL.

### AUTHOR CONTRIBUTIONS

Yvonne Poeschl, Rebekka Sontowski, Nicole M. van Dam designed the project, Anne-Marie Cortesero provided the starting culture of the insects, Rebekka Sontowski, Heiko Vogel performed the laboratory work, Yvonne Poeschl, Cervin Guyomar, Yu Okamura, Heiko Vogel performed the data processing and analysis, Yvonne Poeschl created the figures, Yvonne Poeschl, Rebekka Sontowski, Nicole M. van Dam, Heiko Vogel wrote a first version. All authors contributed to the writing process.

### BENEFIT-SHARING STATEMENT

Benefits generated: A research collaboration was developed with scientists from France providing a starting colony of *D. radicum* flies 7 years ago, all collaborators are included as coauthors. The results of research have been shared with the provider and the broader scientific community (see above). Additionally, benefits from this research accrue from the sharing of our data and results on public databases as described above.

## DATA AVAILABILITY STATEMENT

Genome sequences of *D. radicum* have been submitted to the National Center for Biotechnology (NCBI, WGS project: JAEKOY000000000, GenBank assembly accession: GCA_021234595.1) within the BioProject PRJNA655405. Raw sequence data used for genome assembly (PacBio sequences and Illumina Hi-C sequences) have been made available at NCBI (https://www.ncbi.nlm.nih.gov/sra) in the genome-related BioProject (PRJNA655405), whereas raw sequence data used for annotation (Illumina RNASeq sequences) have been deposited in BioProject PRJNA736225. Sample metadata have also been deposited in the corresponding BioProjects using the package Invertebrate version 1.0. Final genome annotation, respective annotations by GeMoMa, Cufflinks and BRAKER, information on support of gene predictions by external evidence, and also functional transcript annotations made by InterProScan and PANNZER2 are available via Zenodo (https://doi.org/10.5281/zenodo.5706211).

Additionally, the genome sequence of *Wolbachia* was submitted to the NCBI and linked to the BioProject (PRJNA655405) of the genome assembly of *D. radicum*. Corresponding sample metadata was deposited using the package Microbe version 1.0. The coassembled genome of *D. radicum's* endosymbiont is available under GenBank assembly accession GCA_021609905.1.

## CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## ORCID

*Rebekka Sontowski* https://orcid.org/0000-0001-5791-8814
*Yvonne Poeschl* https://orcid.org/0000-0002-6727-6891
*Yu Okamura* https://orcid.org/0000-0001-6765-4998
*Heiko Vogel* https://orcid.org/0000-0001-9821-7731
*Cervin Guyomar* https://orcid.org/0000-0003-2707-2541
*Nicole M. van Dam* https://orcid.org/0000-0003-2622-5446

## REFERENCES

Alexa, A., & Rahnenfuhrer, J. (2020). topGO: enrichment analysis for gene ontology. *R package version, 2.42.0(0)*.

Allema, B., Hoogendoorn, M., van Beek, J., & Leendertse, P. (2017). *Neonicotinoids in European agriculture. Main applications, main crops and scope for alternatives*. Retrieved from Culemborg, The Netherlands.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Arbeitman, M. N., Furlong, E. E. M., Imam, F., Johnson, E., Null, B. H., Baker, B. S., Krasnow, M. A., Scott, M. P., & Davis, R. W. & White, K. P. (2002). Gene expression during the life cycle of drosophila melanogaster. *Science*, *297*(5590), 2270–2275. https://doi.org/10.1126/science.1072152

Assembly [Internet]. : National Library of Medicine (US), National Center for Biotechnology Information; 2012 – [cited 2021 Jun 10]. Available from: https://www.ncbi.nlm.nih.gov/assembly/ [dataset].

Baldo, L., Hotopp, J. C. D., Jolley, K. A., Bordenstein, S. R., Biber, S. A., Choudhury, R. R., Hayashi, C., Maiden, M. C., Tettelin, H. & Werren, J. H. (2006). Multilocus sequence typing system for the endosymbiont *Wolbachia pipientis*. *Applied and Environmental Microbiology*, *72*(11), 7098–7110. https://doi.org/10.1128/AEM.00731-06

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 1165–1188. https://doi.org/10.1214/aos/1013699998

Beutel, R. G., Friedrich, F., Yang, X.-K., & Ge, S.-Q. (2013). *Insect morphology and phylogeny: a textbook for students of entomology*. Walter de Gruyter Berlin/Boston.

Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G. A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D. H., Letunic, I., Marchler-Bauer, A., … Finn, R. D. (2020). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*, *49*(D1), D344–D354. https://doi.org/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. https://doi.org/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. 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), https://doi.org/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. https://doi.org/10.1016/S0960-9822(00)00792-2

Capinera, J. L. (2008). *Encyclopedia of entomology*, 2nd 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). 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. https://doi.org/10.1007/s10886-015-0601-y

Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J.-M., Taly, J.-F., & Notredame, C. (2011). T-Coffee: A web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Research*, *39*(suppl_2), W13–W17. https://doi.org/10.1093/nar/gkr245

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., Batut, P., Chaisson, M., & Gingeras, T. R. (2012). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., Aiden, A. P. & Aiden, E. L. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*(6333), 92–95. https://doi.org/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. 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. https://doi.org/10.2135/cropsci2005.0371

Ferry, A., Dugravot, S., Delattre, T., Christides, J.-P., Auger, J., Bagnères, A.-G., Poinsot, D., & Cortesero, A.-M. (2007). Identification of a widespread monomolecular odor differentially attractive to several *Delia radicum* ground-dwelling predators in the field. *Journal of Chemical Ecology*, *33*(11), 2064–2077. https://doi.org/10.1007/s10886-007-9373-3

Finch, S. (1978). Volatile plant chemicals and their effect on host plant finding by the cabbage root fly (Delia Brassicae). *Entomologia Experimentalis Et Applicata*, *24*(3), 350–359. https://doi.org/10.1111/j.1570-7458.1978.tb02793.x

Finch, S., & Ackley, C. M. (1977). Cultivated and wild host plants supporting populations of the cabbage root fly. *Annals of Applied Biology*, *85*(1), 13–22. https://doi.org/10.1111/j.1744-7348.1977.tb00626.x

Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, *3*, 294–299.

Fournet, S., Stapel, J., Kacem, N., Nenon, J., & Brunel, E. (2000). Life history comparison between two competitive Aleochara species in the cabbage root fly, *Delia radicum*: implications for their use in biological control. *Entomologia Experimentalis Et Applicata*, *96*(3), 205–211. https://doi.org/10.1046/j.1570-7458.2000.00698.x

Gehlenborg, N. (2019). UpSetR: a more scalable alternative to venn and euler diagrams for visualizing intersecting sets. Available at: cran.r-project.org/package=UpSetR. (21 March 2021, date last accessed).

Gouinguene, S. P. D., & Städler, E. (2005). Comparison of the sensitivity of four Delia species to host and non-host plant compounds. *Physiological Entomology*, *30*(1), 62–74. https://doi.org/10.1111/j.0307-6962.2005.00432.x

Gouinguené, S. P. D., & Städler, E. (2006). Comparison of the egg-laying behaviour and electrophysiological responses of *Delia radicum* and *Delia floralis* to cabbage leaf compounds. *Physiological Entomology*, *31*(4), 382–389. https://doi.org/10.1111/j.1365-3032.2006.00532.x

Griffiths, G. (1991). *Economic assessment of cabbage maggot damage in canola in Alberta*. Paper presented at the Proceedings of the GCIRC Eighth International Rapeseed Congress.

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, *36*(9), 2896–2898. https://doi.org/10.1093/bioinformatics/btaa025

Hare, E. E., & Johnston, J. S. (2011). Genome size determination using flow cytometry of propidium iodide-stained nuclei. In V. Orgogozo, & M. V. Rockman (Eds.), *Molecular Methods for Evolutionary Genetics* (pp. 3–12). Humana Press.

Hartman, T. P. V., & Southern, D. I. (1995). Genome reorganization from polyteny to polyploidy in the nurse cells found in onion fly (*Delia antiqua*) and cabbage root fly (*Delia radicum*) ovaries (Diptera, Anthomyiidae). *Chromosome Research*, *3*(5), 271–280. https://doi.org/10.1007/BF00713064

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2015). BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, *32*(5), 767–769. https://doi.org/10.1093/bioinformatics/btv661

Hoff, K. J., Lomsadze, A., Borodovsky, M., & Stanke, M. (2019). *Whole-genome annotation with BRAKER* Gene prediction, Vol. *1962* (pp. 65–95). Springer.

