nature ecology & evolution

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41559-019-1054-4

In the format provided by the authors and unedited.

Co-option of wing-patterning genes underlies the evolution of the treehopper helmet

Cera R. Fisher*, Jill L. Wegrzyn and Elizabeth L. Jockusch

Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, USA. *e-mail: cera.fisher@uconn.edu

The treehopper helmet, a morphological novelty, evolved via co-option of wing-patterning genes

C. R. Fisher[1,2]*, Jill L. Wegrzyn[1,3], Elizabeth L. Jockusch[1]

[1]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs CT USA

[2]ORCID ID: 0000-0001-7449-9076

[3]ORCID ID: 0000-0001-5923-0888

*corresponding author at cera.fisher@uconn.edu

Supplemental information

Supplementary Methods

1. Scaling TPM between two species with different numbers of annotated transcripts

based on Musser & Wagner (2015), JEZ:B

TPM, transcripts per million mapped, is a way of normalizing RNAseq data that accounts for differences in library size by scaling the abundance of a transcript (the "counts") to the total number of transcripts assumed to be present in the transcriptome. In short, TPM = count for transcript "i" / total number of annotated transcripts * a scaling factor.

Necessarily, TPM from one species does not map to the TPM of another species if their transcriptomes are of different sizes, which they almost certainly are, so the values for the species with the smaller transcriptome will be inflated relative to the species with the larger transcriptome. According to Musser & Wagner[1], these values can be rescaled by calculating a scaling factor α:

$$\alpha = 10^{-6} \sum_{j=1}^{N1} tpm(Aj)$$

Where N1 = the number of transcripts for species B (the smaller set), *j* = all the transcripts in the set, and tpm(Aj) is the transcripts per million for species A for each of the genes *j* in the set.

Here is how I interpret this to work in my transcriptomes for *Entylia carinata* and *Homalodisca vitripennis*.

1) Read in the un-normalized TPM values from RSEM/edgeR.

```
library("dplyr")
options(stringsAsFactors = FALSE)
Ecar.TPM <- read.table("ECEF_Ref_Filt.gene.TMM.EXPR.matrix", sep="\t", header
=TRUE)
colnames(Ecar.TPM)
```

```
##  [1] "X"            "ECEF_Abd"     "ECEF_Eye"     "ECEF_Leg"
##  [5] "ECEF_Meso"    "ECEF_Ovi"     "ECEF_Pro"     "ECEF_Wing2"
##  [9] "ECEF_Wing3"   "ECFisC_Abd"   "ECFisC_Eye"   "ECFisC_Leg"
## [13] "ECFisC_Meso"  "ECFisC_Ovi"   "ECFisC_Pro"   "ECFisC_Wing2"
## [17] "ECFisC_Wing3" "ECty4_Abd"    "ECty4_Eye"    "ECty4_Leg"
## [21] "ECty4_Meso"   "ECty4_Ovi"    "ECty4_Pro"    "ECty4_Wing2"
## [25] "ECty4_Wing3"
```

```
colnames(Ecar.TPM)[1] <- "ECid"
```

```
Hvit.TPM <- read.table("HV102_Ref_Filt.gene.TMM.EXPR.matrix", sep="\t", heade
r=TRUE)
colnames(Hvit.TPM)
```

```
##  [1] "X"           "HV102_Abd"   "HV102_Eye"   "HV102_Leg"   "HV102_Meso"
##  [6] "HV102_Ovi"   "HV102_Pro"   "HV102_Wing2" "HV102_Wing3" "HV3a_Abd"
## [11] "HV3a_Eye"    "HV3a_Leg"    "HV3a_Meso"   "HV3a_Ovi"    "HV3a_Pro"
## [16] "HV3a_Wing2"  "HV3a_Wing3"  "HV7_Abd"     "HV7_Leg"     "HV7_Meso"
## [21] "HV7_Pro"     "HV7_Wing2"
```

```
colnames(Hvit.TPM)[1] <- "HVid"
```

2) Calculate α

```
HvitN1 <- as.numeric(length(Hvit.TPM$HVid))
                                                # 17,607
EcarN2 <- as.numeric(length(Ecar.TPM$ECid))
                                                # 18,652
sumN2.EcarTPM <- sum(Ecar.TPM[,c(2:25)])/24
sumN2.EcarTPM
```

```
## [1] 1008495
```

```
## Sum of TPM for any given sample should, by definition, be ~1,000,000
# The average TPM for Ecar is the sum divided by the number of transcripts.
Ec.avg.TPM <- sumN2.EcarTPM/EcarN2

## Getting the sum of Ecar TPM for the # of transcripts in Hvit's transcripto
me
## i.e, multiplying the average times 19,126
sum.ECavgTPM.HvitN1 <- Ec.avg.TPM * HvitN1

# To get alpha, divide that amount by 1,000,000
alpha <- sum.ECavgTPM.HvitN1 * (10**(-6))

alpha
```

```
## [1] 0.9519929
```

