



UPPSALA
UNIVERSITET

UPTEC X 19004

Examensarbete 30 hp
September 2020

An RNA comparison study between
the Amazonian, Centro-American
and Orinocan semispecies of
Drosophila paulistorum

Erik Hedman



UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
<http://www.teknat.uu.se/student>

Abstract

An RNA comparison study between the Amazonian, Centro-American and Orinocan semispecies of *Drosophila paulistorum*

Erik Hedman

Differential expression analysis can be a powerful method to investigate expressed differences between closely related species. Our ambition is to highlight differentially expressed nuclear genes to explain the hybrid incompatibilities among the Amazonian, Centro-American and Orinocan semispecies of *Drosophila paulistorum*. RNA sequencing (RNA-seq) establishes the foundation of the study where we first evaluate the influence of two distinct alignment references. We discover the benefits of concatenating a *de novo* assembly instead of using the genome reference of a close relative. The bioinformatic pipeline handles the interesting inclusion of *D. melanogaster* and *D. willistoni*, where their contribution assists in the search for previously studied speciation genes. Among the down- and upregulated subsets we can see a diverse mix of general biological processes such as regulatory functions and transcriptional factors. In the end we uncover potential indications to why the Amazonian seems to be the least compatible semispecies to produce hybrids. This study provides a competitive working frame for comparative RNA-seq studies between closely related species.

Handledare: Lisa Klasson
Ämnesgranskare: Manfred Grabherr
Examinator: Jan Andersson
ISSN: 1401-2138, UPTEC X 19004

Populärvetenskaplig sammanfattning

Den centrala dogmen beskriver inom molekylärbiologin de mest fundamentala processer som kontinuerligt sker i alla celler hos samtliga organismer i livets träd. Replikation av DNA till DNA, transkription av DNA till RNA och slutligen translation av RNA till protein. Beroende på organism och en mängd olika faktorer som till exempel ålder, föda och levnadsmiljö, ökar eller minskar produktiviteten i den centrala dogmens olika steg. DNA, även känt som arvsmassa eller genom, består till stor del av fyra byggstenar kallade nukleotider. Dessa har ett socker, en eller flera fosfatgrupper och en utav de fyra kvävebaserna: adenin (A), tymin (T), cytosin (C) och guanin (G). RNA kan till stor del beskrivas som transkriptomet och är då delar av det översatta genomet. Det finns även här fyra byggstenar, men med två skillnader från DNA. Nukleotiderna har ett annat socker i RNA och kvävebasen tymin är i RNA utbytt mot uracil (U). Beroende på hur dessa byggstenar sitter sammanlänkade varierar informationen i arvsmassan och då även översättningen från DNA till RNA och från RNA till protein.

2001 publicerades den första kompletta avläsningen av ett mänskligt genom, en så kallad helgenomsekvensering, och den totala kostnaden landade på drygt \$2.7 miljarder. Det tog cirka 13 år att sekvensera alla byggstenar i människans DNA och ska ses som ett fantastiskt startskott för alla naturvetare runt om i hela världen. Idag närmar sig liknande studier en kostnad på \$1000 och en tidsåtgång på runt 2-3 dygn. Detta möjliggör nya dimensioner för forskningen och vi kan idag studera molekylära skillnader/likheter mellan olika organismer i allt större utsträckning och noggrannhet.

I det här projektet jämförs RNA mellan tre underarter hos en av bananflugans flitigt studerade medlemmar, *Drosophila paulistorum*. Målet med projektet är att belysa vilka gener som är olika mycket aktiverade i de tre underarterna och på så vis försöka förklara varför de skiljer sig åt. Bananflugan är en modellorganism och har sedan mitten på 1900-talet genomgått en mängd olika studier för att öka förståelsen kring generella egenskaper hos eukaryota organismer. Tack vare bananflugans levnadssätt och levnadsmiljö utgör den en exemplarisk modell för att forska på artbildning, något som är en komplex och oerhört viktig grundpelare inom biologin.

De allra flesta Nobelpris är resultatet av en lång resa som någonstans började i en forskargrupp med fokus på grundforskning. Tasuku Honjo vann 2018 års Nobelpris i medicin, där han med sina kunskaper inom immunologi lyckades skapa en mycket effektiv behandling av bland annat hud- och lungcancer. Utan en genuin förståelse för hur immunförsvarets olika komponenter fungerar, hade Tasuku och hans kollegor inte kunnat bygga de broar som tog dem till de nya upptäckterna. Konceptuellt gäller detsamma för all form av grundforskning. Det vi lär oss om bananflugans olika gener kan på sikt utveckla ny kunskap inom till exempel medicin och neurovetenskap. Tillsammans med den nya tekniken och den gedigna arbetsmoral som hittas i forskare likt den 72-årige Tasuku Honjo, dröjer det inte länge förrän vi har nya svar på gamla frågor.

Table of Contents

Abbreviations.....	9
1 Introduction.....	11
1.1 <i>Drosophila paulistorum</i>	11
1.2 Project goals	11
1.2.1 RNA-seq alignment reference	12
1.2.2 Mapping bias and functional annotation.....	12
2 Materials & Methods.....	13
2.1 Differential expression analysis	14
2.1.1 Evaluation of aligned reads	14
2.2 Differential expression analysis	14
2.2.1 Reference matrix.....	15
2.2.2 Reference count matrix	15
2.2.3 Condition table	15
2.2.4 Differential expression analysis	15
2.3 Intersection of DE subsets	16
2.4 Annotation table.....	16
2.5 GO enrichment.....	17
3 Results	17
3.1 Alignment of RNA-seq	17
3.2 Differential expression analysis	18
3.3 Intersections of DE subsets.....	19
3.4 Annotation table.....	20
3.5 GO enrichment.....	20
4 Discussion	24
5 Acknowledgment.....	26
6 References.....	27
7 Supplements.....	31

Abbreviations

AMCAOR	Amazonian, Central American, Orinocan
BAM	Binary Alignment Map
BLAST	Basic Local Alignment Search Tool
BWA	Burrows-Wheeler Aligner
DNA	Deoxyribonucleic acid
DE	Differential expression
FASTQ	Text-based format for storing biological sequence
FlyBase	A database for <i>Drosophila</i> genetics and molecular biology
GO	Gene Ontology
NCBI	National Center of Biotechnology Information
NGS	Next Generation Sequencing
RNA	Ribonucleic acid
L2FC	Log 2 Fold Change

1 Introduction

Speciation is a complex cornerstone in biology and can be explained as the origin of reproductive barriers among populations. There are two kinds of reproductive barriers, both resulting in reduced fitness in hybrids. Extrinsic reproductive isolation is dependent on environmental factors and takes form through ecological or sexual selection. In contrast, intrinsic reproductive isolation is not dependent on the environment. These barriers are triggered by genetic drift, or genomic conflict, which causes genetic incompatibilities (Seehausen *et al.* 2014). Emphasized by the Dobzhansky-Muller model, hybrid incompatibilities are initiated by the interaction between nuclear genes that have been functionally diverged. The genes also must be separated over time in their respective hybridizing species to fulfil the model. This concept provides key aspects for research questions regarding incipient speciation (Orr 1996). Another linkage to speciation and hybrid incompatibilities is the regulation of gene expression, which is inherently based on interactions between loci. Differences in gene expression may play a major role in intrinsic post-zygotic isolation (Mack & Nachman 2017).

1.1 *Drosophila paulistorum*

This thesis will be focusing on a classical example of incipient speciation: the Neotropical fruit fly *Drosophila paulistorum*. The phylogram of the *Drosophila* species groups *D. paulistorum* within the willistoni group (Spassky *et al.* 1971). The more famous *D. melanogaster* is another relative, and together with the obscura group, they form sister clade with the willistoni group (*Drosophila* 12 Genomes Consortium *et al.* 2007). The *D. paulistorum* superspecies consists of six known semispecies: Andean-Brazilian (AB), Amazonian (AM), Centro American (CA), Interior (IN), Orinocan (OR), and Transitional (TR). Although, AB, IN and TR will not take part of this study. The geographical distribution for the semispecies is partially overlapping, and their morphology is inseparable. Despite this, they are seemingly incompatible. Both pre- and post-mating barriers keep the semispecies from hybridizing (Coyne & Orr 1996, Miller *et al.* 2010). In other words, a vast majority of intercrosses of the six semispecies result in no offspring or sterile male hybrids (Dobzhansky & Pavlovsky 1966).

1.2 Project goals

The observed frequency of sterile male hybrids suggests that intrinsic reproductive barriers might have a considerable impact on the *Drosophila* speciation. The ambition is to explain the hybrid incompatibilities by highlighting differentially expressed nuclear genes among the three semispecies AM, CA and OR. To address this ambition, we apply RNA-seq along with differential expression analysis. Additionally, we want to investigate the influence of various RNA-seq alignment references. We need to cautiously handle the mapping bias, since we are

aligning three semispecies to the same reference. Finally, we want to improve the functional annotation for non-model organisms. Figure 1 presents an overview of the project workflow.

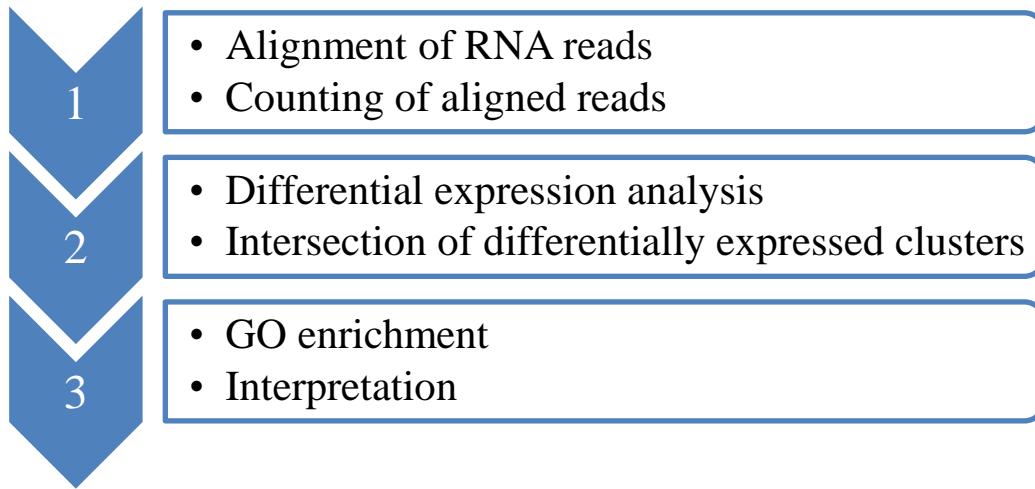


Figure 1. An overview of the project workflow divided into three major blocks.

1.2.1 RNA-seq alignment reference

In cases where a genome reference is available for the studied organism the initial identification of exons should be possible by mapping the RNA reads onto the reference. However, when an appropriate genome reference is not available, the quantification could be accomplished by first assembling the RNA reads de novo into contigs and then aligning the reads onto the assembled transcriptome (Conesa *et al.* 2016). Parekh *et al.* (2018), constructed a high-quality reference genome by including several closely related species (Human, Chimpanzee, Gorilla, Orangutan, Marmoset and Macaque). All the reads from each specie were then cross mapped onto the high-quality reference. Parekh *et al.* (2018) justified that cross-mapping quantification of expression levels are likely to be accurate, and sometimes even preferable for closely related species.

Taking this into consideration, we will apply two separate alignment references: a genome reference and a de novo assembly. The three semispecies have no available genome references. Therefore, we will use a close relative. The phylogram of the *Drosophila* species presents *D. willistoni* as the best fit genome reference for AM, CA and OR. The second reference regards the possibility of finding exons present in the three semispecies but absent in *D. willistoni*. This challenge is tested by using similar ideas as in Parekh *et al.* where we concatenate a high-quality de novo assembly, using the transcripts from AM, CA and OR.

1.2.2 Mapping bias and functional annotation

For genome-wide quantification of RNA, one of the technical hurdles lie in the need to map short RNA reads back to their correct locations in the reference (Degner *et. al* 2009). We challenge this issue by comparing the alignment results from two RNA-seq alignment software. Moreover, a genome reference of a close relative and a concatenated de novo assembly create additional aspects to handle. For an example, the genome reference might result in a high number of unmapped reads due to lower similarity. This leads to fewer differentially expressed

genes, which reduces the possibilities of downstream analysis. If only one (out of three) de novo assembly is applied as reference, the semispecies used as reference would probably present higher numbers of mapped reads compared to the other two semispecies. In this case, the observed expression levels could be dominated by technical issues rather than biological differences. To avoid this, we concatenate the three de novo assemblies into one reference. Hopefully, this justify a more balanced mapping basis.

Finally, orthologous groups can provide various kinds of valuable information for studies of comparative genomics (Li *et al.* 2003). Therefore, this study explores the benefits of using a reference matrix which is matched with the counted RNA reads. The reference matrix is based on BLAST searches, comparing proteins from the species *D. melanogaster*, *D. willistoni* and the three semispecies. This approach results in clustered genes and compensates for some of the RNA reads aligning to multiple loci across the concatenated de novo assembly. In other words, the clusters work as safety nets for the gene counts within each cluster. Another advantage is the inclusion of *D. melanogaster*, which is a superior model of animal genetics etc. (*Drosophila* 12 Genomes Consortium *et al.* 2007), hence a more enriched genome annotation. We believe this strategy can decrease the mapping bias and simultaneously increase the overall number of annotated genes for this comparison study of *D. paulistorum*.

2 Materials & Methods

The initial dataset consisted of 72 RNA FASTQ files of 125 bp paired end reads generated by Illumina HiSeq2500. Before the start of the project, the files had been quality-checked with FastQC (Andrew S. 2010) and trimmed using Trimmomatic (Bolger *et al.* 2014). Additionally, the de novo assemblies of AM, CA and OR had been completed with Trinity v.2.1.1 (Grabherr *et al.* 2011). To find out more details of living flies and RNA extraction, preprocessing of data, assembly of the RNA reads, evaluation and quality checks of the assemblies see materials & methods in Baião *et al.* (2019). Table 1 presents a simplified overview of the initial dataset.

Table 1. Data table of the paired end FASTQ files used in the project. *Each tissue has three samples and each sample has forward and reverse strand, adding up to 72 files. Abd = Abdomens.

Semispecie	Sex	Tissue*	Sex	Tissue*
Amazonian	Female	Abd/Head	Male	Abd/Head
Centro American	Female	Abd/Head	Male	Abd/Head
Orinocan	Female	Abd/Head	Male	Abd/Head

2.1 Differential expression analysis

The RNA-seq alignment was performed using two aligners: STAR v2.5.2b (Dobin *et al.* 2014) and BWA mem v0.7.10 (Li 2013). The STAR aligner performs spliced transcripts alignment where non-canonical splices are handled. Furthermore, it distinguishes between two main categories of alignments, uniquely mapped and mapped onto multiple loci, given an identity threshold. Unmapped reads are reads that either did not align at all or would have been aligned to a large number of loci. BWA mem does not discover spliced junctions, although a fast algorithm for mapping low-divergent sequences of high quality. Both aligners detect chimeric transcripts.

