Supplementary Material for Florida Entomologist 100: 42–51

Tian, Chuan-Bei, Dong Wei, Lin-Fan Xiao, Wei Dou, Huai Liu, and Jin-Jun Wang—Comparative transcriptome analysis of three *Bactrocera dorsalis* (Diptera: Tephritidae) organs to identify functional genes in the male accessory glands and ejaculatory duct.

Abstract

The insect male accessory glands/ejaculatory duct (MAG/ED) are important tissues of the male reproductive system. The MAG/ED's functions in reproduction have been well studied in *Drosophila* (Diptera: Drosophilidae) but remain largely unknown in the important agricultural pest *Bactrocera dorsalis* (Hendel) (Diptera: Tephritidae). In the present study, we re-assembled the transcriptome datasets of *B. dorsalis*'s fat body, testis, and MAG/ED and compared these tissue-specific transcriptomes. Clean reads from these transcriptome data sets were de novo re-assembled and clustered into 31,782 unigenes (average 922 bp). In total, 21,306 unigenes were functionally annotated by Blasting against online databases. Comparative transcriptomic analysis identified numerous genes that were identified with the expressed tissue-bias patterns. Some MAG/ED-specific genes potentially involved in spermatozoa motility and capacitation (e.g., perlucin, glucose dehydrogenase, lipase), mating regulation (pheromone-binding protein–related protein), and immunity (lectin) were identified in *B. dorsalis*. The expressions of some of these genes were further validated by realtime quantitative polymerase chain reaction at transcriptional level. All of these identifications will help us to explore the physiological regulation of mating and reproduction in *B. dorsalis* in the future.

Key Words: oriental fruit fly; comparative transcriptome; accessory gland protein; seminal fluid protein; reproduction

Resumen

Las glándulas accesorias masculinas/conducto eyaculador (GAM/CE) de los insectos son tejidos importantes del sistema reproductivo masculino. Las funciones de GAM/CE en la reproducción han sido bien estudiadas en *Drosophila* (Diptera: Drosophilidae), pero siguen siendo poco conocidas en la importante plaga agrícola *Bactrocera dorsalis* (Hendel) (Diptera: Tephritidae). En el presente estudio, re-ensamblamos los datos de transcriptoma de la grasa del cuerpo, los testículos y los GAM/CE de *B. dorsalis* y comparamos estos transcriptomas específicos de tejidos. Las lecturas claras de estos conjuntos de datos de transcriptoma se volvieron a ensamblar de nuevo y se agruparon en 31.782 unigenes (promedio de 922 bp). En total, 21.306 unigenes fueron anotados funcionalmente por Blasting contra bases de datos en línea. El análisis transcriptómico comparativo identificó numerosos genes que se identificaron con los patrones de sesgo de tejido expresado. En *B. dorsalis* se identificaron algunos genes especıficos de GAM/CE potencialmente implicados en motilidad y capacitación de espermatozoides (por ejemplo, perlucina, deshidrogenasa de glucosa, lipasa), regulación de apareamiento (proteına relacionada con proteına de unión a feromonas) e inmunidad (lectina). Las expresiones de algunos de estos genes fueron aún más validados por la reacción en cadena de la polimerasa cuantitativa en tiempo real a nivel transcripcional. Todas estas identificaciones nos ayudarán a explorar la regulación fisiológica del apareamiento y la reproducción en *B. dorsalis* en el futuro.

Palabras Clave: mosca de la fruta oriental; transcriptoma comparativo; proteína accesoria de las glándulas; proteína fluida seminal; reproducción

47 **Table S1.** Oligonucleotide primers used for quantitative real-time PCR.

