MOLECULAR ECOLOGY
RESOURCES

Supplemental Information for:

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

Rebekka Sontowski, Yvonne Poeschl, Yu Okamura, Heiko Vogel, Cervin Guyomar, Anne-Marie Cortesero, Nicole M. van Dam
Corresponding author: nicole.vandam@idiv.de
Please, contact the corresponding author for supplementary figures in high resolution.

Table of Contents:

MOLECULAR ECOLOGY
RESOURCES

1. Figures

DNA extraction from *D. radicum* for PacBio sequencing

Fly tissue was added to a lysis buffer containing 400 nM NaCl, 20 mM Tris HCl (pH 8.0), 30 mM EDTA (pH 8.0) and carefully mixed. Afterward, 125 µl SDS (10%) and 15 µl Proteinase K (10 mg/ml) were added to the mixture and incubated overnight at 55°C. The next day, samples were centrifuged for 30 min at 4000 x g at room temperature (RT) and the supernatant was transferred to a new tube. We added 5 µl RNAse A (2 mg/ml) to the supernatant, mixed it carefully and incubated the samples for 1 h at 37°C. After this step, the lysed tissue was washed twice with 250 µl Phenol:Chloroform:Isoamylalcohol (25:24:1, equilibrated with 10 mM Tris, pH 8.0 and 1 mM EDTA) and carefully mixed on an Intelligent Mixer for 10 min at RT. In the next step, the samples were centrifuged for 10 min at 4,000 x g at RT and the aqueous phase was transferred to a new tube. The aqueous solution was washed with 250 µl Chloroform, carefully mixed and centrifuged at 4,000 x g, at RT for 10 min. The last washing step was repeated twice. Afterward, the gDNA was precipitated by adding 450 µl precooled 96% ethanol, carefully mixed and centrifuged at 4,000 x g at 6°C for 20 min. The supernatant was discarded and the pellet was washed twice with 400 µl 70% cold ethanol. Samples were shortly spun, dried for 10 min at 37°C and dissolved in 50 µl 1x TE-buffer.

Figure S1 DNA extraction protocol.

Protocol used to extract total DNA from *Delia radicum* adults, which was further used for the PAC-Bio library preparation and sequencing.

MOLECULAR ECOLOGY
RESOURCES

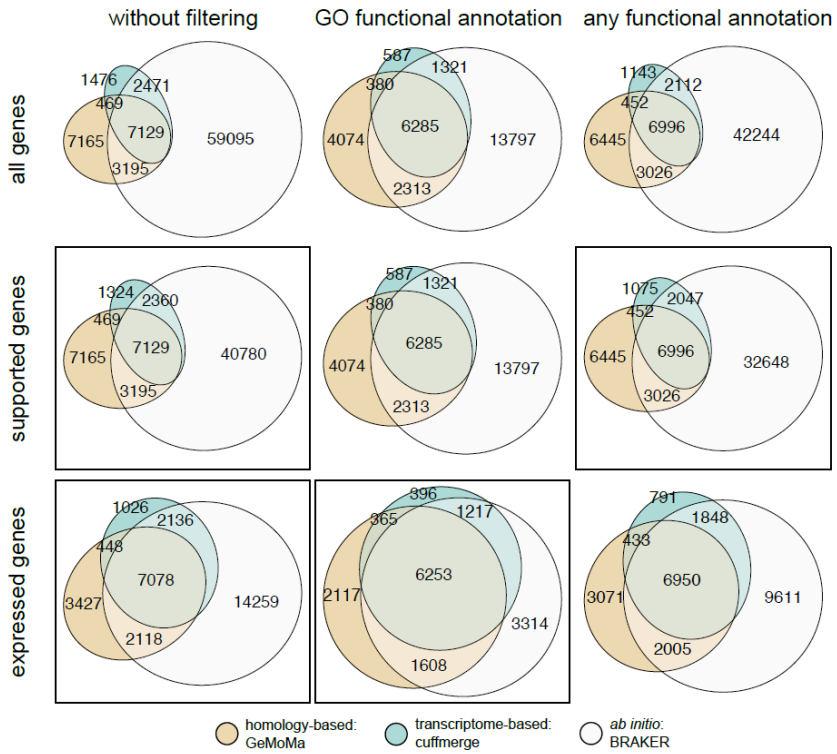


Figure S2 Venn diagrams containing the numbers of genes in the *Delia radicum* genome predicted by homology-based, transcriptome-based or *ab initio* approaches, or a combination thereof.