The STAR alignment was run with both the genome reference of *D. willistoni* (downloaded from FlyBase), and the concatenated de novo assembly. From now on the concatenated de novo assembly will be referred to as the AMCAOR reference. Baião G.C. provided the scripts for running the alignments with STAR. Default parameters were used, except intron lengths and number of allowed multimapped reads. *D. willistoni* was run with “--alignIntronMin 30” and “--alignIntronMax 185000”, according to an annotation release from NCBI. The AMCAOR reference used “--alignIntronMin 1” and “--alignIntronMax 2” since absence of introns in the transcriptome. Threshold for multimapped reads was changed from default “10” to “30”.

The BWA mem alignment was performed by Baião G.C. and the AMCAOR reference was used. Default settings.

2.1.1 Evaluation of aligned reads

The performance of the two aligners was analysed using SAMtools (Li *et al.* 2009). The number of mapped reads were extracted with **samtools view -F 4 file.bam | wc -l**. The BAM file was viewed, the mapped reads were selected with the **-F 4** command, and the number of rows were counted. The mapped reads were handled with the following command: **samtools view -F 4 file.bam | cut -f 3 | grep AM_ | wc -l**. After viewing the BAM file and selecting the mapped reads, the third column was cut out (containing the information of where the mapped reads are aligned in the AMCAOR reference), the **AM_** pattern was selected using **grep**, and finally the number of rows with the **AM** tag were counted. The same procedure was applied for reads mapping to **CA_** and **OR_**.

2.2 Differential expression analysis

The aligned reads using STAR were considered insufficient and put aside. Exclusively the BWA mem results were counted with featureCounts (Liao *et al.* 2014). Multimappers were specified, otherwise default settings were applied. The result file from featureCounts was transferred to RStudio v3.5.2 (RStudio Team 2016).

2.2.1 Reference matrix

Before running the differential expression analysis, we wanted to minimize the mapping bias further. Therefore, Baião G.C. created a reference matrix by clustering genes based on BLAST searches comparing proteins of the species *D. melanogaster*, *D. willistoni*, AM, CA and OR. The BLAST results were parsed to remove hits which were less likely to correspond to matches between orthologs. This means we removed hits in which “hit length” was lower than 60 % of “query length”, or vice versa, and hits in which “alignment length” was lower than 80 % of either “query length” or “hit length”. After parsing, orthoMCL v2.0.9 (Li *et al.* 2003) was run with default parameters using an inflation value of 1.5.

2.2.2 Reference count matrix

The reference matrix was matched against the count file from featureCounts. For each cluster, every gene with a hit got their counts from the count table, resulting in multiple rows for the cluster in progress. When all the genes in one cluster were handled, the rows of counts were summed up column wise, and next cluster went through the same procedure. This resulted in a new matrix with 37 columns and the same number of rows as the reference matrix. The new columns were the cluster ID and the 36 different samples, 12 from each semispecies. From now on this matrix is called reference count matrix. Supplementary tables S10:S12 present subsets of example matrices showing the process.

2.2.3 Condition table

Last step before running the differential expression analysis, initial conditions were specified. 36 rows with the column names: sample, ssp, condition, sex, and tissue. The complete pairwise comparison between the three semispecies was obtained by setting AM as the first reference, comparing tissues and sex from CA and OR. The second reference was set to CA, comparing tissue and sex from OR.

2.2.4 Differential expression analysis

To interpret differential data signals from RNA sequencing with good statistical power, it is required to estimate the variability throughout the complete data universe using a suitable error model. The genes are statistically tested if the observed difference in expression level are higher than what would be expected just due to natural random variation (Love *et al.* 2014). This procedure was performed using the R package DESeq2 (Love *et al.* 2014).

The reference count matrix from 2.2.2 and the condition table from 2.2.3 were compiled with the DESeq2 function in RStudio. Log2 fold change (L2FC) was used as an effect size estimate. L2FC is symmetric around 0 and L2FC > 1 was used to define the clusters with at least two times higher expression levels for the first semispecies in AMCA, AMOR and CAOR. L2FC < -1 was used to define the clusters with at least two times higher expression levels for the second semispecies in the pairwise comparisons. Threshold for an adjusted p-value was arbitrarily set to < 0.01, instead of the default value < 0.05. Subsets of DE clusters for all the pairwise comparisons were saved to files.

2.3 Intersection of DE subsets

After the DE analysis we wanted to investigate and clarify the results. Therefore, upSetR (Lex *et al.* 2014) was used to present the number of shared clusters between the subsets. The intersections of the DE AM clusters with higher expression levels compared to both CA and OR, were defined as upregulated in AM. The DE AM clusters with lower expression levels compared to both CA and OR, were defined as downregulated in AM. The same rules were applied for CA and OR. This procedure highlighted most of the DE clusters specific for each semispecies. If a cluster was down- or upregulated in three out of four samples within a semispecies, it was defined as globally down- or upregulated.

Important clarification for this study, the definition of regulated clusters can also represent absence and presence of clusters, and not necessarily the true down- and upregulation of the clustered genes. See table 2 for an overview of the criteria used to create the new subsets. From now on we will use down- and upregulation to describe the DE subsets.

Table 2. A condensed table defining the criteria of down- and upregulated subsets for the three semispecies. Each gene set consists of four different samples: female abdomens, female heads, male abdomens, and male heads. In total there are 24 subsets.

Gene set	Pairwise comparison 1	Pairwise comparison 2
AM Down	AMCA L2FC < -1	AMOR L2FC < -1
AM Up	AMCA L2FC > 1	AMOR L2FC > 1
CA Down	AMCA L2FC > 1	CAOR L2FC < -1
CA Up	AMCA L2FC < -1	CAOR L2FC > 1
OR Down	AMOR L2FC > 1	CAOR L2FC > 1
OR Up	AMOR L2FC < -1	CAOR L2FC < -1

2.4 Annotation table

The annotation table for this study was created using the contig annotation from Baião *et al.* (2019). In short, Baião *et al.* (2019) performed two independent strategies. Interproscan was run for the GO term annotation. The second strategy blasted all the contigs to a database with genes from *D. melanogaster*, *D. willistoni*, *Wolbachia*, *Saccharomyces cerevisiae* and several *Drosophila* gut bacteria. The contigs not fulfilling specified threshold were discarded. The new

annotation table was structured with the cluster ID followed by the corresponding GO term and gene name. If a cluster failed getting annotation it was excluded from the annotation table. However, we tried to find orthologous genes in *D. melanogaster* to annotate as much as possible. The cluster IDs were then used as keys to annotate all the down- and upregulated subsets.

2.5 GO enrichment

After the differential expression analysis, the gene ontology (GO) enrichment was run on the 24 down- and upregulated subsets. The enrichment analysis was used to find which GO terms were over-represented (or under-represented) in the different subsets. TopGO (Alexa *et al.* 2019) was used to perform the GO enrichment and the annotation table created in 2.4 was applied as the reference universe. The GO enrichment procedure followed the topGO tutorial, available at Bioconductor. In short, each of the 24 subsets were run together with the reference universe. Default statistics of classic Fisher, “weight01” algorithm and biological processes (BP) were specified to generate the results. In the end, additional threshold of classic Fisher value $< 0,05$ was used to extract the result files. Definitions of highlighted GO terms were gathered at www.geneontology.org and www.ebi.ac.uk/QuickGo/.

The GO enrichment was accompanied with the gene annotation. All subsets were sorted by ascending L2FC, and if a gene was present in at least three out of four subsets it was defined as globally down- or upregulated for the corresponding semispecie.

3 Results

Most of this study revolved around what reference to use for the RNA-seq alignment and how to compare the statistics between the two aligners. Furthermore, a great part focused on the definition of the down- and upregulated subsets and how to explain the biological differences despite the abundance of unannotated genes.

3.1 Alignment of RNA-seq

First out is the alignment statistics. Table 3 presents the percentage of mapped reads for all samples. We accomplished an overall 15-28 % increase of mapped reads by changing the reference from *D. willistoni* to the AMCAOR reference. Female samples mapped slightly better compared to male samples, using the *D. willistoni* reference. The AMCAOR reference resulted in a more similar distribution between the sexes. Comparing the different tissues, abdomen samples have a few percentages higher than head samples, except the CA female abdomens. Changing aligner to BWA mem resulted in 19-27 % higher number of mapped reads compared to the STAR run using the same reference. In fact, almost 100 % of all reads for both sexes were aligned running BWA mem with the AMCAOR reference. Interestingly, we got an

increased mapping specificity of 5-10 % by changing aligner from STAR to BWA mem (table 4). Hopefully, this clarifies why we continued with the mapped reads from BWA mem and excluded the STAR results from downstream analysis.

Table 3. The average percentage of mapped reads. In blue, the STAR aligner was run with *D. willistoni* and the AMCAOR reference. In red, the BWA mem aligner was run with the AMCAOR reference. Forward slash separates the numbers between females (F) and males (M). BWA mem resulted in almost 100 % of mapped reads for both sexes, hence no separation of numbers. Each number refers to the mean value of three samples abd1, abd2, abd3 and head1, head2, head3.

F / M	AM abd	CA abd	OR abd	AM head	CA head	OR head
STAR D. wil	66 / 54	56 / 54	61 / 50	58 / 59	59 / 55	57 / 53
STAR AMCAOR	81 / 79	73 / 79	76 / 78	79 / 78	76 / 74	74 / 70
BWA AMCAOR	~100	~100	~100	~100	~100	~100

Table 4. The average percentage of mapping specificity for each semispecie. STAR and BWA mem aligner were run with the AMCAOR reference. The numbers explain to which semispecie the mapped reads are aligned.

	STAR AM transcript	STAR CA transcript	STAR OR transcript	BWA AM transcript	BWA CA transcript	BWA OR transcript
AM reads	58	16	26	63	14	23
CA reads	24	50	26	18	60	22
OR reads	26	18	56	19	18	63

3.2 Differential expression analysis

The reference matrix from 2.2.1 had a total of 31356 clusters, where the biggest cluster had 431 genes. 60 % of the clusters had two or more genes and the remaining 40 % consisted of single genes not assigned to any cluster.

Figure 2 presents the number of DE clusters in each sample. The pairwise comparison AMCA presents more DE clusters than both AMOR and CAOR. CA has more DE clusters than all comparisons except for CAOR abdomens. In 8 out of 12 cases male samples present a slightly higher numbers of DE clusters compared to its corresponding female sample. All head samples have higher number of DE clusters compared to their corresponding abdomen sample.

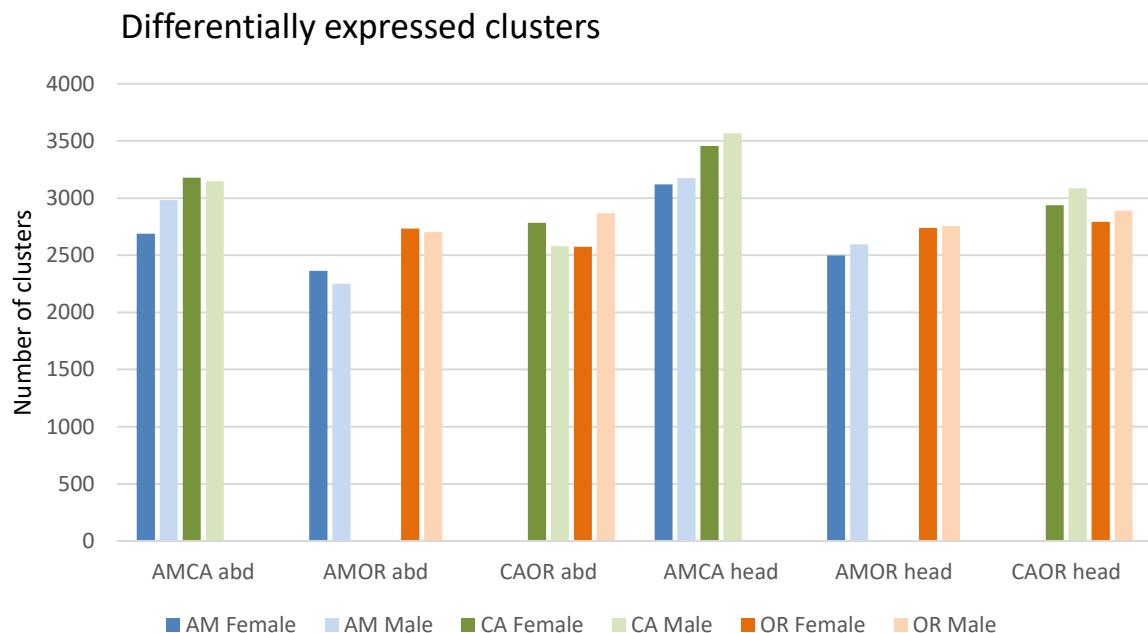


Figure 2. Bar plot of the DE clusters. The reference count matrix from the BWA mem run and the AMCAOR reference was used. Numbers are from the DESeq2 function using the pairwise comparison between semispecies, sex and tissue. Each group has a reference which is AM or CA. The DE clusters for the reference are defined with $L2FC > 1$. The DE clusters for the comparing semispecies are defined with $L2FC < -1$. Additional threshold of adjusted p-value < 0.01 was used for all groups.

3.3 Intersections of DE subsets

The upSetR plots display the number of shared clusters between the DE subsets. These results can be studied in the supplementary figure S3 and S4. Table 5 presents the statistics of the down- and upregulated female subsets, whereas the corresponding male subsets can be studied in the supplementary table S13. In general, the upregulated subsets have twice as many clusters as the downregulated subsets. The downregulated subsets have 11-26 % of single gene clusters, and the upregulated subsets have 57-70 % of single gene clusters. Less than 1 % of the single gene clusters in the downregulated subsets are associated with its corresponding semispecies. The single gene clusters are evenly distributed between the two other semispecies. In the upregulated subsets, almost 100 % of the single gene clusters are associated with its corresponding semispecies. This is true for all subsets.

11-16 % of the upregulated subsets have annotation, whereas the downregulated subsets have between 23-31 % of annotation. The number of annotated clusters correlates with the number of clusters containing *D. melanogaster* and/or *D. willistoni*. Only a few clusters without genes from *D. melanogaster* and *D. willistoni* managed to get annotated using the annotation table.

Table 5. Overview of the down- and upregulated female subsets. TOT presents the total number of clusters in each subset followed by the percentage of single gene clusters. AM presents the total percentage of clusters with AM genes followed by the percentage of single gene AM clusters. The same goes for CA and OR. MEL presents the total percentage of clusters with melanogaster genes. The same goes for WIL, which stands for willistoni. ANN presents the total percentage of annotated clusters in each subset.