No.	Pathway	All genes with pathway annotation (13782)	Pathway ID	Level 1	Level 2
1	Metabolic pathways	1919 (13.92%)	ko01100	Metabolism	Global map
2	Pathways in cancer	566 (4.11%)	ko05200	Human Diseases	Cancers: Overview
3	Regulation of actin cytoskeleton	486 (3.53%)	ko04810	Cellular Processes	Cell motility
4	Focal adhesion	478 (3.47%)	ko04510	Cellular Processes	Cell communication
5	Purine metabolism	428 (3.11%)	ko00230	Metabolism	Nucleotide metabolism
6	RNA transport	422 (3.06%)	ko03013	Genetic Information Processing	Translation
7	HTLV-I infection	403 (2.92%)	ko05166	Human Diseases	Infectious diseases: Viral
8	Huntington's disease	399 (2.9%)	ko05016	Human Diseases	Neurodegenerative diseases
9	Endocytosis	354 (2.57%)	ko04144	Cellular Processes	Transport and catabolism
10	Epstein-Barr virus infection	352 (2.55%)	ko05169	Human Diseases	Infectious diseases: Viral
11	Lysine degradation	350 (2.54%)	ko00310	Metabolism	Amino acid metabolism
12	Spliceosome	344 (2.5%)	ko03040	Genetic Information Processing	Transcription
13	Vascular smooth muscle contraction	341 (2.47%)	ko04270	Organismal Systems	Circulatory system
14	MAPK signaling pathway	328 (2.38%)	ko04010	Environmental Information Processing	Signal transduction
15	Influenza A	314 (2.28%)	ko05164	Human Diseases	Infectious diseases: Viral
16	Alzheimer's disease	313 (2.27%)	ko05010	Human Diseases	Neurodegenerative diseases

No.	Pathway	All genes with pathway annotation (13782)	Pathway ID	Level 1	Level 2
17	Amoebiasis	294 (2.13%)	ko05146	Human Diseases	Infectious diseases: Parasitic
18	Cell cycle	294 (2.13%)	ko04110	Cellular Processes	Cell growth and death
19	Ubiquitin mediated proteolysis	292 (2.12%)	ko04120	Genetic Information Processing	Folding, sorting and degradation
20	Calcium signaling pathway	291 (2.11%)	ko04020	Environmental Information Processing	Signal transduction
21	Herpes simplex infection	288 (2.09%)	ko05168	Human Diseases	Infectious diseases: Viral
22	Phagosome	286 (2.08%)	ko04145	Cellular Processes	Transport and catabolism
23	Vibrio cholerae infection	283 (2.05%)	ko05110	Human Diseases	Infectious diseases: Bacterial
24	Wnt signaling pathway	283 (2.05%)	ko04310	Environmental Information Processing	Signal transduction
25	Tight junction	281 (2.04%)	ko04530	Cellular Processes	Cell communication
26	Protein processing in endoplasmic reticulum	280 (2.03%)	ko04141	Genetic Information Processing	Folding, sorting and degradation
27	Dilated cardiomyopathy	279 (2.02%)	ko05414	Human Diseases	Cardiovascular diseases
28	Hypertrophic cardiomyopathy (HCM)	269 (1.95%)	ko05410	Human Diseases	Cardiovascular diseases
29	RNA degradation	268 (1.94%)	ko03018	Genetic Information Processing	Folding, sorting and degradation
30	Transcriptional misregulation in cancer	266 (1.93%)	ko05202	Human Diseases	Cancers: Overview

Table S2. Top 30 pathways with highest numbers of unigenes as determined by Kyoto encyclopedia of gene and genomes (KEGG) analysis in the re-assembled transcriptome from 3 tissues of *Bactrocera dorsalis*.

geneID	gene Length	Testis _FPKM	Fatbody _FPKM	MAG/ED _FPKM	Nr-annotation	Swissprot-annotation
Unigene4743_All	1133	66.7842	18.8583	42524.8291	CG5867 [Drosophila melanogaster]	Circadian clock-controlled protein OS=Drosophila melanogaster
Unigene12723_All	699	4.4184	1.8563	3171.7019	Aggrecan core protein [Crassostrea gigas]	C-type lectin domain family 4 member K OS=Homo sapiens
Unigene6415_All	623	5.8369	0.6943	3103.4225	predicted protein [Culex quinquefasciatus]	Protein TsetseEP OS=Glossina palpalis palpalis
Unigene5480_All	721	7.8071	1.9796	2932.813	AAEL011612-PB [Aedes aegypti]	Perlucin-like protein OS=Mytilus galloprovincialis
Unigene12065_All	1387	24.7451	2.9936	2477.811	GK25805 [Drosophila willistoni]	Vitellogenin-1 OS=Ceratitis capitat
Unigene8523_All	1989	0.6261	1.6309	2260.8618	GA15794 [Drosophila pseudoobscura pseudoobscura]	Venom serine protease OS=Polistes dominula
Unigene4782_All	1052	2.3202	6.1671	1822.8825	CG3604 [Drosophila melanogaster]	Kunitz-type proteinase inhibitor kalicludin-3 OS=Anemonia sulcata
Unigene12506_All	752	1.5898	2.0706	1635.2022	GF22327 [Drosophila ananassae]	Vitellogenin-1 OS=Ceratitis capitata
Unigene18301_All	2596	0.2111	2.2159	1501.4639	gag-pol polyprotein precursor [Drosophila melanogaster]	--
Unigene5896_All	1343	3.1157	1.256	1030.2893	Perlucin [Crassostrea gigas]	--
Unigene7540_All	695	1.7918	0.3734	677.8261	incilarin A [Haliotis discus discus]	Collectin-12 OS=Bos taurus
Unigene22749_All	687	1.5227	1.3221	593.8936	GI24672 [Drosophila mojavensis]	Protease inhibitor 2 OS=Cenchritis muricatus
Unigene20831_All	648	0	0.1335	511.2208	GJ14891 [Drosophila virilis]	--
Unigene5218_All	1245	2.2006	0.2779	476.9481	GJ17329 [Drosophila virilis]	Lipase 1 OS=Drosophila melanogaster
Unigene20875_All	1152	0.0865	0.413	442.2946	GH17321 [Drosophila grimshawi]	Plasma kallikrein OS=Homo sapiens