This value for α, 0.95199…. etc. is very close to the ratio of the smaller number of transcripts to the larger:

```
checksum <- HvitN1/EcarN2
checksum
```

```
## [1] 0.9439738
```

```
checksum - alpha
```

```
## [1] -0.008019049
```

Scaling Hvit.TPM

```
HV_scaled <- Hvit.TPM[, -1]
row.names(HV_scaled) <- Hvit.TPM[, 1]
Hvit.TPM.scaled <- HV_scaled * alpha
dim(HV_scaled)

## [1] 17607    21

head(Hvit.TPM[, 3])

## [1]  55.633 244.292   9.388  58.149   0.000  34.698

# 55.633 244.292 9.388 58.149 0.000 34.698
head(Hvit.TPM.scaled[, 2])

## [1]  52.962220 232.564246   8.937309  55.357434   0.000000  33.032249

# 52.962220 232.564246 8.937309 55.357434 0.000000 33.032249
```

Multiplying by α results in our Hvit TPM numbers being just a little smaller, though it will matter a lot for some of the outrageously large numbers:

```
which(Hvit.TPM[, 3] == max(Hvit.TPM[, 3]))
## [1] 10605
# 10605
Hvit.TPM[, 3][10605]
## [1] 124817.4
# 124817.4

Hvit.TPM.scaled[, 2][10605]
## [1] 118825.2
# 118825.2
```

That's a difference of 5,992.2 transcripts per million.

Normalize for different library sizes within species

Finally, we want to normalize for size using DESeq2, and then write the resulting matrices to file for filtering to single copy orthologs and other downstream analysis.

```
write.table(Hvit.TPM.scaled, file = "Hvit.90.genes.TPM.scaled2.matrix",
    sep = "\t")

library(DESeq2)
colData <- read.table("Hemiptera_SampleInformation_colData.txt",
    header = TRUE)
hvCol <- (colData[c(42:62), ])
hvTPM <- as.matrix(Hvit.TPM.scaled)
storage.mode(hvTPM) = "integer"
hv.dds <- DESeqDataSetFromMatrix(hvTPM, colData = na.omit(hvCol),
          design = ~Tissue + Pool)
```

```
hv.dds <- estimateSizeFactors(hv.dds)
hv.dds <- estimateDispersions(hv.dds)
hvScaledNorm <- as.data.frame(assay(hv.dds), normalized = TRUE)
write.table(hvScaledNorm, "Hvit_Genes_Scaled_SizeNormed_Integer_TPM.matrix",
    sep = "\t")



ECid <- Ecar.TPM[, 1]
Ecar.TPM <- Ecar.TPM[, -1]
rownames(Ecar.TPM) <- ECid
ec.dds <- as.matrix(Ecar.TPM)
ecCol <- colData[1:24, ]
storage.mode(ec.dds) = "integer"

ec.dds <- DESeqDataSetFromMatrix(ec.dds, colData = ecCol, design = ~Tissue +
    Pool)

ec.dds <- estimateSizeFactors(ec.dds)
ec.dds <- estimateDispersions(ec.dds)
ecNorm <- as.data.frame(assay(ec.dds), normalized = TRUE)
write.table(ecNorm, "Ecar_Genes_SizeNormed_Integer_TPM.matrix", sep = "\t")
```