FEMALE	TOT	AM	CA	OR	MEL	WIL	ANN
AM_abd_Down	681 (18)	30 (< 1)	88 (44)	89 (56)	22	22	25
CA_abd_Down	730 (24)	87 (54)	27 (< 1)	86 (46)	24	24	26
OR_abd_Down	502 (15)	93 (60)	90 (40)	33 (< 1)	29	30	31
AM_abd_Up	1143 (63)	99 (99)	21 (1)	20 (0)	13	14	14
CA_abd_Up	1537 (57)	27 (1)	98 (98)	27 (1)	14	15	16
OR_abd_Up	1084 (61)	23 (0)	22 (1)	99 (99)	12	13	14
AM_head_Down	712 (18)	25 (< 1)	88 (44)	90 (56)	22	21	23
CA_head_Down	792 (21)	86 (48)	26 (0)	89 (52)	24	24	26
OR_head_Down	518 (13)	92 (46)	81 (54)	25 (0)	21	21	24
AM_head_Up	1297 (63)	99 (100)	19 (0)	19 (0)	11	12	12
CA_head_Up	1427 (70)	25 (0)	100 (100)	25 (< 1)	11	11	13
OR_head_Up	1162 (62)	21 (1)	20 (1)	99 (98)	11	12	13

3.4 Annotation table

The annotation table resulted in 9046 rows and even though we tried finding orthologs in *D. melanogaster* a total of 22310 (71 %) clusters were excluded due to no annotation. Unfortunately, most of the smaller clusters with potentially novel *D. paulistorum* genes were excluded and practically none of the single gene clusters got annotated.

3.5 GO enrichment

The GO enrichment resulted in a diverse mix of general GO terms, but also very specific GO terms such as establishment of imaginal disc-derived wing hair orientation. Some of the more general GO terms are metabolic process, multicellular organismal process, and response to stimulus. The number of GO terms in each set varies from 10 to 47. In general, OR has fewer number GO terms compared to AM and CA. Head samples have less GO terms than the abdomen samples. Tables 6:9 show the top-five GO terms for female samples. The complete

GO enrichment of the 24 down- and upregulated subsets are placed in the supplementary tables S14:S37.

Both AM and OR present protein phosphorylation as a downregulated BP (table 6). Protein phosphorylation is a post-translational modification of proteins, causing an altered structure confirmation of the protein. This can activate, deactivate, or modify the protein function (Cohen 2002). OR show regulation of biosynthetic processes as upregulated for both abdomen and head samples (table 7, 9). The biosynthetic processes modulate the rate, frequency or extent of the chemical reactions resulting in formation of substances. Positive regulation of cellular metabolic process is another upregulated BP in OR, in both abdomen and head samples (table 7, 9). This regulation activates or increases the frequency, rate, or extent of the chemical reactions by which cells transform chemical substances. Peptide transport is upregulated in AM abdomens (table 7), whereas protein transport is downregulated in AM head samples (table 8). CA presents response to stimulus as downregulated in abdomens (table 7). This refer to any process that results in a change in state or activity of a cell because of a stimulus. All BP definitions are gathered from www.geneontology.org and www.ebi.ac.uk/QuickGo/.

Table 6. Female abdomens downregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

	GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
AM						
1	GO:0008152	metabolic process	4404	79	74,16	0,0091
2	GO:0006396	RNA processing	568	14	9,56	0,0462
3	GO:0006468	protein phosphorylation	417	13	7,02	0,0378
4	GO:0035107	appendage morphogenesis	368	10	6,2	0,0327
5	GO:0007015	actin filament organization	158	6	2,66	0,0126
CA						
1	GO:0050896	response to stimulus	2731	60	52,14	0,0179
2	GO:0016055	Wnt signalling pathway	138	7	2,63	0,0219
3	GO:0060446	branching involved in open tracheal system development	77	4	1,47	0,0375
4	GO:0051168	nuclear export	73	2	1,39	0,0191
5	GO:0007254	JNK cascade	70	4	1,34	0,0209
OR						
1	GO:0006366	transcription by RNA polymerase II	646	12	10,34	0,0132
2	GO:0006468	protein phosphorylation	417	11	6,68	0,0153
3	GO:0030036	actin cytoskeleton organization	312	11	4,99	0,0306
4	GO:0044770	cell cycle phase transition	171	7	2,74	0,0156
5	GO:0048599	oocyte development	162	4	2,59	0,0205

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table 7. Female abdomens upregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

	GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
AM						
1	GO:0015833	peptide transport	392	6	6,32	0,03231
2	GO:0008283	cell proliferation	379	10	6,11	0,04272
3	GO:0048737	imaginal disc-derived appendage development	376	12	6,06	0,04581
4	GO:0040008	regulation of growth	345	7	5,56	0,04071
5	GO:0043067	regulation of programmed cell death	241	3	3,89	0,04826
CA						
1	GO:0120039	plasma membrane bounded cell projection morphogenesis	650	22	15,65	0,04621
2	GO:0006357	regulation of transcription by RNA polymerase II	596	20	14,35	0,04119
3	GO:0030855	epithelial cell differentiation	432	14	10,4	0,00167
4	GO:0030707	ovarian follicle cell development	346	11	8,33	0,04966
5	GO:0000226	microtubule cytoskeleton organization	326	14	7,85	0,01978
OR						
1	GO:0032501	multicellular organismal process	3425	58	51,17	0,02736
2	GO:0009889	regulation of biosynthetic process	1207	17	18,03	0,01522
3	GO:0031325	positive regulation of cellular metabolic process	698	16	10,43	0,00062
4	GO:0010604	positive regulation of macromolecule metabolic process	688	14	10,28	0,00721
5	GO:0006508	proteolysis	488	15	7,29	0,01213

Table 8. Female heads downregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

	GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
AM						
1	GO:0015031	protein transport	372	10	6,09	0,0332
2	GO: 0007431	salivary gland development	200	7	3,27	0,0107
3	GO:0001933	negative regulation of protein phosphorylation	66	4	1,08	0,0239
4	GO:0042059	negative regulation of epidermal growth factor receptor signalling pathway	39	3	0,64	0,0255
5	GO:0007112	male meiosis cytokinesis	34	3	0,56	0,0177
CA						
1	GO:0008340	determination of adult lifespan	168	8	3,41	0,02082
2	GO:0042461	photoreceptor cell development	136	6	2,76	0,00572
3	GO:0007298	border follicle cell migration	127	6	2,58	0,04416
4	GO:0008358	maternal determination of anterior/posterior axis, embryo	106	5	2,15	0,00236
5	GO:0032543	mitochondrial translation	98	6	1,99	0,03953
OR						
1	GO:0017145	stem cell division	113	3	1,27	0,0273
2	GO:0048149	behavioral response to ethanol	62	3	0,7	0,0325
3	GO:0042157	lipoprotein metabolic process	55	3	0,62	0,0437
4	GO:0008039	synaptic target recognition	50	4	0,56	0,008
5	GO:0001737	establishment of imaginal disc-derived wing hair orientation	28	2	0,32	0,0393

Table 9. Female heads upregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

	GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
AM						
1	GO:0045892	negative regulation of transcription, DNA-templated	352	10	5,26	0,00654
2	GO:0006109	regulation of carbohydrate metabolic process	215	4	3,21	0,00325
3	GO:0007298	border follicle cell migration	127	5	1,9	0,04133
4	GO:0002121	inter-male aggressive behaviour	71	4	1,06	0,0214
5	GO:0008586	imaginal disc-derived wing vein morphogenesis	62	6	0,93	0,00031
CA						
1	GO:0000902	cell morphogenesis	766	20	14,26	0,0291
2	GO:0006979	response to oxidative stress	110	6	2,05	0,0351
3	GO:0010721	negative regulation of cell development	94	4	1,75	0,0169
4	GO:0007274	neuromuscular synaptic transmission	89	6	1,66	0,0046
5	GO:0007427	epithelial cell migration, open tracheal system	45	4	0,84	0,0227
OR						
1	GO:0009889	regulation of biosynthetic process	1207	18	17,46	0,0145
2	GO:0032774	RNA biosynthetic process	1095	18	15,84	0,0423
3	GO:0031325	positive regulation of cellular metabolic process	698	12	10,1	0,0424
4	GO:0007591	molting cycle, chitin-based cuticle	95	5	1,37	0,0115
5	GO:0090174	organelle membrane fusion	73	3	1,06	0,0424

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

The number of genes associated with each BP term was small. However, a few annotated genes were highlighted based on high L2FC values and presence.

The receptor of activated protein kinase C1 (*Rack1*) was observed as downregulated in all AM samples (except in male abdomen), compared to both CA and OR. This gene has been reported to suppress the regulation of protein phosphorylation (Dopie *et al.* 2015), but also to be an essential gene at multiple steps of development, particularly in oogenesis (Kadomas *et al.* 2007). Bag of marbles (*bam*) was found as upregulated in all AM samples (except in female head). *Bam* encodes a protein involved in gametogenesis and has been shown to be necessary to promote germ-line stem cells and cystoblast differentiation (Ji *et al.* 2017). Going forward, the gene *boca* was presented as downregulated in all CA samples. This evolutionary conserved *Drosophila* gene encodes an endoplasmic reticulum protein that is vital for the intracellular trafficking of low-density lipoprotein receptors (Culi and Mann, 2003). Another downregulated CA gene (except in female head) was the *BthD* selenoprotein. *BthD* has been reported to be highly expressed in the developing salivary gland, whereas the loss of *BthD* results in compromised salivary gland morphogenesis and reduced animal viability (Kwon *et al.* 2003). S6 kinase like (*S6KL*) was one of the upregulated CA genes (except in female abdomen) and has been reported to act as a negative regulator of bone morphogenetic protein (BMP) signalling. Results have demonstrated that *S6KL* regulates synaptic development and function

by promoting proteasomal degradation of the BMP receptor thickveins (Zhao *et al.* 2015). All CA samples expressed bifocal (*bif*) as an upregulated gene. *Bif* is a putative cytoskeletal regulator and by its interactions with the protein phosphatase-1 (PP1), *bif* mediates normal photoreceptor morphology in *Drosophila* (Babu *et al.* 2005). Finally, OR presented *Nedd4* as upregulated in all samples. *Nedd4* encodes an E3 Ubiquitin ligase where various studies imply its function to negatively regulate the Notch signalling pathway (Sakata *et al.* 2004, Wilkin *et al.* 2004, Jaekel and Klein, 2006, Zhu *et al.* 2017).

Two genes previously reported to be involved in speciation were found in AM and CA. Nucleoporin 98-96kD (*Nup98-96*) was upregulated in AM male abdomens and O-6-alkylguanine-DNA alkyltransferase (*agt*) was downregulated in CA male abdomens. Speciation genes are defined to cause reproductive isolation and are normally very rapidly evolving, often driven by positive Darwinian selection. They fall into many functional classes, although a role in transcriptional regulation could prove most common (Orr 2005). *Nup98-96* has been involved in hybrid sterility between *D. melanogaster* and *D. simulans* (Presgraves *et al.* 2003). The protein consists of two nucleoporins (Nup98 and Nup96) and operates in the nuclear pore complex (NPC). The NPCs are the only site of cytonuclear trafficking of RNAs and proteins and they are one of the largest macromolecular complexes in eukaryotic cells. Nup96 and Nup98 are both found on the cytoplasmic and nucleoplasmic sides of the NPC. However, Nup96 is stably bound at the NPC where it seems to have a structural role, whereas Nup98 is a mobile protein and shuttles on and off the NPC. Both nucleoporins operate in RNA export (Presgraves *et al.* 2003). *Agt* has been associated with hybrid male sterility between *D. simulans* and *D. mauritiana*. It is a small, intronless and fast evolving gene, and the molecular function is annotated as methylated-DNA-protein-cysteine S-methyltransferase activity (Araripe *et al.* 2010).

4 Discussion

The original objectives outlined in the introduction have been successfully accomplished. We have investigated the influence of two RNA-seq alignment references. We have cautiously diminished the mapping bias, and at the same time tested an alternative strategy to improve the functional annotation for *D. paulistorum*. The mapping sensitivity was enhanced by changing the reference from *D. willistoni* to the concatenated de novo assembly. Due to this change, we saw less benefits of using the STAR aligner, since its default parameters are designed for genome references (handling of introns, alternative splicing etc.). BWA mem resulted in an increased mapping specificity of 5-10 %, compared to STAR. We believe the reference matrix with clustered genes was the correct decision for this comparison study. Around 80 % of the annotation was derived through the orthologous genes in *D. melanogaster* and *D. willistoni*. In the end, we were left with the GO enrichment, a diverse mix of biological processes and the associated gene annotation. The ambition was to shed light on the hybrid incompatibilities by

pinpointing DE nuclear genes. Unluckily, the complexity of such investigation directed us to an alternative more general route while evaluating the results.

Overall, CA has the largest amount of DE clusters. According to previous studies (Zanini *et al.* 2018), the time since CA diverged from AM and OR is almost double the time of divergence between AM and OR. Additionally, CA is geographically separated from AM and OR. This might explain why CA presents higher numbers of DE clusters. Time and space have probably affected CA quite a lot. Going forward, Dobzhansky and Pavlovsky (1966) presented AM as the least compatible semispecie to produce hybrids, since only a small portion of all inter-crosses involving AM females or males resulted in progeny. CA females were described as less discriminating, and the OR females were intermediate in this respect. If we look at the diverse results from the GO enrichment, we can see that AM presents reproduction related GO terms as upregulated (inter-male aggressive behaviour and germ-line stem cell population maintenance). These observations might indicate an interesting hypothesis to the hybrid incompatibility involving the AM semispecie. Furthermore, OR and AM share several GO terms across the subsets, which is not that common between neither OR and CA nor AM and CA.

One notable difference between the semispecies is that OR seems to be more homogeneous regarding globally upregulated GO terms. AM and CA are more sex and tissue specific. OR and AM share some globally upregulated GO terms. However, these GO terms are only globally regulated in one of the semispecies. For an example, positive regulation of cellular metabolic process is upregulated in all OR samples, but only in AM male heads. Imaginal disc-derived wing vein morphogenesis is upregulated in all AM samples except in male heads, whereas in OR only female abdomens present this GO term as upregulated. Two GO terms were highlighted among the upregulated CA head samples (cell morphogenesis, and epithelial cell migration, open tracheal system). Otherwise, CA presented more interesting findings among the downregulated GO terms. Response to stimulus, nuclear export, JNK cascade, and photoreceptor cell development were all globally downregulated GO terms across most of the CA samples. The conspicuous finding is the nuclear export, since this can be connected to the speciation gene *Nup98-96*, which is upregulated in AM male abdomens. Regulated differences between nuclear genes might not fulfil the original Dobzhansky Muller model, but it may play a major role in intrinsic post-zygotic isolation (Mack & Nachman 2017). Therefore, I believe it would be very interesting to investigate further. For instance, related genes or GO terms could be used as guidance when looking for patterns between AM and CA. Another idea is to continue the development of the reference matrix to enhance the annotation. The speciation gene concept was discovered late in the project, and it would be very interesting to see if any improvement of the reference matrix could result in further detection of other speciation genes. Examples of previously studied speciation genes within *Drosophila* are *OdsH*, *Hmr* and *tmy* (Orr 2005), and *Zhr*, *Lhr* and *JYalpha* (Araripe *et al.* 2010). The common denominator for all these genes is their involvement in hybrid sterility and/or inviability. Genes leading to sterility in hybrids would likely be involved in some aspect of reproduction, whereas genes causing hybrid

inviability may have important regulatory, housekeeping, or developmental functions (Araripe *et al.* 2010)

Zanini *et al.* hypothesize that the origin and further diversification of the *willistoni* subgroup took place in the Amazonian area, the North region of South America. Even though the difference in mapping specificity was small, both AM and CA seemed to prefer OR when mapping outside of their own transcriptome in the concatenated de novo assembly (table 4). AM accomplished slightly higher number of mapped reads against the *willistoni* reference, compared to CA and OR (table 5). However, the differences were too small and any indication regarding relatedness due to mapping specificity must be ignored.