geneID	gene Length	Testis _FPKM	Fatbody _FPKM	MAG/ED _FPKM	Nr-annotation	Swissprot-annotation
Unigene20808_All	699	0.0713	0.3713	408.3381	CRLBP homologous protein [Phormia regina]	Pheromone-binding protein-related protein 5 OS=Drosophila melanogaster
Unigene22824_All	362	0.8256	0.1195	405.3389	Aggrecan core protein [Crassostrea gigas]	Lymphocyte antigen 75 OS=Mesocricetus auratus
Unigene20845_All	1163	2.6556	0.4463	401.4347	GF21595 [Drosophila ananassae]	Venom allergen 3 OS=Solenopsis richteri
Unigene20775_All	1230	0.567	0.2813	390.4096	GK15364 [Drosophila willistoni]	Angiopoietin-related protein 1 OS=Homo sapiens
Unigene20920_All	625	0.2391	0.4844	381.038	GA24335 [Drosophila pseudoobscura pseudoobscura]	--
Unigene22777_All	1177	0	0.0367	363.4332	GJ14891 [Drosophila virilis]	Vitellogenin-1 OS=Drosophila melanogaster
Unigene20895_All	779	3.3252	0.0555	357.0085	Transposable element Tc3 transposase [Lepeophtheirus salmonis]	--
Unigene20911_All	964	0.2584	0.3589	335.2118	GA23387 [Drosophila pseudoobscura pseudoobscura]	Fibroleukin OS=Bos taurus
Unigene8806_All	1855	1.4769	0.4896	256.9336	GJ20598 [Drosophila virilis]	Lysyl oxidase homolog 2B OS=Danio rerio
Unigene20804_All	615	0.405	0.0703	234.7203	GM26528 [Drosophila sechellia]	--
Unigene22798_All	897	0.0555	0.0482	128.4426	serine protease [Eupolyphaga sinensis]	Trypsin alpha-4 OS=Lucilia cuprina
Unigene22901_All	2131	0.1636	0.1827	92.4334	GI10599 [Drosophila mojavensis]	--
Unigene20784_All	2061	0.1934	0.3568	75.6391	GK17743 [Drosophila willistoni]	Glycine receptor subunit alpha-2 OS=Rattus norvegicus
Unigene20837_All	1966	0.152	0.682	69.428	GI24327 [Drosophila mojavensis]	Acyl-CoA desaturase 1 OS=Rattus norvegicus