The numbers of genes in the diagrams are based on: (top) all predicted genes, (middle) all predicted genes that are supported by external evidence, and (bottom) expressed genes (Transcript Per Million (TPM) value ≥ 1); (left) total number of genes, number of genes with (middle) a functional annotation based on GO annotation and (right) any functional annotation, which includes GO annotation and/or protein family or domain annotation. Boxed Venn diagrams are presented in Figure 4.

MOLECULAR ECOLOGY
RESOURCES

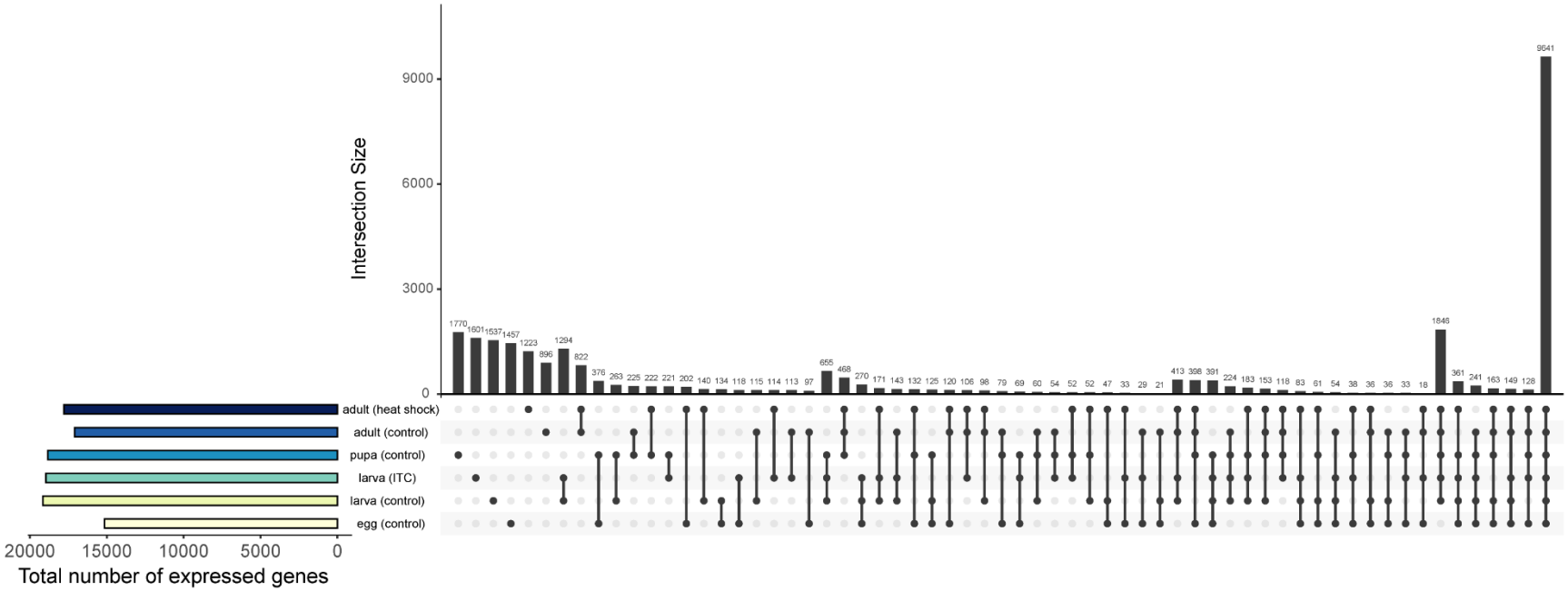
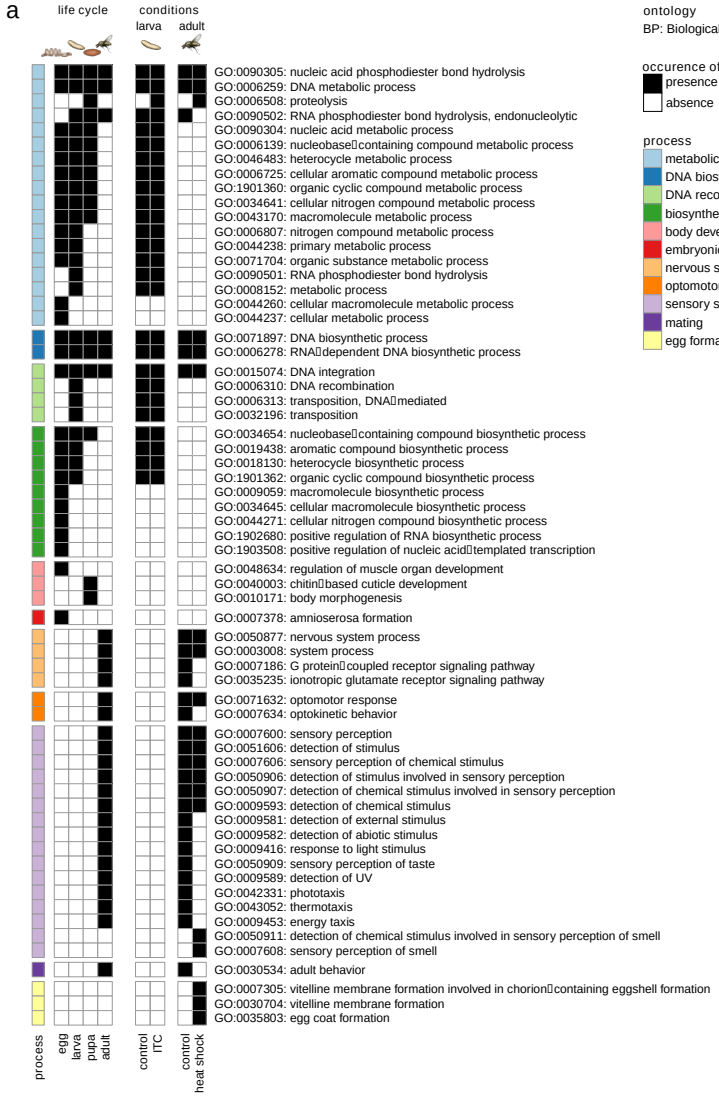