Overall, it has been a joyful ride with both ups and downs, probably much similar to the flying pattern of our beloved little flies. I believe the bioinformatic pipeline constructed for this project has great potential to serve as an excellent frame for most comparative genomic studies between closely related species. Hopefully, our contribution will assist in future projects of similar character.

5 Acknowledgment

A special acknowledgement to the people making this project possible.

Subject reader Manfred Grabherr has ensured the relevance and quality of the research. His expertise within the field secured the structure for planning the project, but he also assisted with insightful notes during the interpretation of the results. Thank you, Manfred!

Supervisor Lisa Klasson, who accepted my enquiry of master thesis in her research group. She offered profound guidance and expertise throughout the project. It has been a joyful life lesson working with Lisa, her positive attitude and friendly leadership. I sincerely hope future students will get the opportunity to learn, discuss and evaluate ambiguous results just like we did in this project. Thank you, Lisa!

PhD student Guilherme Costa Baião, who provided day-to-day knowledge throughout the project. He helped with scripting, discussions, and essential support when progression was hard to present. I was very fortunate having such a patient and pedagogical guide. Thank you, Guilherme!

Furthermore, I want to thank all the great people in the Department of Molecular Evolution. You know who you are, thank you!

Finally, I want to thank Jan Andersson and Lena Henriksson for coordinating the master thesis course. I also want to highlight their day to day work where they constantly improve the master's programme in molecular biotechnology engineering. Thank you, Jan and Lena!

6 References

- Alexa A, Rahnenfuhrer J. 2019. topGO: Enrichment Analysis for Gene Ontology. R package version 2.34.0. doi 10.18129/B9.bioc.topGO.
- Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. URL: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
- Araripe LO, Montenegro H, Lemos B, Hartl DL. 2010. Fine-scale genetic mapping of a hybrid sterility factor between *Drosophila simulans* and *D. mauritiana*: the varied and elusive functions of “speciation genes”. BMC Evolutionary Biology, doi 10.1186/1471-2148-10-385.
- Babu K, Bahri S, Alphey L, Chia W. 2005. Bifocal and PP1 interaction regulates targeting of the R-cell growth cone in *Drosophila*. Developmental Biology 288: 372-386.
- Baião GC, Schneider DI, Miller WJ, Klasson L. 2019. The effect of *Wolbachia* on gene expression in *Drosophila paulistorum* and its implications for symbiont-induced host speciation. BMC Genomics, doi 10.1186/s12864-019-5816-9.
- Benjamin AM, Nichols M, Burke TW, Ginsburg GS, Lucas JE. 2014. Comparing reference-based RNA-seq mapping methods for non-human primate data. BMC Genomics, doi 10.1186/1471-2164-15-570.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120.
- Cohen P. 2002. The origins of protein phosphorylation. Nature Cell Biology 4: E127-E130.
- Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Wojciech S. M, Gaffney J. D, Elo L. L, Zhang X, Mortazavi A. 2016. A survey best practices for RNA-seq data analysis. Genome Biology, doi 10.1186/s13059-016-0881-8.
- Coyne JA, Orr HA. 1996. Patterns of speciation in *Drosophila*. Evolution 51: 295-303.
- Culi J, Mann RS. 2003. Boca, an endoplasmic reticulum protein required for Wingless signaling and trafficking of LDL receptor family members in *Drosophila*. Cell 112: 343-354.
- Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK. 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics 25: 3207-3212.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15-21.

Dobzhansky T, Pavlovsky O. 1966. Experiments of incipient species of the *Drosophila paulistorum* complex. *Genetics* 55: 141-156.

Drosophila 12 genome consortium, 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450: 203-218.

Dopie J, Rajakylä EK, Joensuu MS, Huet G, Ferrantelli E, Xie T, Jäälinöja H, Jokitalo E, Vartiainen MK. 2015. Genome-wide RNAi screen for nuclear actin reveals a network of cofilin regulators. *Journal of Cell Science* 128: 2388-2400.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Linblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology* 29: 644-652.

Jaekel R, Klein T. 2006. The *Drosophila* Notch inhibitor and tumor suppressor gene lethal (2) giant discs encodes a conserved regulator of endosomal trafficking. *Developmental Cell* 11: 655-669.

Ji S, Li C, Hu L, Liu K, Mei J, Luo Y, Tao Y, Xia Z, Sun Q, Chen D. 2017. Bam-dependent deubiquitinase complex can disrupt germ-line stem cell maintenance by targeting cyclin A. *Proceedings of the National Academy of Sciences of the United States of America* 114: 6316-6321.

Kadrmas JL, Smith MA, Pronovost SM, Beckerle MC. 2007. Characterization of RACK1 function in *Drosophila* development. *Development Dynamics* 236: 2207-2215.

Kwon SY, Badenhorst P, Martin-Romero FJ, Carlson BA, Paterson BM, Gladyshev VN, Lee BJ, Hatfield DL. 2003. The *Drosophila* selenoprotein BthD is required for survival and has a role in salivary gland development. *Molecular Cell Biology* 23: 8495-8504.

Lex A, Gehlenborg N, Strobel H, Vuillemot R, Pfister H. 2014. UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics* 20: 1983-1992.

Li H. 2013. Aligning sequence reads, clone sequence and assembly contigs with BWA-MEM. ArXiv: 1303.3997.

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*, doi 10.1093/bioinformatics/btp352.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research* 13: 2178-2189.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30: 923-930.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, doi 10.1186/s13059-014-0550-8.

Mack KL, Nachman MW. 2017. Gene Regulation and Speciation. *Trends in Genetics* 33: 68-80.

Miller WJ, Ehrman L, Schneider D. 2010. Infectious Speciation Revisited: Impact of Symbiont-Depletion on Female Fitness and Mating Behavior of *Drosophila paulistorum*. *PLoS Pathogens*, doi 10.1371/journal.ppat.1001214.

Orr HA. 1996. Dobzhansky, Bateson, and the Genetics of Speciation. *Genetics* 144: 1331-1335.

Orr HA. 2005. The genetics basis of reproductive isolation: Insights from *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America* 102: 6522-6526.

Parekh S, Vieth B, Ziegenhain C, Enard W, Hellman I. 2018. Strategies for quantitative RNA-seq analyses among closely related species. *BioRxiv*, doi 10.1101/297408.

Presgraves D, Balagopalan L, Abmayr S, Orr HA. 2003. Adaptive evolution drives divergence of a hybrid inviability gene between two species of *Drosophila*. *Nature* 423: 715-719.

RStudio Team 2016. RStudio: Integrated Development for R. RStudio Inc., Boston, MA. URL: <https://www.rstudio.com>.

Sakata T, Sakaguchi H, Tsuda L, Higashitani A, Aigaki T, Matsuno K, Hayashi S. 2004. *Drosophila* Nedd4 regulates endocytosis of Notch and suppresses its ligand-independent activation. *Current Biology* 14: 2228-2236.

Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, Peichel CL, Saetre GP, Bank C, Bränström A, Brelsford A, Clarkson CS, Eroukmanoff F, Feder JL, Fischer MC, Foote AD, Franchini P, Jiggins CD, Jones FC, Lindholm AK, Lucek K, Maan ME, Marques DA, Martin SH, Matthews B, Meier JI, Möst M, Nachman MW, Nonaka E, Rennison DJ, Schwarzer J, Watson ET, Westram AM, Widmer A. 2014. Genomics and the origin of species. *Nature Reviews Genetics* 15: 176-192.

Spassky B, Richmond RC, Perez-Salas S, Pavlovsky O, Mourao CA, Hunter AS, Hoenigsberg H, Dobzhansky Th, Ayala FJ. 1971. Geography of the sibling species related to *Drosophila willistoni*, and of the semispecies of the *Drosophila paulistorum* complex. *Evolution* 25: 129-143.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10: 57-63.

Wilkin MB, Carbery AM, Fostier M, Aslam H, Mazaleyrat SL, Higgs J, Myat A, Evans DA, Cornell M, Baron M. 2004. Regulation of Notch endosomal sorting and signaling by *Drosophila* Nedd4 family proteins. *Current Biology* 14: 2237-2244.

Zanini R, Müller MJ, Vieira GC, Valiati VH, Deprá M, Valente VLDS. 2018. Combining morphology and molecular data to improve *Drosophila paulistorum* (Diptera, Drosophilidae) taxonomic status. *Fly* 12: 81-94.

Zhao G, Wu Y, Du L, Li W, Xiong Y, Yao A, Wang Q, Zhang YQ. 2015. *Drosophila* S6 Kinase Like Inhibits Neuromuscular Junction Growth by Downregulating the BMP Receptor Thickveins. *PLoS Genetics*, doi 10.1371/journal.pgen.1004984.

Zhu JY, Heidersbach A, Kathiriya IS, Garay BI, Ivey KN, Srivastave D, Han Z, King IN. 2017. The E3 ubiquitin ligase Nedd4/Nedd4L is directly regulated by microRNA 1. *Development* 144: 866-875.

7 Supplements

Table S10. A subset example of the reference matrix.

ClusterID	Genes
Cluster1	dwil FBpp0375213, dwil FBpp0376914, dpam m.16071, dpor m.74000, ...
Cluster2	dpam m.49646, dPCA m.79969, dPCA m.64192, ...
Cluster1018	dpam m.41219, dpam m.41336, dmel FBpp0070005, ...
ClusterN	dPCA m.59457, dpor m.28400
NoCluster1	dPCA m.55531
NoClusterM	dpor m.33457

Table S11. A subset example of the count matrix from featureCounts.

Genes	abd1	abd2	abd3	head1	head2	head3
dpam m.16071	0	0	0	1	7	7
dpor m.74000	0	0	0	12	9	5
dpam m.41336	34	11	21	89	60	24
dpam m.41219	47	23	32	130	127	71
dPCA m.55531	0	0	0	0	0	0
dpor m.33457	315	102	230	542	343	186

Table S12. A subset example of the corresponding reference count matrix.

ClusterID	AM_F abd1	AM_F abd2	AM_F abd3	AM_F head1	AM_F head2	AM_F head3
Cluster1	0	0	0	13	16	12
Cluster1081	81	34	53	219	187	95
NoCluster1	0	0	0	0	0	0
NoClusterM	315	102	230	542	343	186

Table S13. Overview of the down- and upregulated male subsets. TOT presents the total number of clusters in each subset followed by the percentage of single gene clusters. AM presents the total percentage of clusters with AM genes followed by the percentage of single gene AM clusters. The same goes for CA and OR. MEL presents the total percentage of clusters with melanogaster genes. The same goes for WIL, which stands for willistoni. ANN presents the total percentage of annotation in each subset.

MALE	TOT	AM	CA	OR	MEL	WIL	ANN
AM_abd_Down	692 (16)	33 (<1)	90 (50)	89 (50)	25	26	30
CA_abd_Down	823 (22)	88 (55)	26 (0)	86 (45)	24	26	28
OR_abd_Down	458 (11)	95 (64)	92 (36)	31 (0)	29	30	31
AM_abd_Up	1150 (61)	99 (100)	23 (<1)	23 (<1)	15	15	16
CA_abd_Up	1427 (59)	25 (1)	99 (98)	24 (1)	13	14	15
OR_abd_Up	1155 (61)	24 (<1)	23 (<1)	99 (100)	12	13	14
AM_head_Down	728 (17)	26 (<1)	89 (47)	90 (53)	23	22	24
CA_head_Down	827 (26)	86 (73)	22 (<1)	84 (27)	21	22	23
OR_head_Down	519 (12)	94 (58)	91 (42)	26 (<1)	21	21	24
AM_head_Up	1338 (63)	99 (99)	19 (<1)	18 (<1)	10	11	11
CA_head_Up	1716 (62)	21 (1)	99 (98)	21 (1)	10	10	11
OR_head_Up	1170 (63)	20 (0)	19 (1)	99 (99)	11	12	13

AMCA_AM corresponds to the DE clusters with a L2FC > 1. AMCA_CA corresponds to the DE clusters with a L2FC < -1, and so on. CAOR_CA and AMCA_CA define the upregulated clusters for CA compared to both AM and OR. CAOR_OR and AMCA_AM define the downregulated clusters for CA compared to both AM and OR, and so on.

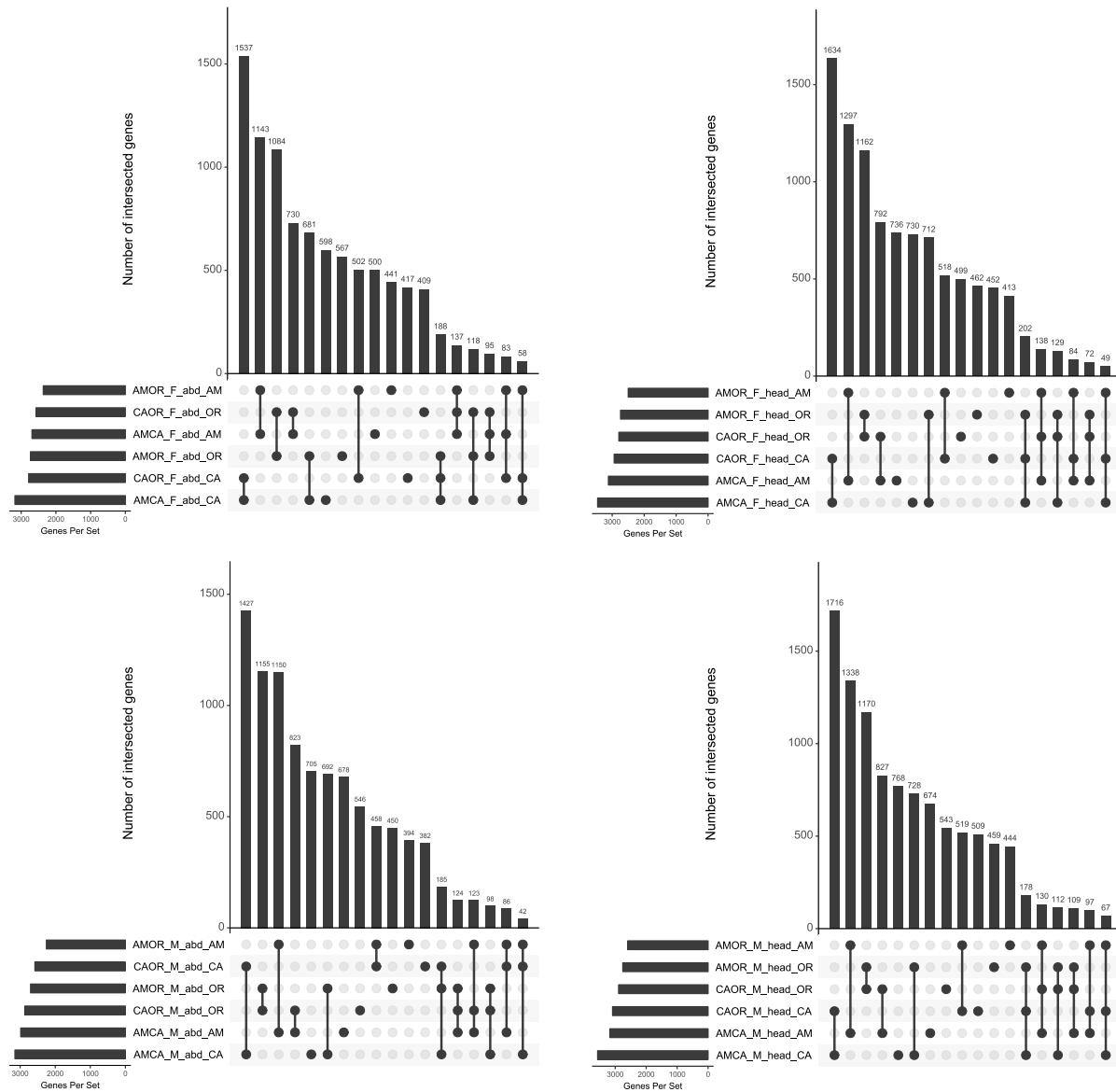


Figure S3. A panel plot from upSetR with female abdomens, heads, and male abdomens, heads.