Hopkins, R. J., van Dam, N. M., & van Loon, J. J. (2009). Role of glucosinolates in insect-plant relationships and multitrophic interactions. *Annual Review of Entomology*, *54*, 57–83. https://doi.org/10.1146/annurev.ento.54.110807.090623

Izumi, S., Yano, K., Yamamoto, Y., & Takahashi, S. Y. (1994). Yolk proteins from insect eggs: structure, biosynthesis and programmed degradation during embryogenesis. *Journal of Insect Physiology*, *40*(9), 735–746. https://doi.org/10.1016/0022-1910(94)90001-9

Jaenicke, R., Heber, U., Franks, F., Chapman, D., Griffin, M. C. A., Hvidt, A., Cowan, D. A., Laws, R. M., & Franks, F. (1990). Protein structure and function at low temperatures. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *326*(1237), 535–553. https://doi.org/10.1098/rstb.1990.0030

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Kapranas, A., Sbaiti, I., Degen, T., & Turlings, T. C. J. (2020). Biological control of cabbage fly *Delia radicum* with entomopathogenic nematodes: Selecting the most effective nematode species and testing a novel application method. *Biological Control*, *144*, 104212 https://doi.org/10.1016/j.biocontrol.2020.104212.

Katoh, K., Misawa, K., Kuma, K. i., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O., & Grau, J. (2018). Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics*, *19*(1), 189. https://doi.org/10.1186/s12859-018-2203-5

Keilwagen, J., Wenk, M., Erickson, J. L., Schattat, M. H., Grau, J., & Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research*, *44*(9), e89. https://doi.org/10.1093/nar/gkw092

Kergunteuil, A., Dugravot, S., Danner, H., van Dam, N. M., & Cortesero, A. M. (2015). Characterizing volatiles and attractiveness of five brassicaceous plants with potential for a 'push-pull' strategy toward the cabbage root fly, *Delia radicum*. *Journal of Chemical Ecology*, *41*(4), 330–339. https://doi.org/10.1007/s10886-015-0575-9

Kissen, R., Rossiter, J. T., & Bones, A. M. (2009). The 'mustard oil bomb': not so easy to assemble?! Localization, expression and distribution of the components of the myrosinase enzyme system. *Phytochemistry Reviews*, *8*(1), 69–86. https://doi.org/10.1007/s11101-008-9109-1

Kolde, R. (2019). Package 'pheatmap'. *R package*, *1.0.12*.

Konecka, E., Olszanowski, Z., & Koczura, R. (2019). Wolbachia of phylogenetic supergroup E identified in oribatid mite Gustavia microcephala (Acari: Oribatida). *Molecular Phylogenetics and Evolution*, *135*, 230–235. https://doi.org/10.1016/j.ympev.2019.03.019

Konopka, T. (2020). *umap: Uniform Manifold Approximation and Projection. R package version*, *0.2.7.0*.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via

adaptive k-mer weighting and repeat separation. *Genome Research*, *Gr*, *215087*, 215116.

Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*, *19*(9), 1639–1645. https://doi.org/10.1101/gr.092759.109

Lachaise, T., Ourry, M., Lebreton, L., Guillerm-Erckelboudt, A.-Y., Linglin, J., Paty, C., Chaminade, V., Marnet, N., Aubert, J., Poinsot, D., & Cortesero, A.-M. & Mougel, C. (2017). Can soil microbial diversity influence plant metabolites and life history traits of a rhizophagous insect? A demonstration in oilseed rape. *Insect Science*, *24*(6), 1045–1056. https://doi.org/10.1111/1744-7917.12478

Lamy, F., Dugravot, S., Cortesero, A. M., Chaminade, V., Faloya, V., & Poinsot, D. (2018). One more step toward a push-pull strategy combining both a trap crop and plant volatile organic compounds against the cabbage root fly *Delia radicum*. *Environmental Science and Pollution Research*, *25*(30), 29868–29879. https://doi.org/10.1007/s11356-017-9483-6

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lo, N., Casiraghi, M., Salati, E., Bazzocchi, C., & Bandi, C. (2002). How many wolbachia supergroups exist? *Molecular Biology and Evolution*, *19*(3), 341–346. https://doi.org/10.1093/oxfordjournals.molbev.a004087