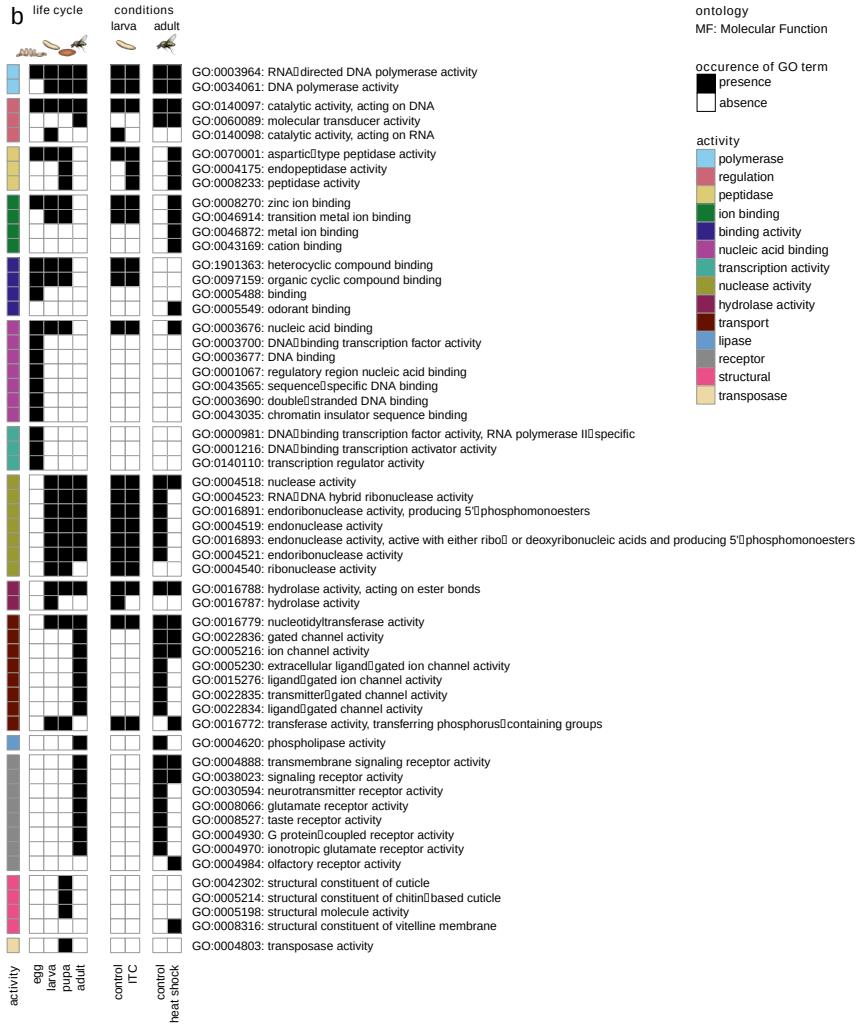
Figure S3 UpSet plot showing the number of expressed genes (Transcript Per Million (TPM) value ≥ 1) in all intersection sets (vertical bar plot).