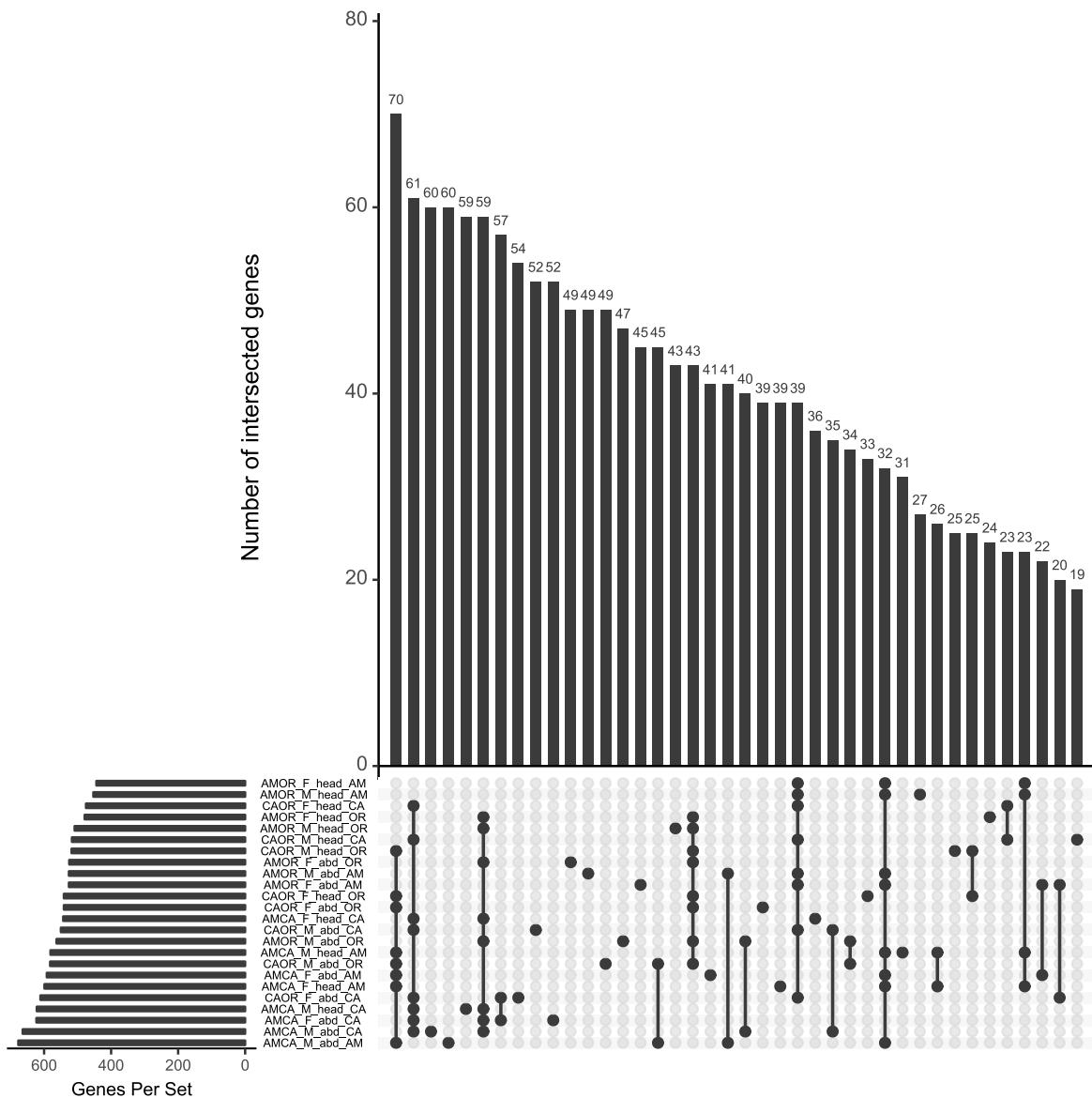


Figure S4. upSetR plot of all intersected DE clusters, exclusively with annotation.

Table S14. AM female abdomens downregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0008152 metabolic process	4404	79	74,16	0,0091
2	GO:0006396 RNA processing	568	14	9,56	0,0462
3	GO:0006468 protein phosphorylation	417	13	7,02	0,0378
4	GO:0035107 appendage morphogenesis	368	10	6,2	0,0327
5	GO:0007015 actin filament organization	158	6	2,66	0,0126
6	GO:0001933 negative regulation of protein phosphorylation	66	4	1,11	0,0252
7	GO:0032868 response to insulin	54	3	0,91	0,0331
8	GO:0006888 ER to Golgi vesicle-mediated transport	54	4	0,91	0,0391
9	GO:0045185 maintenance of protein location	47	3	0,79	0,0093
10	GO:0035023 regulation of Rho protein signal transduction	42	3	0,71	0,0476
11	GO:0042059 negative regulation of epidermal growth factor receptor signaling pathway	39	3	0,66	0,0275
12	GO:1905330 regulation of morphogenesis of an epithelium	35	3	0,59	0,004
13	GO:0048806 genitalia development	35	2	0,59	0,0333
14	GO:0007112 male meiosis cytokinesis	34	3	0,57	0,0191
15	GO:0034401 chromatin organization involved in regulation of transcription	33	4	0,56	0,0488
16	GO:0070868 heterochromatin organization involved in chromatin silencing	30	3	0,51	0,0136
17	GO:0000132 establishment of mitotic spindle orientation	23	3	0,39	0,0065
18	GO:0008069 dorsal/ventral axis specification, ovarian follicular epithelium	18	2	0,3	0,0361
19	GO:0045840 positive regulation of mitotic nuclear division	17	2	0,29	0,0325
20	GO:0045450 bicoid mRNA localization	16	2	0,27	0,029
21	GO:0008585 female gonad development	15	2	0,25	0,0256
22	GO:0000466 maturation of 5,8S rRNA from tricistronic rRNA transcript	14	2	0,24	0,0224
23	GO:0031290 retinal ganglion cell axon guidance	14	2	0,24	0,0224
24	GO:0045880 positive regulation of smoothened signaling pathway	12	2	0,2	0,0166
25	GO:0060250 germ-line stem-cell niche homeostasis	11	2	0,19	0,014
26	GO:0016556 mRNA modification	9	2	0,15	0,0094
27	GO:0006309 apoptotic DNA fragmentation	9	2	0,15	0,0094
28	GO:0048133 male germ-line stem cell asymmetric division	8	2	0,13	0,0074
29	GO:0003407 neural retina development	8	2	0,13	0,0074
30	GO:0007147 female meiosis II	7	2	0,12	0,0332
31	GO:0031118 rRNA pseudouridine synthesis	6	2	0,1	0,004
32	GO:0097039 protein linear polyubiquitination	5	2	0,08	0,0027

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S15. CA female abdomens downregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0050896	response to stimulus	2731	60	52,14
2	GO:0016055	Wnt signaling pathway	138	7	2,63
3	GO:0060446	branching involved in open tracheal system development	77	4	1,47
4	GO:0051168	nuclear export	73	2	1,39
5	GO:0007254	JNK cascade	70	4	1,34
6	GO:0046777	protein autophosphorylation	52	5	0,99
7	GO:0003015	heart process	50	5	0,95
8	GO:0006869	lipid transport	48	4	0,92
9	GO:0048085	adult chitin-containing cuticle pigmentation	44	4	0,84
10	GO:0019991	septate junction assembly	41	3	0,78
11	GO:0030717	oocyte karyosome formation	39	3	0,74
12	GO:0046716	muscle cell cellular homeostasis	37	3	0,71
13	GO:0071805	potassium ion transmembrane transport	26	4	0,5
14	GO:0008101	decapentaplegic signaling pathway	25	3	0,48
15	GO:0006414	translational elongation	21	3	0,4
16	GO:0009649	entrainment of circadian clock	19	3	0,36
17	GO:0008069	dorsal/ventral axis specification, ovarian follicular epithelium	18	3	0,34
18	GO:0060857	establishment of glial blood-brain barrier	18	3	0,34
19	GO:0003333	amino acid transmembrane transport	17	2	0,32
20	GO:0015986	ATP synthesis coupled proton transport	17	2	0,32
21	GO:0002027	regulation of heart rate	15	2	0,29
22	GO:0060361	flight	14	2	0,27
23	GO:0048747	muscle fiber development	14	2	0,27
24	GO:0035309	wing and notum subfield formation	14	2	0,27
25	GO:0030322	stabilization of membrane potential	13	3	0,25
26	GO:0061343	cell adhesion involved in heart morphogenesis	12	2	0,23
27	GO:0006122	mitochondrial electron transport, ubiquinol to cytochrome c	12	2	0,23
28	GO:0045498	sex comb development	12	2	0,23
29	GO:0097105	presynaptic membrane assembly	11	2	0,21
30	GO:0008366	axon ensheathment	10	2	0,19
31	GO:0034446	substrate adhesion-dependent cell spreading	9	2	0,17
32	GO:0035155	negative regulation of terminal cell fate specification, open tracheal system	7	2	0,13

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S16. OR female abdomens downregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0006366 transcription by RNA polymerase II	646	12	10,34	0,0132
2	GO:0006468 protein phosphorylation	417	11	6,68	0,0153
3	GO:0030036 actin cytoskeleton organization	312	11	4,99	0,0306
4	GO:0044770 cell cycle phase transition	171	7	2,74	0,0156
5	GO:0048599 oocyte development	162	4	2,59	0,0205
6	GO:0007297 ovarian follicle cell migration	145	4	2,32	0,0329
7	GO:0043484 regulation of RNA splicing	117	4	1,87	0,0105
8	GO:0016180 snRNA processing	26	2	0,42	0,047
9	GO:0006821 chloride transport	22	2	0,35	0,0477
10	GO:0051310 metaphase plate congression	21	3	0,34	0,0036
11	GO:0007307 eggshell chorion gene amplification	20	2	0,32	0,04
12	GO:0031440 regulation of mRNA 3'-end processing	12	2	0,19	0,0151
13	GO:0001178 regulation of transcriptional start site selection at RNA polymerase II promoter	9	2	0,14	0,0085
14	GO:0031573 intra-S DNA damage checkpoint	9	2	0,14	0,0085
15	GO:0035542 regulation of SNARE complex assembly	8	2	0,13	0,0067
16	GO:0043144 snoRNA processing	8	2	0,13	0,0159
17	GO:0006760 folic acid-containing compound metabolic process	8	2	0,13	0,0316
18	GO:0006513 protein monoubiquitination	7	2	0,11	0,0051
19	GO:0035094 response to nicotine	6	2	0,1	0,0037
20	GO:0045835 negative regulation of meiotic nuclear division	6	2	0,1	0,0037

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S17. AM female abdomens upregulated GO enrichment sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0015833 peptide transport	392	6	6,32	0,03231
2	GO:0008283 cell proliferation	379	10	6,11	0,04272
3	GO:0048737 imaginal disc-derived appendage development	376	12	6,06	0,04581
4	GO:0040008 regulation of growth	345	7	5,56	0,04071
5	GO:0043067 regulation of programmed cell death	241	3	3,89	0,04826
6	GO:0007264 small GTPase mediated signal transduction	202	8	3,26	0,02561
7	GO:0006260 DNA replication	104	4	1,68	0,02054
8	GO:0007173 epidermal growth factor receptor signaling pathway	88	5	1,42	0,00713
9	GO:0002121 inter-male aggressive behaviour	71	5	1,15	0,00563
10	GO:0043547 positive regulation of GTPase activity	67	4	1,08	0,02269
11	GO:0099111 microtubule-based transport	66	3	1,06	0,01601
12	GO:0008586 imaginal disc-derived wing vein morphogenesis	62	7	1	0,00006
13	GO:0046843 dorsal appendage formation	55	5	0,89	0,00185
14	GO:0007443 Malpighian tubule morphogenesis	47	4	0,76	0,02103
15	GO:0070374 positive regulation of ERK1 and ERK2 cascade	42	3	0,68	0,02981
16	GO:0018212 peptidyl-tyrosine modification	36	4	0,58	0,01584
17	GO:0007494 midgut development	34	3	0,55	0,01705
18	GO:0010454 negative regulation of cell fate commitment	31	2	0,5	0,03187
19	GO:0000912 assembly of actomyosin apparatus involved in cytokinesis	29	2	0,47	0,01606
20	GO:0032008 positive regulation of TOR signaling	20	2	0,32	0,04058
21	GO:0007638 mechanosensory behavior	18	2	0,29	0,03337
22	GO:0055070 copper ion homeostasis	16	2	0,26	0,02672
23	GO:0050806 positive regulation of synaptic transmission	15	2	0,24	0,02363
24	GO:0007394 dorsal closure, elongation of leading-edge cells	15	2	0,24	0,02363
25	GO:0030381 chorion-containing eggshell pattern formation	13	2	0,21	0,01793
26	GO:0030433 ubiquitin dependent ERAD pathway	13	2	0,21	0,01793
27	GO:0048172 regulation of short-term neuronal synaptic plasticity	12	3	0,19	0,00081
28	GO:0040015 negative regulation of multicellular organism growth	12	2	0,19	0,01533
29	GO:0050975 sensory perception of touch	11	2	0,18	0,01291
30	GO:0000712 resolution of meiotic recombination intermediates	10	2	0,16	0,01067
31	GO:0061099 negative regulation of protein tyrosine kinase activity	9	2	0,15	0,00863
32	GO:0050974 detection of mechanical stimulus involved in sensory perception	7	2	0,11	0,00514
33	GO:0006313 transposition, DNA-mediated	7	3	0,11	0,01590
34	GO:0000335 negative regulation of transposition, DNA-mediated	6	2	0,1	0,00371
35	GO:0071786 endoplasmic reticulum tubular network organization	5	2	0,08	0,00250