geneID	gene Length	Testis _FPKM	Fatbody _FPKM	MAG/ED _FPKM	Nr-annotation	Swissprot-annotation
Unigene20897_All	2378	0.0419	0.2728	61.4777	GD16846 [Drosophila simulans]	Sodium-coupled monocarboxylate transporter 1 OS=Xenopus laevis
Unigene20848_All	1778	0.056	0.0487	57.5769	GL20599 [Drosophila persimilis]	Venom carboxylesterase-6 OS=Apis mellifera
Unigene22918_All	2417	0.1855	0.5368	55.1356	GM26533 [Drosophila sechellia]	Protein Skeletor, isoforms D/E OS=Drosophila melanogaster
Unigene20842_All	566	0.088	0.0764	30.3034	FLJ37770-like protein [Acromyrmex echinatior]	Putative uncharacterized protein FLJ37770 OS=Homo sapiens
Unigene22860_All	681	0	0.0635	22.7466	GJ13625 [Drosophila virilis]	Glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase 1 OS=Drosophila melanogaster
Unigene22872_All	2034	0.1469	0.0851	19.5361	GH17733 [Drosophila grimshawi]	Small G protein signaling modulator 1 OS=Mus musculus
Unigene20909_All	607	0	0	18.7144	GG11290 [Drosophila erecta]	--
Unigene21261_All	702	0.1419	0	16.0539	GH18792 [Drosophila grimshawi]	--
Unigene22873_All	1513	0.0658	0	14.2742	RE30781p [Drosophila melanogaster]	Small G protein signaling modulator 1 OS=Homo sapiens
Unigene22735_All	988	0	0.0438	11.4522	GI18492 [Drosophila mojavensis]	Polypeptide N-acetylgalactosaminyltransferase 5 OS=Caenorhabditis elegans
Unigene20841_All	415	0	0	10.9274	FLJ37770-like protein [Acromyrmex echinatior]	Putative uncharacterized protein FLJ37770 OS=Homo sapiens PE=5 SV=1
Unigene22186_All	956	0.0521	0	10.8962	GA26432 [Drosophila pseudoobscura pseudoobscura]	--
Unigene22193_All	826	0	0.0524	8.4799	GA29189 [Drosophila pseudoobscura pseudoobscura]	--
Unigene21357_All	656	0	0	7.2552	GE11021 [Drosophila yakuba]	--

geneID	gene Length	Testis _FPKM	Fatbody _FPKM	MAG/ED _FPKM	Nr-annotation	Swissprot-annotation
Unigene22349_All	679	0	0	6.0175	mariner transposase [Buenoa sp. HMR-1997]	--
Unigene23158_All	731	0.0681	0	4.9138	GH17326 [Drosophila grimshawi]	Lachesin OS=Drosophila melanogaster
Unigene22122_All	801	0	0	4.8207	GM18689 [Drosophila sechellia]	DNA-directed RNA polymerase II subunit RPB1 OS=Caenorhabditis elegans
Unigene23869_All	574	0	0	2.6596	GF16852 [Drosophila ananassae]	Octopamine receptor OS=Heliothis virescens
Unigene20751_All	1320	0	0.0328	1.1225	Camar1 transposase [Chymomyza amoena]	Mariner Mos1 transposase OS=Drosophila mauritiana
Unigene18921_All	1273	0	0	0.4938	GM10508 [Drosophila sechellia]	Muscle LIM protein Mlp84B OS=Drosophila melanogaster
Unigene13119_All	856	0.0582	0.0505	0.4721	FLJ37770-like protein [Acromyrmex echinatior]	Putative uncharacterized protein FLJ37770 OS=Homo sapiens PE=5 SV=1
Unigene1378_All	1550	0.0964	0.0558	0.4345	GJ13058 [Drosophila virilis]	Cysteine-rich with EGF-like domain protein 2-A OS=Xenopus laevis
Unigene1757_All	1791	0.0834	0.0724	0.1253	aldehyde dehydrogenase type III, isoform Q [Drosophila melanogaster]	Aldehyde dehydrogenase family 3 member B1 OS=Rattus norvegicus
Unigene20810_All	611	0	0	0.0735	ion transport peptide, isoform C [Drosophila melanogaster]	Ion transport peptide-like OS=Schistocerca gregaria

52

Table S3. Distinct unigenes information with specific protein prediction in NCBI nr or Swiss-Prot databases. Gene length, expression value in each sample, and annotation in NCBI nr or Swiss-Prot databases were provided in details.

Repeats	4	5	6	7	8	9	10	11	12	total
TG	-	-	154	49	44	16	1	2	2	268
TTG	-	177	55	15	3					250
AC	-	-	114	69	34	23	3	2	1	246
CAA	-	173	52	10	2					237
TGT	-	163	36	7	2					208
GT	-	-	98	41	17	11	6	6	1	180
AT	-	-	100	34	18	5	8	3	4	172
TA	-	-	92	45	15	5		5	1	163
ACA	-	110	40	7	3					160
CA	-	-	85	40	15	3	2	2		147
TGC	-	79	21	5	2					107
AAC	-	66	19	6						91
GTT	-	56	16	7	1					80
CAG	-	57	17	3						77
GCT	-	44	14	4	1					63
GCA	-	38	8	2	2					50
CCA	-	23	18	3						44
TGG	-	20	19	2						41
GAT	-	19	13		2					34
CAC	-	21	8	3						32
AAT	-	24	6	1						31
AGC	-	18	7	2	4					31
GGT	-	27	4							31
GAC	-	26	3		1					30
TCA	-	22	2	2	2					28
AAG	-	23	4							27
GTG	-	18	6	3						27
CAT	-	17	8	1						26
GA	-	-	13	5		4	1	2		25
CTG	-	11	9	4	1					25
TGA	-	18	6	1						25
TTA	-	19	6							25
CT	-	-	7	6	3	4	2	2		24
ACC	-	15	5	1						21
ATG	-	13	6	1	1					21
GCC	-	20								20
AG	-	-	5	8	2	2	2			19
TC	-	-	7	3	2	4	1		1	18
ATT	-	15	2	1						18