The total number of expressed genes within each life stage or additional stress condition is given in the horizontal bar plot on the left.

MOLECULAR ECOLOGY
RESOURCES



MOLECULAR ECOLOGY
RESOURCES



MOLECULAR ECOLOGY
RESOURCES

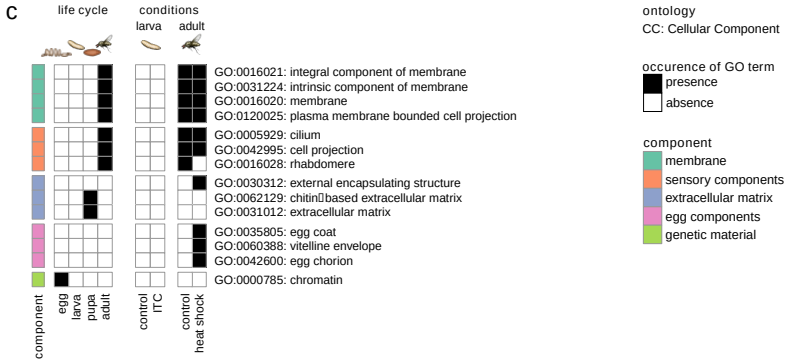


Figure S4 Gene ontology analyses of biological process (BP, a), molecular function (MF, b) and cellular components (CC, c) ontologies, based on expressed genes *(*Transcripts Per Million (TPM) value ≥ 1).

Results are shown in three respective heatmaps, where rows are labeled by GO terms and columns with life stages and/or conditions. Each cell in the heatmap shows the presence or absence of a GO term annotated for expressed genes in a life stage and/or condition. Only GO terms that were significantly overrepresented in a GO-enrichment analysis (Fisher's exact test, $P<0.05$ after correction with Benjamini- Yekutieli) are considered. Go terms are sorted into generic categories, where categories are indicated by the colors in the leftmost column. In all heatmaps the block with four columns to the left shows GO terms per life stage (egg, larva, pupa and adult) under control conditions, whereas the four columns to the right show the GO annotation of expressed genes of larvae and adults under control and stressed conditions. For easier comparison of control and stress conditions, columns showing results for larvae and adult under control conditions are duplicated.
ITC = larvae fed with 2-phenylethyl isothiocyanate.