2. Train a Poisson-distributed Linear Discriminant Analysis Classifier with MLSeq

R (version 3.6.1) code used to create PLDA classifier and separate species markers from other genes. The below block of code is intended to run in a parallel processing environment such as a high-performance computing cluster, though a moderately powerful workstation that can provide at least six threads for processing will be sufficient. This code uses the MLSeq wrapper package[2] to train a PLDA classifier[3] to find the minimal set of genes that are sufficient to distinguish one species from another. More information and test data files are located at https://github.com/fishercera/TreehopperSeq.

```
#~~~~ Cera Fisher (2019) MIT License, use what you like
### MLSeq - finding species marker genes with machine learning
options(stringsAsFactors = FALSE)
library("dplyr") # dplyr_0.8.3
library("DESeq2") # DESeq2_1.24.0
library("MLSeq") # MLSeq_2.2.1


#Read in TPM matrix
MergedCounts <- read.delim("SCOsTPM_2.matrix", sep="\t", header=TRUE)
cts <- as.matrix(MergedCounts[,2:46])
storage.mode(cts) = "integer"
class <- data.frame(condition = factor(rep(c("Ecar","Hvit"),c(24,21))))

## Setting up a Class object for DESeq2
set.seed(2128)
vars <- sort(apply(cts, 1, var, na.rm = TRUE), decreasing=TRUE)
data <- cts # Operating on the whole data set
## You can randomly select the set of test samples, but
## with a small sample set, it's possible to get too many of
## one species; I chose to hand curate the test samples,
## but the below two lines will do it randomly if you uncomment them.
# nTest <- ceiling(ncol(data) * 0.3)
# ind <- sample(ncol(data), nTest, FALSE)
 GoodInd <- read.table("good_ind2.txt", sep="\t")     ## Comment these three
 ind <- as.vector(GoodInd)                            ## out if you do not
 ind <- ind[,1]                                       ## hand curate.
data.train <- as.matrix(data[ ,-ind] + 1)
data.test <- as.matrix(data[ ,ind] + 1)
classtr <- data.frame(condition = class[-ind, ])
classts <- data.frame(condition = class[ind, ])
cts.train <- as.matrix(cts[ ,-ind] + 1)

#Make DESeq objects
cts.trainS4 <- DESeqDataSetFromMatrix(countData = cts.train, colData=classtr,
design=formula(~condition))
featureData <- data.frame(gene=MergedCounts$OrthoID)
mcols(cts.trainS4) <- DataFrame(mcols(cts.trainS4), featureData)
```

```
mcols(cts.trainS4) <- DataFrame(mcols(cts.trainS4), data.frame(HVid=MergedCou
nts$HVid))
mcols(cts.trainS4) <- DataFrame(mcols(cts.trainS4), data.frame(ECid=MergedCou
nts$ECid))

# Set the parameters for the classifier
ctrl.PLDA <- discreteControl(method="repeatedcv", number=30, tuneLength=100,
repeats=100000, parallel=TRUE)

## We're setting tuneLength=100 to let the classifier
## try multiple different tuning parameters(rho).
## It will settle on the one that gives the sparsest model
## with the highest accuracy.

fit.all.PLDA <- classify(cts.trainS4, method="PLDA", preProcess="deseq-vst",
  control=discreteControl(ctrl.PLDA))
plot(fit.all.PLDA)
trained(fit.all.PLDA)

### During my run --
# The optimum model is obtained when rho = 24.08461 with
# an overall accuracy of Accuracy = 0.9730 over folds.
# On the average 320.48 out of 7635 features was used
# in the classifier.
## Use the selectedGenes function to pick the genes that are
## most species biased.
Markers <- selectedGenes(fit.all.PLDA)
Markers.Counts <- MergedCounts[Markers,]
 write.table(Markers.Counts, "New_SCOs_Tuned_MergedCounts_SelectedMarkers_tun
eLength100.txt")
 ## Filter out the ones that are not species biased.
 UnselectedGenes <- MergedCounts[(which(!(MergedCounts$OrthoID %in% MergedCou
nts[Markers, ]$OrthoID))), ]
 write.table(UnselectedGenes, "New_SCOs_TunedMergedCounts_UnselectedGenes_tun
eLength100.txt", sep="\t", quote=FALSE)
```