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S18. CA female abdomens upregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0120039	plasma membrane bounded cell projection morphogenesis	650	22	15,65 0,04621
2	GO:0006357	regulation of transcription by RNA polymerase II	596	20	14,35 0,04119
3	GO:0030855	epithelial cell differentiation	432	14	10,4 0,00167
4	GO:0030707	ovarian follicle cell development	346	11	8,33 0,04966
5	GO:0000226	microtubule cytoskeleton organization	326	14	7,85 0,01978
6	GO:0006629	lipid metabolic process	316	10	7,61 0,02140
7	GO:0051493	regulation of cytoskeleton organization	152	8	3,66 0,02682
8	GO:0006836	neurotransmitter transport	129	7	3,11 0,0060
9	GO:0099504	synaptic vesicle cycle	126	7	3,03 0,01437
10	GO:0007619	courtship behaviour	113	3	2,72 0,04989
11	GO:0060179	male mating behaviour	106	3	2,55 0,01486
12	GO:0045471	response to ethanol	84	6	2,02 0,01481
13	GO:0007605	sensory perception of sound	72	5	1,73 0,02932
14	GO:0042067	establishment of ommatidial planar polarity	63	5	1,52 0,04268
15	GO:0051298	centrosome duplication	41	5	0,99 0,02237
16	GO:0016198	axon choice point recognition	35	3	0,84 0,04729
17	GO:0007430	terminal branching, open tracheal system	34	3	0,82 0,04762
18	GO:0010950	positive regulation of endopeptidase activity	28	2	0,67 0,02403
19	GO:0001737	establishment of imaginal disc-derived wing hair orientation	28	3	0,67 0,02893
20	GO:0035050	embryonic heart tube development	26	3	0,63 0,02378
21	GO:0007428	primary branching, open tracheal system	23	3	0,55 0,01707
22	GO:0007187	G protein-coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger	23	3	0,55 0,04722
23	GO:0032509	endosome transport via multivesicular body sorting pathway	21	3	0,51 0,01328
24	GO:0046669	regulation of compound eye retinal cell programmed cell death	21	3	0,51 0,04721
25	GO:0008543	fibroblast growth factor receptor signalling pathway	20	3	0,48 0,01158
26	GO:0008069	dorsal/ventral axis specification, ovarian follicular epithelium	18	3	0,43 0,00859
27	GO:0034587	piRNA metabolic process	15	2	0,36 0,04929
28	GO:0031290	retinal ganglion cell axon guidance	14	4	0,34 0,00027
29	GO:0045143	homologous chromosome segregation	14	3	0,34 0,00805
30	GO:0035076	ecdysone receptor-mediated signaling pathway	14	2	0,34 0,04339
31	GO:0007189	adenylate cyclase-activating G protein-coupled receptor signalling pathway	13	2	0,31 0,03778
32	GO:0007523	larval visceral muscle development	13	2	0,31 0,03778
33	GO:0008333	endosome to lysosome transport	12	2	0,29 0,03247
34	GO:0045571	negative regulation of imaginal disc growth	12	2	0,29 0,03247
35	GO:2000274	regulation of epithelial cell migration, open tracheal system	12	2	0,29 0,03247
36	GO:0045198	establishment of epithelial cell apical/basal polarity	11	2	0,26 0,02749
37	GO:0045186	zonula adherens assembly	11	2	0,26 0,02749

38	GO:0010824	regulation of centrosome duplication	10	2	0,24	0,02398
39	GO:0061099	negative regulation of protein tyrosine kinase activity	9	2	0,22	0,01857
40	GO:0015837	amine transport	8	2	0,19	0,01467
41	GO:0046514	ceramide catabolic process	8	2	0,19	0,01467
42	GO:0035331	negative regulation of hippo signaling	8	2	0,19	0,01467
43	GO:0045705	negative regulation of salivary gland boundary specification	8	2	0,19	0,01467
44	GO:0042135	neurotransmitter catabolic process	7	2	0,17	0,01118
45	GO:0051415	microtubule nucleation by interphase microtubule organizing center	6	2	0,14	0,00811
46	GO:0048730	epidermis morphogenesis	5	2	0,12	0,0055
47	GO:0030497	fatty acid elongation	5	2	0,12	0,0055

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S19. OR female abdomens upregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0032501	multicellular organismal process	3425	58	51,17 0,02736
2	GO:0009889	regulation of biosynthetic process	1207	17	18,03 0,01522
3	GO:0031325	positive regulation of cellular metabolic process	698	16	10,43 0,00062
4	GO:0010604	positive regulation of macromolecule metabolic process	688	14	10,28 0,00721
5	GO:0006508	proteolysis	488	15	7,29 0,01213
6	GO:0007391	dorsal closure	115	5	1,72 0,04940
7	GO:0090174	organelle membrane fusion	73	2	1,09 0,04418
8	GO:0048149	behavioural response to ethanol	62	6	0,93 0,00031
9	GO:0008586	imaginal disc-derived wing vein morphogenesis	62	4	0,93 0,01360
10	GO:0046843	dorsal appendage formation	55	3	0,82 0,04871
11	GO:0007218	neuropeptide signaling pathway	48	3	0,72 0,03460
12	GO:0048085	adult chitin-containing cuticle pigmentation	44	3	0,66 0,02765
13	GO:0046666	retinal cell programmed cell death	32	2	0,48 0,02953
14	GO:1905037	autophagosome organization	32	2	0,48 0,04397
15	GO:0000132	establishment of mitotic spindle orientation	23	2	0,34 0,04567
16	GO:0000027	ribosomal large subunit assembly	22	3	0,33 0,00408
17	GO:0043648	dicarboxylic acid metabolic process	22	2	0,33 0,04211
18	GO:0006821	chloride transport	22	2	0,33 0,04211
19	GO:0007318	pole plasm protein localization	18	2	0,27 0,02899
20	GO:0006096	glycolytic process	17	2	0,25 0,02602
21	GO:0046887	positive regulation of hormone secretion	17	2	0,25 0,02948
22	GO:0035690	cellular response to drug	16	2	0,24 0,02318
23	GO:0045938	positive regulation of circadian sleep/wake cycle	15	2	0,22 0,02048
24	GO:1901616	organic hydroxy compound catabolic process	14	2	0,21 0,04387
25	GO:0043113	receptor clustering	13	2	0,19 0,01551
26	GO:0030322	stabilization of membrane potential	13	2	0,19 0,01551
27	GO:0043112	receptor metabolic process	12	2	0,18 0,01325
28	GO:0048260	positive regulation of receptor-mediated endocytosis	11	2	0,16 0,01115
29	GO:0046845	branched duct epithelial cell fate determination, open tracheal system	9	2	0,13 0,00744
30	GO:0060573	cell fate specification involved in pattern specification	9	2	0,13 0,04385
31	GO:0007438	oenocyte development	8	2	0,12 0,00585
32	GO:1900242	regulation of synaptic vesicle endocytosis	8	2	0,12 0,00585
33	GO:0045823	positive regulation of heart contraction	7	2	0,1 0,00443

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S20. AM female head downregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0015031 protein transport	372	10	6,09	0,0332
2	GO:0007431 salivary gland development	200	7	3,27	0,0107
3	GO:0001933 negative regulation of protein phosphorylation	66	4	1,08	0,0239
4	GO:0042059 negative regulation of epidermal growth factor receptor signalling pathway	39	3	0,64	0,0255
5	GO:0007112 male meiosis cytokinesis	34	3	0,56	0,0177
6	GO:0000132 establishment of mitotic spindle orientation	23	3	0,38	0,006
7	GO:0031122 cytoplasmic microtubule organization	22	2	0,36	0,0496
8	GO:0009649 entrainment of circadian clock	19	2	0,31	0,0379
9	GO:0035332 positive regulation of hippo signaling	17	2	0,28	0,0308
10	GO:0045840 positive regulation of mitotic nuclear division	17	2	0,28	0,0308
11	GO:0045450 bicoid mRNA localization	16	2	0,26	0,0275
12	GO:0000466 maturation of 5,8S rRNA from tricistronic rRNA transcript	14	2	0,23	0,0213
13	GO:0006348 chromatin silencing at telomere	13	2	0,21	0,0184
14	GO:0045880 positive regulation of smoothened signalling pathway	12	2	0,2	0,0158
15	GO:0030709 border follicle cell delamination	11	2	0,18	0,0133
16	GO:0016556 mRNA modification	9	2	0,15	0,0089
17	GO:0007147 female meiosis II	7	2	0,11	0,0322
18	GO:0031118 rRNA pseudouridine synthesis	6	2	0,1	0,0038
19	GO:0032353 negative regulation of hormone biosynthetic process	5	2	0,08	0,0026
20	GO:0035271 ring gland development	5	2	0,08	0,0026

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S21. CA female head downregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0008340 determination of adult lifespan	168	8	3,41	0,02082
2	GO:0042461 photoreceptor cell development	136	6	2,76	0,00572
3	GO:0007298 border follicle cell migration	127	6	2,58	0,04416
4	GO:0008358 maternal determination of anterior/posterior axis, embryo	106	5	2,15	0,00236
5	GO:0032543 mitochondrial translation	98	6	1,99	0,03953
6	GO:0042060 wound healing	91	6	1,85	0,04337
7	GO:0007173 epidermal growth factor receptor signalling pathway	88	8	1,78	0,00012
8	GO:0060446 branching involved in open tracheal system development	77	5	1,56	0,03956
9	GO:0046579 positive regulation of Ras protein signal transduction	73	5	1,48	0,01602
10	GO:0051168 nuclear export	73	2	1,48	0,02033
11	GO:0007254 JNK cascade	70	6	1,42	0,02288
12	GO:0072002 Malpighian tubule development	67	3	1,36	0,03173
13	GO:0046777 protein autophosphorylation	52	5	1,05	0,00631
14	GO:0070374 positive regulation of ERK1 and ERK2 cascade	42	5	0,85	0,0015
15	GO:0008293 torso signaling pathway	38	3	0,77	0,04117
16	GO:0043631 RNA polyadenylation	31	2	0,63	0,04005
17	GO:0007362 terminal region determination	29	3	0,59	0,02034
18	GO:0001941 postsynaptic membrane organization	18	3	0,36	0,01047
19	GO:0015986 ATP synthesis coupled proton transport	17	3	0,34	0,00452
20	GO:0003333 amino acid transmembrane transport	17	2	0,34	0,04554
21	GO:0060361 flight	14	2	0,28	0,0317
22	GO:0030488 tRNA methylation	14	2	0,28	0,0317
23	GO:0035309 wing and notum subfield formation	14	2	0,28	0,0317
24	GO:0016332 establishment or maintenance of polarity of embryonic epithelium	13	2	0,26	0,02753
25	GO:0045498 sex comb development	12	2	0,24	0,02361
26	GO:0045186 zonula adherens assembly	11	2	0,22	0,01994
27	GO:0009631 cold acclimation	9	2	0,18	0,0134
28	GO:0045740 positive regulation of DNA replication	9	2	0,18	0,0134
29	GO:0034446 substrate adhesion-dependent cell spreading	9	2	0,18	0,0134
30	GO:0006750 glutathione biosynthetic process	7	3	0,14	0,00027
31	GO:0006265 DNA topological change	7	2	0,14	0,00803
32	GO:0035155 negative regulation of terminal cell fate specification, open tracheal system	7	2	0,14	0,00803
33	GO:0097688 glutamate receptor clustering	6	2	0,12	0,00581
34	GO:0035193 larval central nervous system remodeling	6	2	0,12	0,00581
35	GO:0030335 positive regulation of cell migration	6	2	0,12	0,00581
36	GO:0071722 detoxification of arsenic-containing substance	5	2	0,1	0,00393

Table S22. OR female head downregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0017145 stem cell division	113	3	1,27	0,0273
2	GO:0048149 behavioral response to ethanol	62	3	0,7	0,0325
3	GO:0042157 lipoprotein metabolic process	55	3	0,62	0,0437
4	GO:0008039 synaptic target recognition	50	4	0,56	0,008
5	GO:0001737 establishment of imaginal disc-derived wing hair orientation	28	2	0,32	0,0393
6	GO:0033301 cell cycle comprising mitosis without cytokinesis	27	2	0,3	0,044
7	GO:0016180 snRNA processing	26	2	0,29	0,0332
8	GO:0006821 chloride transport	22	2	0,25	0,0251
9	GO:0007303 cytoplasmic transport, nurse cell to oocyte	21	2	0,24	0,0229
10	GO:0000387 spliceosomal snRNP assembly	21	2	0,24	0,0229
11	GO:0018345 protein palmitoylation	15	2	0,17	0,012
12	GO:0050806 positive regulation of synaptic transmission	15	2	0,17	0,012
13	GO:0030708 germarium-derived female germ-line cyst encapsulation	13	2	0,15	0,009
14	GO:0035154 terminal cell fate specification, open tracheal system	11	2	0,12	0,0065

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S23. AM female head upregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0045892 negative regulation of transcription, DNA-templated	352	10	5,26	0,00654
2	GO:0006109 regulation of carbohydrate metabolic process	215	4	3,21	0,00325
3	GO:0007298 border follicle cell migration	127	5	1,9	0,04133
4	GO:0002121 inter-male aggressive behaviour	71	4	1,06	0,0214
5	GO:0008586 imaginal disc-derived wing vein morphogenesis	62	6	0,93	0,00031
6	GO:0046843 dorsal appendage formation	55	3	0,82	0,04871
7	GO:0060070 canonical Wnt signaling pathway	39	3	0,58	0,02028
8	GO:0046716 muscle cell cellular homeostasis	37	3	0,55	0,01748
9	GO:0018212 peptidyl-tyrosine modification	36	3	0,54	0,01477
10	GO:0007494 midgut development	34	3	0,51	0,01391
11	GO:0007426 tracheal outgrowth, open tracheal system	31	3	0,46	0,01079
12	GO:0010454 negative regulation of cell fate commitment	31	2	0,46	0,02953
13	GO:0006289 nucleotide-excision repair	30	3	0,45	0,00985
14	GO:0045792 negative regulation of cell size	24	2	0,36	0,04935
15	GO:0048072 compound eye pigmentation	22	3	0,33	0,02021
16	GO:0007179 transforming growth factor beta receptor signalling pathway	22	2	0,33	0,04211
17	GO:1901800 positive regulation of proteasomal protein catabolic process	20	2	0,3	0,0439
18	GO:0055070 copper ion homeostasis	16	2	0,24	0,02318
19	GO:0007394 dorsal closure, elongation of leading edge cells	15	2	0,22	0,02048
20	GO:0007080 mitotic metaphase plate congression	15	2	0,22	0,02048
21	GO:0016339 calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules	14	2	0,21	0,01792
22	GO:0008057 eye pigment granule organization	14	2	0,21	0,01792
23	GO:0035099 hemocyte migration	14	2	0,21	0,01792
24	GO:0007523 larval visceral muscle development	13	2	0,19	0,01551
25	GO:0048172 regulation of short-term neuronal synaptic plasticity	12	2	0,18	0,01325
26	GO:0044331 cell-cell adhesion mediated by cadherin	11	2	0,16	0,01115
27	GO:0051926 negative regulation of calcium ion transport	10	2	0,15	0,00921
28	GO:0042595 behavioural response to starvation	9	2	0,13	0,00744
29	GO:0035001 dorsal trunk growth, open tracheal system	9	2	0,13	0,00744
30	GO:0061099 negative regulation of protein tyrosine kinase activity	9	2	0,13	0,00744
31	GO:0045737 positive regulation of cyclin-dependent protein serine/threonine kinase activity	8	2	0,12	0,00585
32	GO:1901409 positive regulation of phosphorylation of RNA polymerase II C-terminal domain	7	2	0,1	0,00443