MOLECULAR ECOLOGY
RESOURCES

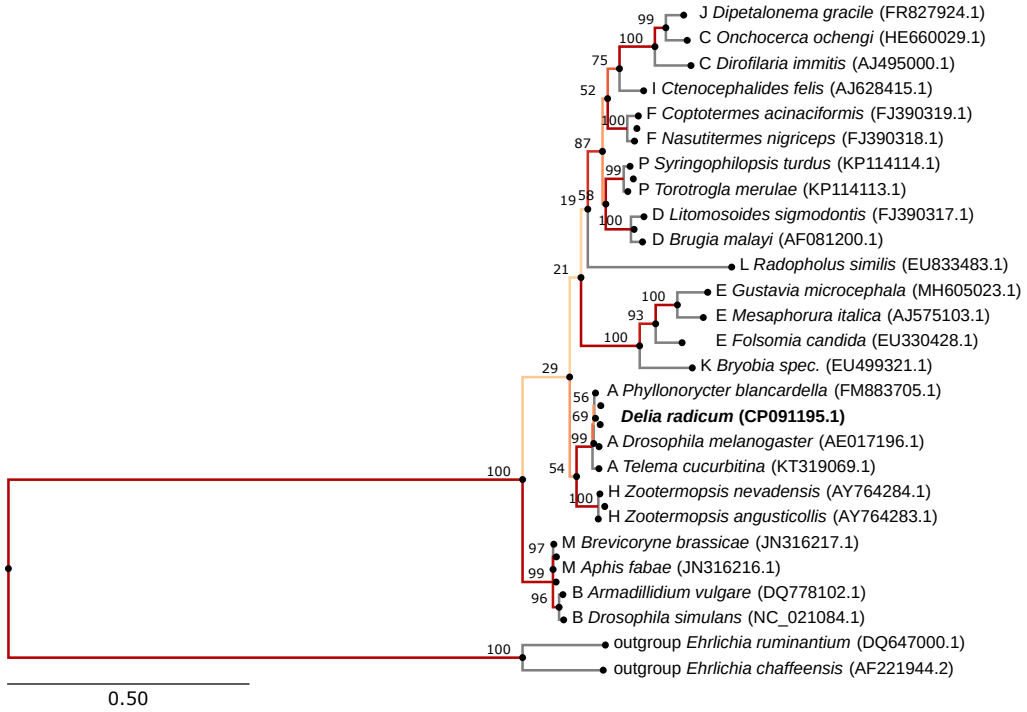


Figure S5 Maximum likelihood reconstruction of *Wolbachia* supergroup phylogeny.

Phylogenetic analyses of 27 *Wolbachia* species are based on the T-Coffee alignment of the sequences of their corresponding *ftsZ* locus with RAxML. Capital letters denote the *Wolbachia* supergroup affiliation, *Wolbachia* strains are labeled according to their host species and are referenced by their respective accession numbers at the National Center for Biotechnology (https://www.ncbi.nlm.nih.gov)([dataset] Assembly, 2012).

MOLECULAR ECOLOGY
RESOURCES

2. Tables

Tables **Table S1**, **Table S4**, **Table S7**, **Table S8**, **Table S11**, and **Table S12** are too large to be included here. They are provided as separate files.

Table S1 64 selected Dipteran species.

All 64 species are fully sequenced and annotated, and information can be obtained from National Center for Biotechnology (https://www.ncbi.nlm.nih.gov)([dataset] Assembly, 2012).

As external file.