Repeats	4	5	6	7	8	9	10	11	12	total
CTT	-	14	4							18
GCG	-	14	2	1						17
GGC	-	14	2							16
TAA	-	10	4		2					16
TACA	-	14								14
CCG	-	11	2							13
TAT	-	9	1	2	1					13
GTC	-	11								11
TCG	-	7	4							11
ATAC	-	11								11
AGA	-	3	7							10
CGC	-	8	2							10
GTA	-	6	4							10
TAC	-	5	5							10
TAG	-	8	1	1						10
TATG	-	10								10
TGTA	-	9	1							10
AGT	-	5	2	2						9
TCT	-	8	1							9
CATA	-	7	1	1						9
CGG	-	8								8
CTA	-	7			1					8
GAA	-	3	1	3	1					8
ACAT	-	7	1							8
CGA	-	4	3							7
CTC	-	3		4						7
TTC	-	7								7
GTTTC	7									7
ATA	-	3	2	1						6
ATC	-	4	1		1					6
GTAT	-	5	1							6
CGT	-	2	3							5
AAAAAC	5									5
ACG	-	2	2							4
ACT	-	3			1					4
GAG	-	3	1							4
ATGT	-	4								4
CG	-	-	3							3
GGA	-	2	1							3
CCT	-	1	1							2
TCC	-	1	1							2
ATTTT	2									2

Repeats	4	5	6	7	8	9	10	11	12	total
CCTAA	2									2
GTTAT	1	1								2
TTGAT	2									2
AAAAAT	2									2
GC	-	-	1							1
AGG	-	1								1
AAAT	-	1								1
AACA	-	1								1
ACCG	-	1								1
CCGA	-	1								1
CCTA	-	1								1
CTGA	-		1							1
CTGT	-	1								1
GTAG	-	1								1
TAGA	-	1								1
TCAG	-	1								1
TCGT	-	1								1
TGAC	-	1								1
TGTC	-	1								1
TTAT	-	1								1
AAACA	1									1
AAATA	1									1
AACAA	1									1
AATTT	1									1
ACATA	1									1
ACATC	1									1
ACTCT	1									1
CAATT	1									1
CACAG	1									1
GAGTA	1									1
GTTTT		1								1
TAAAT	1									1
TAAGT	1									1
TTAAT		1								1
TTAGG	1									1
TTCAG		1								1
TTTAT	1									1
TTTGG	1									1
TTTTA	1									1
AAAACC	1									1
AAACAA	1									1
AAATAC	1									1

Repeats	4	5	6	7	8	9	10	11	12	total
AATACA	1									1
ACAAAA	1									1
ATGCCA	1									1
CACGCG	1									1
CTGTTG	1									1
GCATTT	1									1
GTTGGC	1									1
GTTTTT	1									1
TTAATG	1									1
TTCTGT	1									1
TTGCTG	1									1