Lomsadze, A., Burns, P. D., & Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research*, *42*(15), e119. https://doi.org/10.1093/nar/gku557

Lopez, V., Cortesero, A. M., & Poinsot, D. (2018). Influence of the symbiont *Wolbachia* on life history traits of the cabbage root fly (*Delia radicum*). *Journal of Invertebrate Pathology*, *158*, 24–31. https://doi.org/10.1016/j.jip.2018.09.002

McDonald, R., & Sears, M. (1992). Assessment of larval feeding damage of the cabbage maggot (Diptera: Anthomyiidae) in relation to oviposition preference on canola. *Journal of Economic Entomology*, *85*(3), 957–962. https://doi.org/10.1093/jee/85.3.957

Michaud, S., Marin, R., Westwood, J. T., & Tanguay, R. M. (1997). Cell-specific expression and heat-shock induction of Hsps during spermatogenesis in Drosophila melanogaster. *Journal of Cell Science*, *110*(17), 1989–1997. https://doi.org/10.1242/jcs.110.17.1989

Moretti, S., Armougom, F., Wallace, I. M., Higgins, D. G., Jongeneel, C. V., & Notredame, C. (2007). The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods. *Nucleic Acids Research*, *35*(suppl_2), W645–W648. https://doi.org/10.1093/nar/gkm333

Neveu, N., Krespi, L., Kacem, N., & Nénon, J. P. (2000). Host-stage selection by Trybliographa rapae, a parasitoid of the cabbage root fly *Delia radicum*. *Entomologia Experimentalis Et Applicata*, *96*(3), 231–237. https://doi.org/10.1046/j.1570-7458.2000.00701.x

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment11Edited by. J. Thornton. *Journal of Molecular Biology*, *302*(1), 205–217. https://doi.org/10.1006/jmbi.2000.4042

Nottingham, S. (1988). Host-plant finding for oviposition by adult cabbage root fly, *Delia radicum*. *Journal of Insect Physiology*, *34*(3), 227–234. https://doi.org/10.1016/0022-1910(88)90053-4

Ourry, M., Crosland, A., Lopez, V., Derocles, S. A. P., Mougel, C., Cortesero, A. M., & Poinsot, D. (2021). Influential Insider: Wolbachia, an intracellular symbiont, manipulates bacterial diversity in its insect host. *Microorganisms*, *9*(6), https://doi.org/10.3390/microorganisms9061313

R Core Team (2020). *R: A Language and Environment for Statistical Computing [R]*. R Foundation for Statistical Computing.

Robinson, O., Dylus, D., & Dessimoz, C. (2016). Phylo.io: Interactive viewing and comparison of large phylogenetic trees on the web. *Molecular Biology and Evolution*, *33*(8), 2163–2166. https://doi.org/10.1093/molbev/msw080

Roessingh, P., & Städler, E. (1990). Foliar form, colour and surface characteristics influence oviposition behaviour in the cabbage root fly *Delia radicum*. *Entomologia Experimentalis Et Applicata*, *57*(1), 93–100. https://doi.org/10.1111/j.1570-7458.1990.tb01419.x

Roessingh, P., Städler, E., Fenwick, G., Lewis, J., Nielsen, J. K., Hurter, J., & Ramp, T. (1992). Oviposition and tarsal chemoreceptors of the cabbage root fly are stimulated by glucosinolates and host plant extracts. *Entomologia Experimentalis Et Applicata*, *65*(3), 267–282. https://doi.org/10.1111/j.1570-7458.1992.tb00680.x

Sato, K., & Touhara, K. (2008). Insect olfaction: receptors, signal transduction, and behavior. In S. Korsching, & W. Meyerhof (Eds.), *Chemosensory systems in mammals, fishes, and insects*, Vol. *47* (pp. 203–220). Springer.

Schramm, K., Vassão, D. G., Reichelt, M., Gershenzon, J., & Wittstock, U. (2012). Metabolism of glucosinolate-derived isothiocyanates to glutathione conjugates in generalist lepidopteran herbivores. *Insect Biochemistry and Molecular Biology*, *42*(3), 174–182. https://doi.org/10.1016/j.ibmb.2011.12.002

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: assessing genome assembly and annotation completeness. In M. Kollmar (Ed), *Gene prediction* (Vol. *1962*, pp. 227-245). Springer.

Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000Research*, *4*, 1521. https://doi.org/10.12688/f1000research.7563.1

Sontowski, R., & van Dam, N. M. (2020). Functional variation in dipteran gut bacterial communities in relation to their diet, life cycle stage and habitat. *Insects*, *11*(8), 543. https://doi.org/10.3390/insects11080543

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stanke, M., Schöffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, *7*(1), 62. https://doi.org/10.1186/1471-2105-7-62

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, *35*(11), 1026–1028. https://doi.org/10.1038/nbt.3988

Törönen, P., Medlar, A., & Holm, L. (2018). PANNZER2: A rapid functional annotation web server. *Nucleic Acids Research*, *46*(W1), W84–W88. https://doi.org/10.1093/nar/gky350

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515. https://doi.org/10.1038/nbt.1621

Tsunoda, T., Krosse, S., & van Dam, N. M. (2017). Root and shoot glucosinolate allocation patterns follow optimal defence allocation theory. *Journal of Ecology*, *105*(5), 1256–1266. https://doi.org/10.1111/1365-2745.12793

van Dam, N. M., Tytgat, T. O., & Kirkegaard, J. A. (2009). Root and shoot glucosinolates: A comparison of their diversity, function and interactions in natural and managed ecosystems. *Phytochemistry Reviews*, *8*(1), 171–186. https://doi.org/10.1007/s11101-008-9101-9

van Herk, W. G., Vernon, R. S., Waterer, D. R., Tolman, J. H., Lafontaine, P. J., & Prasad, R. P. (2016). Field evaluation of insecticides for control of cabbage maggot (Diptera: Anthomyiidae) in Rutabaga in Canada. *Journal of Economic Entomology*, *110*(1), 177–185. https://doi.org/10.1093/jee/tow238

Wallace, I. M., O'Sullivan, O., Higgins, D. G., & Notredame, C. (2006). M-Coffee: Combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Research*, *34*(6), 1692–1699. https://doi.org/10.1093/nar/gkl091

Wang, S., Voorrips, R. E., Steenhuis-Broers, G., Vosman, B., & van Loon, J. J. (2016). Antibiosis resistance against larval cabbage root fly, *Delia radicum*, in wild Brassica-species. *Euphytica*, *211*(2), 139–155. https://doi.org/10.1007/s10681-016-1724-0

Werren, J. H., Baldo, L., & Clark, M. E. (2008). Wolbachia: Master manipulators of invertebrate biology. *Nature Reviews Microbiology*, *6*(10), 741–751. https://doi.org/10.1038/nrmicro1969

Werren, J. H., & Windsor, D. M. (2000). Wolbachia infection frequencies in insects: Evidence of a global equilibrium? *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *267*(1450), 1277–1285. https://doi.org/10.1098/rspb.2000.1139

Werren, J. H., Zhang, W., & Guo, L. R. (1995). Evolution and phylogeny of Wolbachia: Reproductive parasites of arthropods. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *261*(1360), 55–63.

West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C., & Banfield, J. F. (2018). Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Research*, *28*(4), 569–580. https://doi.org/10.1101/gr.228429.117

Wiegmann, B. M., Trautwein, M. D., Winkler, I. S., Barr, N. B., Kim, J.-W., Lambkin, C., Bertone, M. A., Cassel, B. K., Bayless, K. M., Heimberg, A. M., Wheeler, B. M., Peterson, K. J., Pape, T., Sinclair, B. J., Skevington, J. H., Blagoderov, V., Caravas, J., Kutty, S. N., Schmidt-Ott, U., … Yeates, D. K. (2011). Episodic radiations in the fly tree of life. *Proceedings of the National Academy of Sciences*, *108*(14), 5690–5695. https://doi.org/10.1073/pnas.1012675108

Wittstock, U., & Gershenzon, J. (2002). Constitutive plant toxins and their role in defense against herbivores and pathogens. *Current Opinion in Plant Biology*, *5*(4), 300–307. https://doi.org/10.1016/S1369-5266(02)00264-9

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Sontowski, R., Poeschl, Y., Okamura, Y., Vogel, H., Guyomar, C., Cortesero, A.-M., & van Dam, N. M. (2022). A high-quality functional genome assembly of *Delia radicum* L. (Diptera: Anthomyiidae) annotated from egg to adult. *Molecular Ecology Resources*, 22, 1954–1971. https://doi.org/10.1111/1755-0998.13594