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S24. CA female head upregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0000902 cell morphogenesis	766	20	14,26	0,0291
2	GO:0006979 response to oxidative stress	110	6	2,05	0,0351
3	GO:0010721 negative regulation of cell development	94	4	1,75	0,0169
4	GO:0007274 neuromuscular synaptic transmission	89	6	1,66	0,0046
5	GO:0007427 epithelial cell migration, open tracheal system	45	4	0,84	0,0227
6	GO:0048085 adult chitin-containing cuticle pigmentation	44	4	0,82	0,0088
7	GO:0019991 septate junction assembly	41	3	0,76	0,0403
8	GO:0042059 negative regulation of epidermal growth factor receptor signalling pathway	39	3	0,73	0,0355
9	GO:0055090 acylglycerol homeostasis	29	3	0,54	0,0184
10	GO:0008543 fibroblast growth factor receptor signalling pathway	20	3	0,37	0,0199
11	GO:0030514 negative regulation of BMP signaling pathway	19	2	0,35	0,0479
12	GO:0008069 dorsal/ventral axis specification, ovarian follicular epithelium	18	3	0,34	0,0042
13	GO:2000242 negative regulation of reproductive process	17	2	0,32	0,039
14	GO:0007031 peroxisome organization	17	2	0,32	0,039
15	GO:0045478 fusome organization	16	2	0,3	0,0348
16	GO:0034587 piRNA metabolic process	15	2	0,28	0,0309
17	GO:0035076 ecdysone receptor-mediated signaling pathway	14	3	0,26	0,002
18	GO:0031290 retinal ganglion cell axon guidance	14	3	0,26	0,002
19	GO:0007189 adenylate cyclase-activating G protein-coupled receptor signalling pathway	13	2	0,24	0,0235
20	GO:0008333 endosome to lysosome transport	12	2	0,22	0,0201
21	GO:0045571 negative regulation of imaginal disc growth	12	2	0,22	0,0201
22	GO:0045186 zonula adherens assembly	11	2	0,2	0,017
23	GO:0033619 membrane protein proteolysis	10	2	0,19	0,0185
24	GO:0010824 regulation of centrosome duplication	10	2	0,19	0,0185
25	GO:0045199 maintenance of epithelial cell apical/basal polarity	9	2	0,17	0,0114
26	GO:0030497 fatty acid elongation	5	2	0,09	0,0033
27	GO:0045761 regulation of adenylate cyclase activity	5	2	0,09	0,0033

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S25. OR female head upregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0009889 regulation of biosynthetic process	1207	18	17,46	0,0145
2	GO:0032774 RNA biosynthetic process	1095	18	15,84	0,0423
3	GO:0031325 positive regulation of cellular metabolic process	698	12	10,1	0,0424
4	GO:0007591 molting cycle, chitin-based cuticle	95	5	1,37	0,0115
5	GO:0090174 organelle membrane fusion	73	3	1,06	0,0424
6	GO:0016246 RNA interference	36	3	0,52	0,0426
7	GO:0035186 syncytial blastoderm mitotic cell cycle	23	2	0,33	0,0431
8	GO:0006584 catecholamine metabolic process	19	2	0,27	0,0302
9	GO:0046887 positive regulation of hormone secretion	17	2	0,25	0,0285
10	GO:0032483 regulation of Rab protein signal transduction	14	2	0,2	0,0169
11	GO:0031000 response to caffeine	12	2	0,17	0,0125
12	GO:0006744 ubiquinone biosynthetic process	11	2	0,16	0,0105
13	GO:0006312 mitotic recombination	10	2	0,14	0,0087
14	GO:1900242 regulation of synaptic vesicle endocytosis	8	2	0,12	0,0055

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S26. AM male abdomens downregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0008152 metabolic process	4404	99	92,44	0,0159
2	GO:0048489 synaptic vesicle transport	90	4	1,89	0,0482
3	GO:0030010 establishment of cell polarity	66	5	1,39	0,0411
4	GO:0045185 maintenance of protein location	47	3	0,99	0,0143
5	GO:0042059 negative regulation of epidermal growth factor receptor signaling pathway	39	3	0,82	0,0479
6	GO:0036335 intestinal stem cell homeostasis	36	3	0,76	0,0391
7	GO:0007112 male meiosis cytokinesis	34	4	0,71	0,0053
8	GO:0045840 positive regulation of mitotic nuclear division	17	2	0,36	0,0485
9	GO:0045450 bicoid mRNA localization	16	2	0,34	0,0433
10	GO:0034587 piRNA metabolic process	15	3	0,31	0,0034
11	GO:0000478 endonucleolytic cleavage involved in rRNA processing	12	2	0,25	0,0209
12	GO:0045880 positive regulation of smoothened signalling pathway	12	2	0,25	0,0252
13	GO:0030709 border follicle cell delamination	11	2	0,23	0,0213
14	GO:0031445 regulation of heterochromatin assembly	10	2	0,21	0,0413
15	GO:0016556 mRNA modification	9	2	0,19	0,0143
16	GO:0008335 female germline ring canal stabilization	7	2	0,15	0,0086
17	GO:0007147 female meiosis II	7	2	0,15	0,0413
18	GO:0032353 negative regulation of hormone biosynthetic process	5	2	0,1	0,0042
19	GO:0097039 protein linear polyubiquitination	5	2	0,1	0,0042

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S27. CA male abdomens downregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0050896 response to stimulus	2731	67	60,56	0,0212
2	GO:0009790 embryo development	586	23	12,99	0,0297
3	GO:0046578 regulation of Ras protein signal transduction	139	9	3,08	0,0267
4	GO:0042461 photoreceptor cell development	136	9	3,02	0,0066
5	GO:0007173 epidermal growth factor receptor signalling pathway	88	5	1,95	0,0171
6	GO:0001763 morphogenesis of a branching structure	79	4	1,75	0,0436
7	GO:0022904 respiratory electron transport chain	73	3	1,62	0,016
8	GO:0046579 positive regulation of Ras protein signal transduction	73	5	1,62	0,0275
9	GO:0046552 photoreceptor cell fate commitment	68	5	1,51	0,0219
10	GO:0072002 Malpighian tubule development	67	3	1,49	0,0374
11	GO:0042706 eye photoreceptor cell fate commitment	65	4	1,44	0,022
12	GO:0055059 asymmetric neuroblast division	53	4	1,18	0,033
13	GO:0003015 heart process	50	4	1,11	0,0321
14	GO:0007218 neuropeptide signaling pathway	48	4	1,06	0,0213
15	GO:0070374 positive regulation of ERK1 and ERK2 cascade	42	4	0,93	0,0136
16	GO:0007354 zygotic determination of anterior/poster axis, embryo	38	5	0,84	0,0155
17	GO:0046716 muscle cell cellular homeostasis	37	3	0,82	0,0481
18	GO:0045924 regulation of female receptivity	35	3	0,78	0,0476
19	GO:0007362 terminal region determination	29	3	0,64	0,0257
20	GO:0008594 photoreceptor cell morphogenesis	21	3	0,47	0,0475
21	GO:0008069 dorsal/ventral axis specification, ovarian follicular epithelium	18	3	0,4	0,0068
22	GO:0001941 postsynaptic membrane organization	18	3	0,4	0,0124
23	GO:0003333 amino acid transmembrane transport	17	3	0,38	0,0058
24	GO:0045450 bicoid mRNA localization	16	3	0,35	0,0049
25	GO:0008299 isoprenoid biosynthetic process	16	2	0,35	0,0479
26	GO:0060361 flight	14	2	0,31	0,0374
27	GO:0048747 muscle fiber development	14	2	0,31	0,0374
28	GO:0035309 wing and notum subfield formation	14	2	0,31	0,0374
29	GO:0032392 DNA geometric change	11	2	0,24	0,0221
30	GO:0006835 dicarboxylic acid transport	11	2	0,24	0,0236
31	GO:0015800 acidic amino acid transport	10	2	0,22	0,0196
32	GO:0070584 mitochondrion morphogenesis	10	2	0,22	0,0196
33	GO:0006491 N-glycan processing	10	2	0,22	0,0196
34	GO:0006941 striated muscle contraction	10	2	0,22	0,0196
35	GO:0035288 anterior head segmentation	9	2	0,2	0,0159
36	GO:0045740 positive regulation of DNA replication	9	2	0,2	0,0159
37	GO:0034446 substrate adhesion-dependent cell spreading	9	2	0,2	0,0159
38	GO:0110117 positive regulation of compound eye photoreceptor cell differentiation	9	2	0,2	0,0437
39	GO:0097688 glutamate receptor clustering	6	2	0,13	0,0069
40	GO:0035193 larval central nervous system remodeling	6	2	0,13	0,0069
41	GO:0051415 microtubule nucleation by interphase microtubule organizing center	6	2	0,13	0,0069

Table S28. OR male abdomens downregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0008104 protein localization	602	11	8,35	0,02995
2	GO:0043604 amide biosynthetic process	462	9	6,41	0,00426
3	GO:0030036 actin cytoskeleton organization	312	9	4,33	0,02713
4	GO:0006403 RNA localization	189	5	2,62	0,02722
5	GO:0044770 cell cycle phase transition	171	3	2,37	0,01392
6	GO:0048599 oocyte development	162	5	2,25	0,00085
7	GO:0017145 stem cell division	113	3	1,57	0,0401
8	GO:0048588 developmental cell growth	86	3	1,19	0,03077
9	GO:0031123 RNA 3'-end processing	77	5	1,07	0,04001
10	GO:0007602 phototransduction	64	4	0,89	0,01415
11	GO:0045746 negative regulation of Notch signaling pathway	51	3	0,71	0,03348
12	GO:0008652 cellular amino acid biosynthetic process	35	2	0,49	0,02743
13	GO:0016180 snRNA processing	26	2	0,36	0,04082
14	GO:0097352 autophagosome maturation	25	2	0,35	0,04653
15	GO:0090502 RNA phosphodiester bond hydrolysis, endonucleolytic	23	4	0,32	0,00938
16	GO:0045167 asymmetric protein localization involved in cell fate determination	22	2	0,31	0,0368
17	GO:0006821 chloride transport	22	2	0,31	0,0368
18	GO:0007303 cytoplasmic transport, nurse cell to oocyte	21	3	0,29	0,00288
19	GO:0000387 spliceosomal snRNP assembly	21	2	0,29	0,03375
20	GO:0042398 cellular modified amino acid biosynthetic process	14	3	0,19	0,00085
21	GO:0045175 basal protein localization	13	2	0,18	0,01347
22	GO:0030708 germarium-derived female germ-line cyst encapsulation	13	2	0,18	0,01347
23	GO:0031440 regulation of mRNA 3'-end processing	12	2	0,17	0,0115
24	GO:0000479 endonucleolytic cleavage of tricistronic rRNA transcript	11	2	0,15	0,00967
25	GO:0007295 growth of a germarium-derived egg chamber	11	2	0,15	0,00967
26	GO:0001682 tRNA 5'-leader removal	10	2	0,14	0,00799
27	GO:0042559 pteridine-containing compound biosynthetic process	9	3	0,12	0,00021
28	GO:0000463 maturation of LSU-rRNA from tricistronic rRNA transcript	9	2	0,12	0,00645
29	GO:0001178 regulation of transcriptional start site selection at RNA polymerase II promoter	9	2	0,12	0,00645
30	GO:0006760 folic acid-containing compound metabolic process	8	4	0,11	0,00018
31	GO:0046653 tetrahydrofolate metabolic process	6	2	0,08	0,00276
32	GO:0009113 purine nucleobase biosynthetic process	5	2	0,07	0,00186

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S29. AM male abdomens upregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0015833 peptide transport	392	5	6,79	0,035
2	GO:0055080 cation homeostasis	107	3	1,85	0,01729
3	GO:0006260 DNA replication	104	4	1,8	0,0062
4	GO:0051236 establishment of RNA localization	104	4	1,8	0,03404
5	GO:0007274 neuromuscular synaptic transmission	89	5	1,54	0,0186
6	GO:0007173 epidermal growth factor receptor signalling pathway	88	5	1,52	0,00869
7	GO:0051403 stress-activated MAPK cascade	73	2	1,26	0,01734
8	GO:0008586 imaginal disc-derived wing vein morphogenesis	62	6	1,07	0,00067
9	GO:0015711 organic anion transport	62	4	1,07	0,03741
10	GO:0000724 double-strand break repair via homologous recombination	46	4	0,8	0,00801
11	GO:0008407 chaeta morphogenesis	45	3	0,78	0,04252
12	GO:0046716 muscle cell cellular homeostasis	37	3	0,64	0,02571
13	GO:0018212 peptidyl-tyrosine modification	36	3	0,62	0,01715
14	GO:0016476 regulation of embryonic cell shape	35	3	0,61	0,04572
15	GO:0007494 midgut development	34	3	0,59	0,02056
16	GO:0006289 nucleotide-excision repair	30	3	0,52	0,01466
17	GO:0000912 assembly of actomyosin apparatus involved in cytokinesis	29	3	0,5	0,01713
18	GO:0032008 positive regulation of TOR signaling	20	2	0,35	0,04614
19	GO:0015893 drug transport	19	2	0,33	0,04199
20	GO:0008103 oocyte microtubule cytoskeleton polarization	18	3	0,31	0,00343
21	GO:0050806 positive regulation of synaptic transmission	15	3	0,26	0,00199
22	GO:0006298 mismatch repair	15	2	0,26	0,02697
23	GO:0030433 ubiquitin-dependent ERAD pathway	13	2	0,23	0,02049
24	GO:0048172 regulation of short-term neuronal synaptic plasticity	12	3	0,21	0,001
25	GO:0043488 regulation of mRNA stability	12	2	0,21	0,01754
26	GO:0000712 resolution of meiotic recombination intermediates	10	3	0,17	0,00056
27	GO:0006123 mitochondrial electron transport, cytochrome c to oxygen	10	2	0,17	0,01223
28	GO:0035001 dorsal trunk growth, open tracheal system	9	2	0,16	0,0099
29	GO:0061099 negative regulation of protein tyrosine kinase activity	9	2	0,16	0,0099
30	GO:0036297 interstrand cross-link repair	8	2	0,14	0,00778
31	GO:0006313 transposition, DNA-mediated	7	2	0,12	0,01721