Diptera in %:

Assembly	Complete	Single (S)	Duplicated (D)	Fragmented (F)	Missing (M)	Total
raw assembly	95.7	57.6	38.1	2.2	2.1	3285
polished assembly (1st round)	97.1	53.4	43.7	1.3	1.6	3285
polished assembly (2nd round)	97.3	53.5	43.8	1.2	1.5	3285
purged assembly	93.5	87.4	6.1	3.1	3.4	3285
chromosome-scale assembly	93.4	92.2	1.2	2.8	3.8	3285

Diptera absolute counts:

Assembly	Complete	Single (S)	Duplicated (D)	Fragmented (F)	Missing (M)	Total
raw assembly	3141	1891	1250	72	72	3285
polished assembly (1st round)	3189	1752	1436	44	52	3285
polished assembly (2nd round)	3196	1758	1438	40	49	3285
purged assembly	3071	2872	199	102	112	3285
chromosome-scale assembly	3070	3030	40	91	124	3285

Table S2 Summary of BUSCO statistics for genome assemblies of *Delia radicum* on the Diptera gene set.

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. Detailed statistics on the assemblies are provided in Table 2. BUSCO statistics for the diptera_odb10.2019-11-20 gene set are given in percentages and absolute numbers for the intermediate and the final assemblies. BUSCO was run in genome mode for these analyses.

MOLECULAR ECOLOGY
RESOURCES

Endopterygota in %:

Assembly	Complete	Single (S)	Duplicated (D)	Fragmented (F)	Missing (M)	Total
raw assembly	98.1	54.1	44	0.8	1.1	2124
polished assembly (1st round)	98.6	50.4	48.2	0.6	0.8	2124
polished assembly (2nd round)	98.6	49.6	49	0.6	0.8	2124
purged assembly	96	89.1	6.9	2.3	1.7	2124
chromosome-scale assembly	95.8	94.1	1.7	2.5	1.7	2124

Endopterygota absolute counts:

Assembly	Complete	Single (S)	Duplicated (D)	Fragmented (F)	Missing (M)	Total
raw assembly	2084	1150	934	18	22	2124
polished assembly (1st round)	2095	1071	1024	12	17	2124
polished assembly (2nd round)	2094	1054	1040	12	18	2124
purged assembly	2040	1893	147	48	36	2124
chromosome-scale assembly	2036	1999	37	54	34	2124

Table S3 Summary of BUSCO statistics for genome assemblies of *D. radicum* on the Endopterygota gene set.

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. Detailed statistics on the assemblies are provided in Table 2. BUSCO statistics for the endopterygota_odb10.2019-11-20 gene set are given in percentages and absolute numbers for the intermediate and the final assemblies. BUSCO was run in genome mode for these analyses.

MOLECULAR ECOLOGY
RESOURCES

Pseudo chromosomes:

Superscaffold	Number of bases	Gaps	Genes (raw)	Genes (supported)
HiC_scaffold_1	328,483,116	1,316	19,908	15,268
HiC_scaffold_2	247,132,447	1,002	14,571	11,309
HiC_scaffold_3	242,504,274	1,228	13,905	10,868
HiC_scaffold_4	241,915,331	1,538	15,073	11,417
HiC_scaffold_5	208,954,149	1,037	12,265	9,603
HiC_scaffold_6	12,881,041	67	1,653	1,258
sum	1,281,870,358	6,188	77,375	59,723
% of total	96.79		95.52	95.68

Second sheet “all scaffolds” as external file.

Table S4 Summary of statistics of the individual six chromosome-scale scaffolds.

Number of bases, gaps and genes are listed for each of the six pseudo-chromosomes. Additionally, number of bases, number of predicted genes and number of genes with support by external evidence are provided for all scaffolds.