56

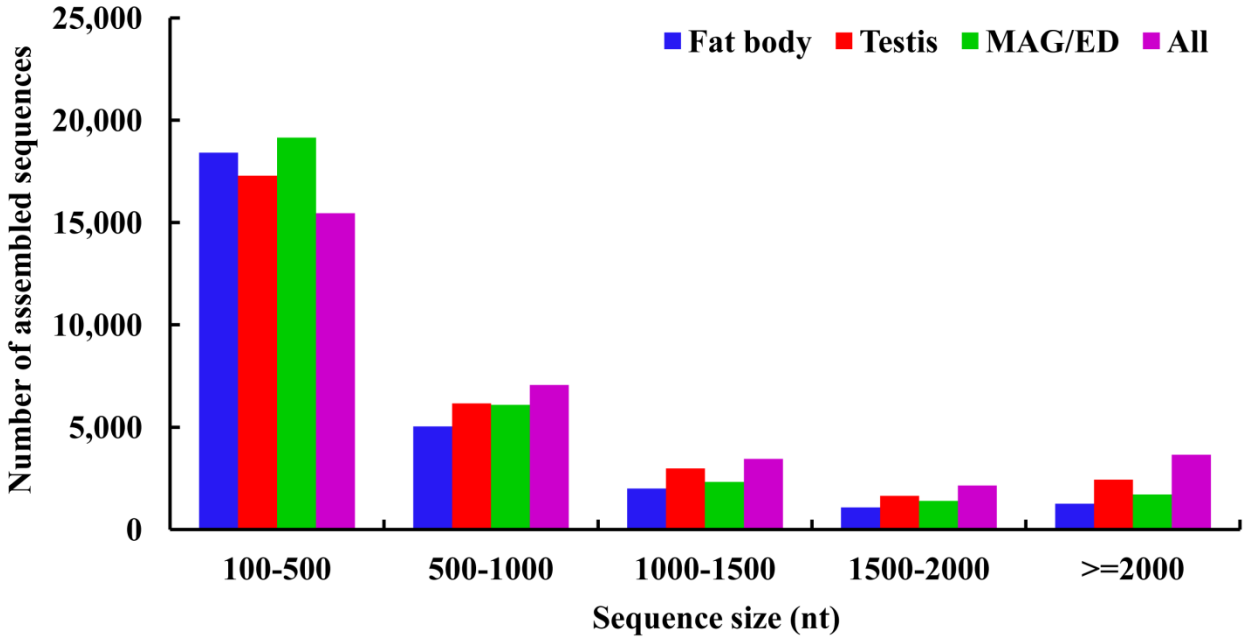

Fig. S1. Statistics of sequence length in 3 tissue transcriptomes of *Bactrocera dorsalis*. nt, nucleotide.

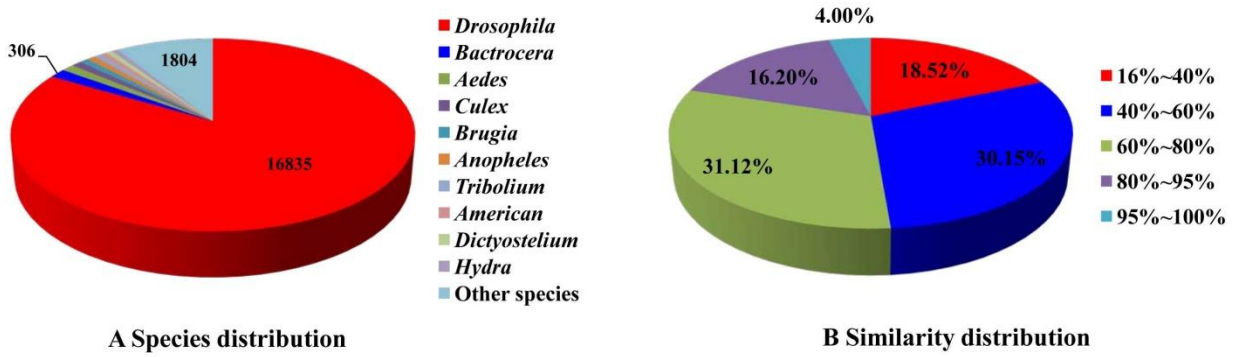

Fig. S2. Species (A) distribution and (B) similarity of unigenes in re-assembled transcriptomes from 3 tissues of *Bactrocera dorsalis*.