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S30. CA male abdomens upregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0051493 regulation of cytoskeleton organization	152	7	3,21	0,021
2	GO:0007274 neuromuscular synaptic transmission	89	6	1,88	0,0078
3	GO:0018393 internal peptidyl-lysine acetylation	84	3	1,77	0,0211
4	GO:0007605 sensory perception of sound	72	5	1,52	0,0178
5	GO:0003015 heart process	50	3	1,06	0,0296
6	GO:0016198 axon choice point recognition	35	2	0,74	0,0417
7	GO:0055090 acylglycerol homeostasis	29	3	0,61	0,0209
8	GO:0016545 male courtship behaviour, veined wing vibration	29	2	0,61	0,0211
9	GO:1905515 non-motile cilium assembly	20	3	0,42	0,0081
10	GO:0008069 dorsal/ventral axis specification, ovarian follicular epithelium	18	3	0,38	0,006
11	GO:0007031 peroxisome organization	17	2	0,36	0,049
12	GO:0035332 positive regulation of hippo signaling	17	2	0,36	0,049
13	GO:0035076 ecdysone receptor-mediated signaling pathway	14	3	0,3	0,0028
14	GO:0016339 calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules	14	2	0,3	0,0341
15	GO:0045143 homologous chromosome segregation	14	2	0,3	0,0341
16	GO:0031290 retinal ganglion cell axon guidance	14	2	0,3	0,0341
17	GO:0022409 positive regulation of cell-cell adhesion	13	2	0,27	0,0297
18	GO:0005980 glycogen catabolic process	13	2	0,27	0,0416
19	GO:0008333 endosome to lysosome transport	12	2	0,25	0,0254
20	GO:0045571 negative regulation of imaginal disc growth	12	2	0,25	0,0254
21	GO:0044331 cell-cell adhesion mediated by cadherin	11	2	0,23	0,0215
22	GO:0035293 chitin-based larval cuticle pattern formation	11	2	0,23	0,0215
23	GO:0045186 zonula adherens assembly	11	2	0,23	0,0215
24	GO:0005978 glycogen biosynthetic process	9	2	0,19	0,0145
25	GO:0061099 negative regulation of protein tyrosine kinase activity	9	2	0,19	0,0145
26	GO:0007210 serotonin receptor signaling pathway	6	2	0,13	0,0063
27	GO:0030497 fatty acid elongation	5	2	0,11	0,0042

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S31. OR male abdomens upregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0009889 regulation of biosynthetic process	1207	28	19,18	0,01481
2	GO:0031325 positive regulation of cellular metabolic process	698	21	11,09	0,04356
3	GO:0045893 positive regulation of transcription, DNA-templated	389	14	6,18	0,02082
4	GO:0007622 rhythmic behaviour	119	4	1,89	0,04652
5	GO:0090174 organelle membrane fusion	73	2	1,16	0,04696
6	GO:0051225 spindle assembly	72	5	1,14	0,0495
7	GO:0048149 behavioural response to ethanol	62	6	0,99	0,00043
8	GO:0016476 regulation of embryonic cell shape	35	3	0,56	0,0391
9	GO:0040001 establishment of mitotic spindle localization	33	3	0,52	0,01028
10	GO:0006821 chloride transport	22	2	0,35	0,04706
11	GO:0000027 ribosomal large subunit assembly	22	2	0,35	0,04706
12	GO:0003383 apical constriction	21	2	0,33	0,04322
13	GO:0031058 positive regulation of histone modification	19	2	0,3	0,03592
14	GO:0046887 positive regulation of hormone secretion	17	3	0,27	0,00025
15	GO:0016575 histone deacetylation	15	2	0,24	0,02298
16	GO:0040034 regulation of development, heterochronic	13	2	0,21	0,01743
17	GO:0000022 mitotic spindle elongation	11	2	0,17	0,01255
18	GO:0045737 positive regulation of cyclin-dependent protein serine/threonine kinase activity	8	2	0,13	0,00659

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S32. AM male heads downregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0048699 generation of neurons	1018	14	17,62	0,0179
2	GO:0009408 response to heat	69	4	1,19	0,0313
3	GO:0001933 negative regulation of protein phosphorylation	66	4	1,14	0,0266
4	GO:0042059 negative regulation of epidermal growth factor receptor signalling pathway	39	3	0,68	0,0295
5	GO:0044706 multi-monicellular organism process	29	2	0,5	0,0173
6	GO:0000132 establishment of mitotic spindle orientation	23	3	0,4	0,007
7	GO:0045840 positive regulation of mitotic nuclear division	17	2	0,29	0,0342
8	GO:0000466 maturation of 5,8S rRNA from tricistronic rRNA transcript	14	2	0,24	0,0236
9	GO:0000478 endonucleolytic cleavage involved in rRNA processing	12	2	0,21	0,0172
10	GO:0045880 positive regulation of smoothened signalling pathway	12	2	0,21	0,0175
11	GO:0016556 mRNA modification	9	2	0,16	0,0099
12	GO:0031118 rRNA pseudouridine synthesis	6	2	0,1	0,0043
13	GO:0032353 negative regulation of hormone biosynthetic process	5	2	0,09	0,0029
14	GO:0097039 protein linear polyubiquitination	5	2	0,09	0,0029

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S33. CA male heads downregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0050896	response to stimulus	2731	71	51,17 0,01543
2	GO:0009790	embryo development	586	18	10,98 0,041
3	GO:0010628	positive regulation of gene expression	465	10	8,71 0,03429
4	GO:0008340	determination of adult lifespan	168	9	3,15 0,00424
5	GO:0042461	photoreceptor cell development	136	5	2,55 0,00494
6	GO:0007619	courtship behaviour	113	6	2,12 0,03042
7	GO:0008358	maternal determination of anterior/posterior axis, embryo	106	5	1,99 0,00201
8	GO:0032543	mitochondrial translation	98	7	1,84 0,00764
9	GO:0042060	wound healing	91	6	1,7 0,03383
10	GO:0007173	epidermal growth factor receptor signalling pathway	88	8	1,65 0,00008
11	GO:0002168	instar larval development	74	5	1,39 0,04513
12	GO:0046579	positive regulation of Ras protein signal transduction	73	7	1,37 0,00041
13	GO:0051168	nuclear export	73	3	1,37 0,01866
14	GO:0007254	JNK cascade	70	5	1,31 0,01989
15	GO:0070374	positive regulation of ERK1 and ERK2 cascade	42	5	0,79 0,00105
16	GO:0008293	torso signaling pathway	38	3	0,71 0,03374
17	GO:0046716	muscle cell cellular homeostasis	37	3	0,69 0,0315
18	GO:0000462	maturational SSU-rRNA from tricistronic rRNA transcript	33	3	0,62 0,02334
19	GO:0007362	terminal region determination	29	3	0,54 0,0165
20	GO:0045433	male courtship behaviour, veined wing generated song production	28	3	0,52 0,015
21	GO:0030641	regulation of cellular pH	22	2	0,41 0,01866
22	GO:0048803	imaginal disc-derived male genitalia morphogenesis	19	2	0,36 0,04843
23	GO:0008069	dorsal/ventral axis specification, ovarian follicular epithelium	18	3	0,34 0,00428
24	GO:0001941	postsynaptic membrane organization	18	2	0,34 0,04386
25	GO:0015986	ATP synthesis coupled proton transport	17	3	0,32 0,00362
26	GO:0003333	amino acid transmembrane transport	17	2	0,32 0,03946
27	GO:0030968	endoplasmic reticulum unfolded protein response	16	2	0,3 0,03524
28	GO:0035073	pupariation	15	2	0,28 0,03122
29	GO:0002027	regulation of heart rate	15	2	0,28 0,03122
30	GO:0042398	cellular modified amino acid biosynthetic process	14	4	0,26 0,00673
31	GO:0048747	muscle fiber development	14	2	0,26 0,02739
32	GO:0030488	tRNA methylation	14	2	0,26 0,02739
33	GO:0035309	wing and notum subfield formation	14	2	0,26 0,02739
34	GO:0016226	iron-sulfur cluster assembly	13	2	0,24 0,02376
35	GO:0098586	cellular response to virus	10	2	0,19 0,01422
36	GO:0006491	N-glycan processing	10	2	0,19 0,01422
37	GO:0060148	positive regulation of posttranscriptional gene silencing	10	2	0,19 0,01422
38	GO:0006941	striated muscle contraction	10	2	0,19 0,01422
39	GO:0006879	cellular iron ion homeostasis	9	2	0,17 0,01152

40	GO:0045740	positive regulation of DNA replication	9	2	0,17	0,01152
41	GO:0034126	positive regulation of MyD88-dependent toll-like receptor signalling pathway	8	2	0,15	0,00907
42	GO:0006750	glutathione biosynthetic process	7	2	0,13	0,00689
43	GO:0045823	positive regulation of heart contraction	7	2	0,13	0,00689
44	GO:0097688	glutamate receptor clustering	6	2	0,11	0,00498
45	GO:0035193	larval central nervous system remodeling	6	2	0,11	0,00498
46	GO:0030335	positive regulation of cell migration	6	2	0,11	0,00498
47	GO:1903078	positive regulation of protein localization to plasma membrane	6	2	0,11	0,00498

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S34. OR male heads downregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0048149 behavioral response to ethanol	62	3	0,76	0,0398
2	GO:0033301 cell cycle comprising mitosis without cytokinesis	27	2	0,33	0,0476
3	GO:0006821 chloride transport	22	2	0,27	0,0291
4	GO:0007303 cytoplasmic transport, nurse cell to oocyte	21	2	0,26	0,0267
5	GO:0007479 leg disc proximal/distal pattern formation	18	2	0,22	0,0199
6	GO:0042073 intraciliary transport	18	2	0,22	0,0199
7	GO:0050806 positive regulation of synaptic transmission	15	2	0,18	0,014
8	GO:0030708 germarium-derived female germ-line cyst encapsulation	13	2	0,16	0,0106
9	GO:0045572 positive regulation of imaginal disc growth	11	2	0,13	0,0076
10	GO:0006513 protein monoubiquitination	7	2	0,09	0,003

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S35. AM male heads upregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0031325 positive regulation of cellular metabolic process	698	12	10,59	0,04468
2	GO:0009617 response to bacterium	213	4	3,23	0,03348
3	GO:0007219 Notch signaling pathway	160	4	2,43	0,01599
4	GO:0007298 border follicle cell migration	127	5	1,93	0,04374
5	GO:0030718 germ-line stem cell population maintenance	74	4	1,12	0,02577
6	GO:0002121 inter-male aggressive behaviour	71	5	1,08	0,00436
7	GO:0099111 microtubule-based transport	66	3	1	0,01506
8	GO:0008586 imaginal disc-derived wing vein morphogenesis	62	6	0,94	0,00033
9	GO:0048190 wing disc dorsal/ventral pattern formation	56	4	0,85	0,01012
10	GO:0046843 dorsal appendage formation	55	4	0,83	0,00951
11	GO:0008587 imaginal disc-derived wing margin morphogenesis	54	4	0,82	0,01318
12	GO:0018212 peptidyl-tyrosine modification	36	3	0,55	0,015
13	GO:0007494 midgut development	34	3	0,52	0,01451
14	GO:0016199 axon midline choice point recognition	33	3	0,5	0,01337
15	GO:0007426 tracheal outgrowth, open tracheal system	31	3	0,47	0,01126
16	GO:0006289 nucleotide-excision repair	30	3	0,46	0,01028
17	GO:0031532 actin cytoskeleton reorganization	29	3	0,44	0,04639
18	GO:0033627 cell adhesion mediated by integrin	23	2	0,35	0,04699
19	GO:0048072 compound eye pigmentation	22	2	0,33	0,04332
20	GO:1901800 positive regulation of proteasomal protein catabolic process	20	2	0,3	0,0446
21	GO:0015893 drug transport	19	2	0,29	0,03303
22	GO:0055070 copper ion homeostasis	16	2	0,24	0,02387
23	GO:0016339 calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules	14	2	0,21	0,01846
24	GO:0008057 eye pigment granule organization	14	2	0,21	0,01846
25	GO:0035099 hemocyte migration	14	2	0,21	0,01846
26	GO:0030720 oocyte localization involved in germarium-derived egg chamber formation	14	2	0,21	0,01846
27	GO:0048172 regulation of short-term neuronal synaptic plasticity	12	2	0,18	0,01366
28	GO:0042595 behavioural response to starvation	9	2	0,14	0,00767
29	GO:0035001 dorsal trunk growth, open tracheal system	9	2	0,14	0,00767
30	GO:0061099 negative regulation of protein tyrosine kinase activity	9	2	0,14	0,00767
31	GO:0071786 endoplasmic reticulum tubular network organization	5	2	0,08	0,00222

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S36. CA male heads upregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0000902	cell morphogenesis	766	25	14,99
2	GO:0120039	plasma membrane bounded cell projection morphogenesis	650	22	12,72
3	GO:0048812	neuron projection morphogenesis	648	21	12,68
4	GO:0030707	ovarian follicle cell development	346	14	6,77
5	GO:0000398	mRNA splicing, via spliceosome	273	10	5,34
6	GO:0007274	neuromuscular synaptic transmission	89	6	1,74
7	GO:0048675	axon extension	65	4	1,27
8	GO:0031935	regulation of chromatin silencing	51	4	1
9	GO:0007427	epithelial cell migration, open tracheal system	45	4	0,88
10	GO:0030717	oocyte karyosome formation	39	3	0,76
11	GO:0007428	primary branching, open tracheal system	23	3	0,45
12	GO:0008543	fibroblast growth factor receptor signalling pathway	20	3	0,39
13	GO:0008069	dorsal/ventral axis specification, ovarian follicular epithelium	18	3	0,35
14	GO:0007032	endosome organization	18	2	0,35
15	GO:0007031	peroxisome organization	17	2	0,33
16	GO:0045478	fusome organization	16	2	0,31
17	GO:0034587	piRNA metabolic process	15	2	0,29
18	GO:0035076	ecdysone receptor-mediated signaling pathway	14	3	0,27
19	GO:0031290	retinal ganglion cell axon guidance	14	3	0,27
20	GO:0008333	endosome to lysosome transport	12	2	0,23
21	GO:0045571	negative regulation of imaginal disc growth	12	2	0,23
22	GO:0045186	zonula adherens assembly	11	2	0,22
23	GO:0090278	negative regulation of peptide hormone secretion	9	2	0,18
24	GO:0051262	protein tetramerization	9	2	0,18
25	GO:0030497	fatty acid elongation	5	2	0,1
26	GO:0045761	regulation of adenylate cyclase activity	5	2	0,1

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value

Table S37. OR male heads upregulated GO enrichment, sorted by number of times a GO term appears in the reference universe, ascending order.

GO.ID	Term	*Annot	*Sign	*Exp	*Fisher
1	GO:0009889 regulation of biosynthetic process	1207	15	16,6	0,0141
2	GO:0032774 RNA biosynthetic process	1095	14	15,06	0,0415
3	GO:0031325 positive regulation of cellular metabolic process	698	14	9,6	0,0394
4	GO:0090174 organelle membrane fusion	73	4	1	0,04
5	GO:0048149 behavioral response to ethanol	62	6	0,85	0,0002
6	GO:0046887 positive regulation of hormone secretion	17	2	0,23	0,0271
7	GO:0032483 regulation of Rab protein signal transduction	14	2	0,19	0,0153
8	GO:0043113 receptor clustering	13	2	0,18	0,0133
9	GO:0007488 histoblast morphogenesis	11	2	0,15	0,0095
10	GO:0009631 cold acclimation	9	2	0,12	0,0063
11	GO:0045737 positive regulation of cyclin-dependent protein serine/threonine kinase activity	8	2	0,11	0,005
12	GO:1900242 regulation of synaptic vesicle endocytosis	8	2	0,11	0,005

*Annot: Number of times a GO term appears in the reference universe. Sign: Number of DE clusters which are annotated with the GO term. Exp: Number of times a GO term would be expected to appear in the DE cluster dataset. *Fisher: classic Fisher value