Diptera in %:

Species	Complete	Single (S)	Duplicated (D)	Fragmented (F)	Missing (M)	Total
Anopheles gambiae	98.9	98	0.9	0.2	0.9	3285
Drosophila melanogaster	99.4	99	0.4	0.2	0.4	3285
Musca domestica	97.1	94.9	2.2	1.2	1.7	3285
Delia radicum	93.4	92.2	1.2	2.8	3.8	3285
Lucilia cuprina	98.1	97.1	1	1	0.9	3285

Diptera absolute counts:

Species	Complete	Single (S)	Duplicated (D)	Fragmented (F)	Missing (M)	Total
Anopheles gambiae	3250	3219	31	8	27	3285
Drosophila melanogaster	3265	3251	14	5	15	3285
Musca domestica	3188	3116	72	39	58	3285
Delia radicum	3070	3030	40	91	124	3285
Lucilia cuprina	3223	3189	34	34	28	3285

Table S5 Summary of BUSCO statistics for 5 Dipteran species.

BUSCO statistics for the diptera_odb10.2019-11-20 gene set are given in percentages and absolute numbers for five Dipteran species. BUSCO was run in genome mode for these analyses. Detailed information on the species is provided in Table 1.

MOLECULAR ECOLOGY
RESOURCES

Endopterygota in %:

Species	Complete	Single (S)	Duplicated (D)	Fragmented (F)	Missing (M)	Total
Spodoptera litura	97.1	96.2	0.9	1.1	1.8	2124
Manduca sexta	95.8	93.8	2	2.4	1.8	2124
Pieris rapae	97.6	97.2	0.4	0.8	1.6	2124
Plutella xylostella	88.3	75.7	12.6	3.4	8.3	2124
Tribolium castaneum	98.8	98.5	0.3	0.6	0.6	2124
Anopheles gambiae	98.6	97	1.6	0.5	0.9	2124
Drosophila melanogaster	99.6	98.7	0.9	0	0.4	2124
Musca domestica	97.8	96	1.8	1.1	1.1	2124
Delia radicum	95.8	94.1	1.7	2.5	1.7	2124
Lucilia cuprina	99.1	98.3	0.8	0.4	0.5	2124

Endopterygota absolute counts:

Species	Complete	Single (S)	Duplicated (D)	Fragmented (F)	Missing (M)	Total
Spodoptera litura	2064	2044	20	23	37	2124
Manduca sexta	2035	1993	42	51	38	2124
Pieris rapae	2073	2064	9	17	34	2124
Plutella xylostella	1874	1607	267	73	177	2124
Tribolium castaneum	2100	2093	7	13	11	2124
Anopheles gambiae	2094	2060	34	10	20	2124
Drosophila melanogaster	2115	2096	19	1	8	2124
Musca domestica	2078	2040	38	24	22	2124
Delia radicum	2036	1999	37	54	34	2124
Lucilia cuprina	2105	2088	17	9	10	2124
intersect		1217				

Table S6 Summary of BUSCO statistics for *D. radicum* and 9 selected insect species on the Endopterygota gene set.

MOLECULAR ECOLOGY
RESOURCES

BUSCO statistics for the endopterygota_odb10.2019-11-20 gene set are given in percentages and absolute numbers for the 10 species. BUSCO was run in genome mode for these analyses. Detailed information on the species is provided in Table 1.

Table S7 List of genes of the endopterygota_odb10.2019-11-20 gene set shared by *D. radicum* and the 9 selected insect species.

As external file.

Table S8 Summary of synteny analysis of *D. radicum* and *D. melanogaster*

The set of chromosomes of *D. radicum* was compared with those of *D. melanogaster* based on gene annotations for *D. melanogaster*. Gene models annotated for *D. melanogaster* were used in GeMoMa to predict homologous genes in *D. radicum* genomic sequences. Resulting genes are listed in the table. Additionally, the list of genes used for generating the circus plot to visualize the synteny for k=2 (at least 2 genes must be consecutive) in Figure 3b is listed.

As external file.