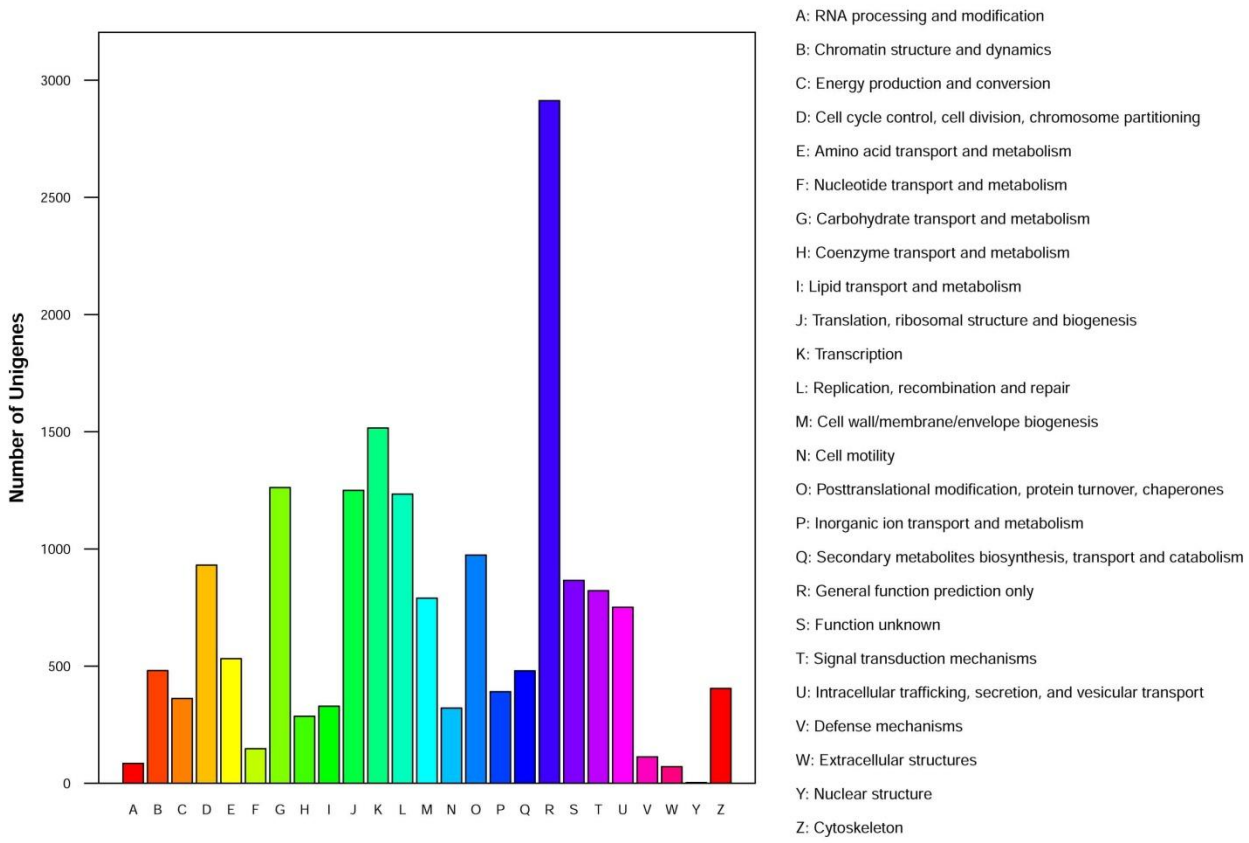

Fig. S3. Clusters of orthologous groups (COG) functional classification of unigenes in re-assembled transcriptome from 3 tissues of *Bactrocera dorsalis*.

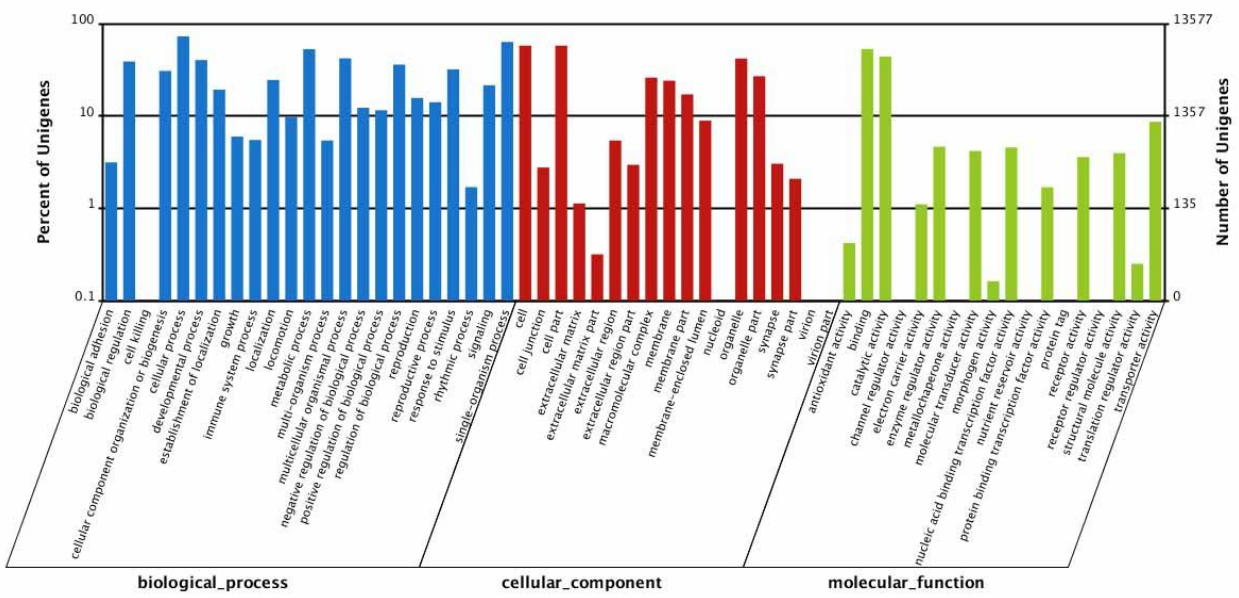

Fig. S4. Gene ontology (GO) classification of unigenes in re-assembled transcriptome from 3 tissues of *Bactrocera dorsalis*.

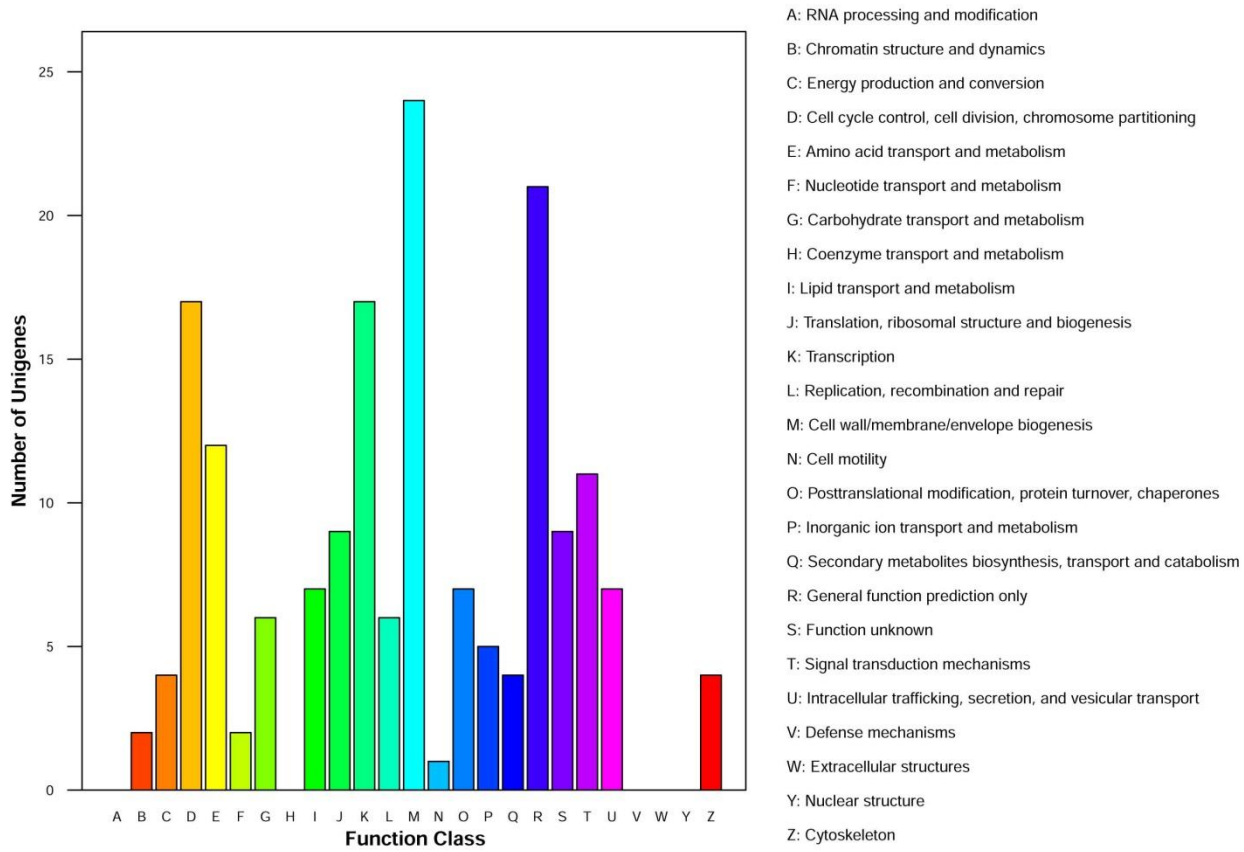

Fig. S5. Tissue expression profiling of 5 possible functional distinct unigenes with a length of <500 bp, which were highly expressed in male accessory glands and ejaculatory duct of *Bactrocera dorsalis*. Relative expression levels were determined as described in Fig. 5.