MOLECULAR ECOLOGY
RESOURCES

Support by external evidence:

	total number of genes	GO annotation		match in Dipteran database		expression state		support by external evidence	
		yes	no	yes	no	TPM ≥ 1	TPM < 1	yes	no
braker	59095	13797	45298	30532	28563	14259	44836	40780	18315
braker:cuffmerge	2471	1321	1150	1877	594	2136	335	2360	111
cuffmerge	1476	587	889	886	590	1026	450	1324	152
to exclude									18578

Table S9 Summary statistics of genes supported by external evidence.

Genes precited by Cuffmerge or BRAKER (Figure 4, Figure S2) were counted to have GO annotation, putative homology to annotated Dipteran proteins or show expression. This external evidence was used to determine supported and unsupported genes. A gene is unsupported if it has no GO annotation, has no putative homology or is not expressed. 18578 unsupported genes were excluded from the set of predicted *D. radicum* genes. (TPM=Transcripts Per Million)

MOLECULAR ECOLOGY
RESOURCES

BUSCO sets in %:

BUSCO set	Subset of genes	Complete	Single (S)	Duplicated (D)	Fragmented (F)	Missing (M)	Total
Insecta	all	97.5	34.2	63.3	1.9	0.6	1367
Endopterygota	all	95.4	31.4	64	3.3	1.3	2124
Diptera	all	93.6	27.7	65.9	3.3	3.1	3285
Insecta	by all 3 approaches	75.8	24.2	51.6	0.8	23.4	1367
Endopterygota	by all 3 approaches	74.6	22	52.6	0.5	24.9	2124
Diptera	by all 3 approaches	71.7	18.4	53.3	0.5	27.8	3285
Insecta	by at least 2 approaches	90.7	30.1	60.6	1.2	8.1	1367
Endopterygota	by at least 2 approaches	89.5	28.2	61.3	1.8	8.7	2124
Diptera	by at least 2 approaches	87.5	24.6	62.9	1.4	11.1	3285

BUSCO sets absolute counts:

BUSCO set	Subset of genes	Complete	Single (S)	Duplicated (D)	Fragmented (F)	Missing (M)	Total
Insecta	all	1332	467	865	26	9	1367
Endopterygota	all	2026	666	1360	71	27	2124
Diptera	all	3075	911	2164	109	101	3285
Insecta	all 3 approaches	1037	331	706	11	319	1367
Endopterygota	all 3 approaches	1584	467	1117	11	529	2124
Diptera	all 3 approaches	2358	606	1752	16	911	3285
Insecta	at least 2 approaches	1239	411	828	17	111	1367
Endopterygota	at least 2 approaches	1901	599	1302	38	185	2124
Diptera	at least 2 approaches	2872	807	2065	46	367	3285

MOLECULAR ECOLOGY
RESOURCES

Table S10 Summary of BUSCO statistics on final genome annotation.

BUSCO statistics for the insecta_odb10.2019-11-20, endopterygota_odb10.2019-11-20 and diptera_odb10.2019-11-20 gene set are given in percentages and absolute numbers for gene predictions of the final genome annotation of the *Delia radicum* genome. Different subsets of genes annotated by GeMoMa, Cufflinks and BRAKER were analyzed. BUSCO was run in protein mode for these analyses. BUSCO results are the same for the final gene set containing all genes and the filtered set containing genes that are supported by external evidence.

Table S11 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 a 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 81,000 putative genes were 62,422 are supported by external evidence. Additionally, the table contains the number of genes that were annotated with any functional annotation, which includes GO annotation and/or protein family or domain annotation. Related gene sets are shown in Figure 4 and Figure S2.

As external file.

MOLECULAR ECOLOGY
RESOURCES

Table S12 Summary of results of GO-enrichment analyses.

Tables contain enrichment results of GO annotations of genes (*Transcripts Per Million (TPM) value* ≥ *1*) expressed in different life stages and/or conditions for each of the three ontologies (biological process, molecular function and cellular component). Enrichment analyses were done with topGO by applying Fisher's exact test and p-value correction with Benjamini- Yekutieli.

As external files.