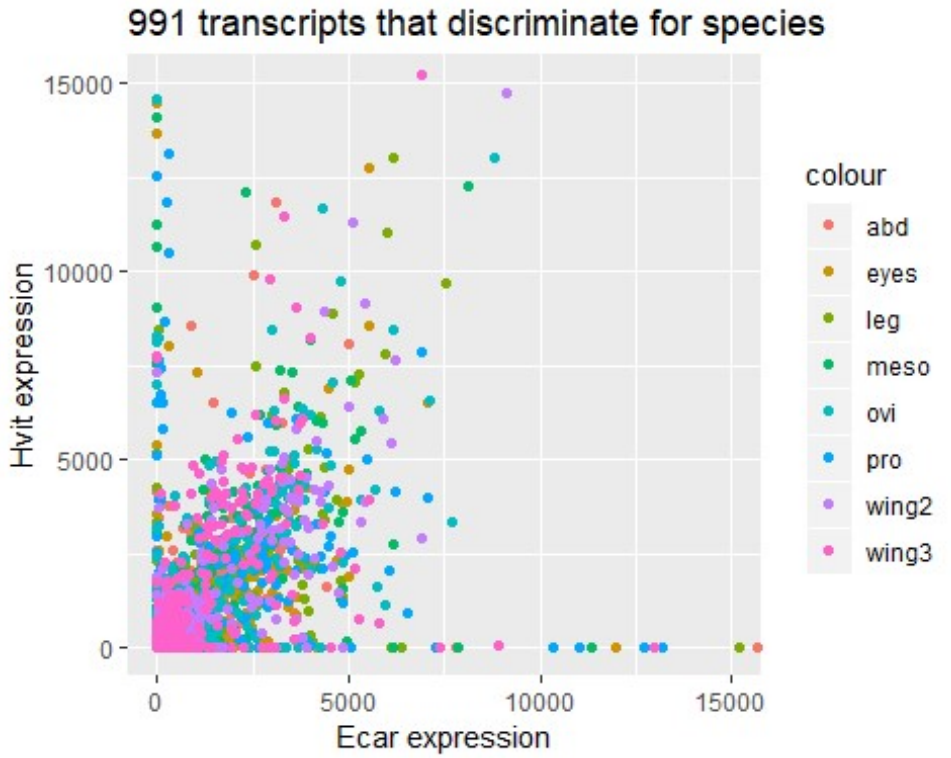
Comparing selected species markers to unbiased genes

The outcome of the PLDA classifier can be seen by plotting expression levels of “marker” genes and “unbiased” genes.

Read in selected markers, filter the counts matrices.

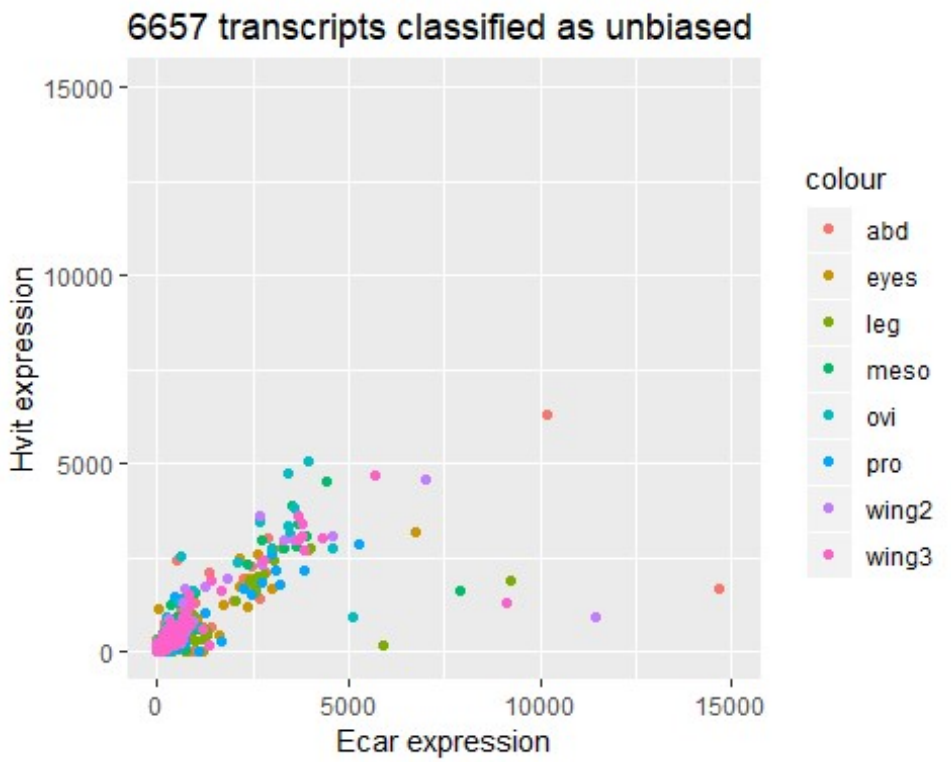
```
selectedMarkers <- read.table("1_Tuned_MergedCounts_SelectedMarkers_10000iter
.txt")
Merged.Counts <- read.table("HV_EC_NewSamples.SingleCopyOrthogroups.MergedTPM
.txt")
unselected <- filter(Merged.Counts,
                     !(Merged.Counts$OrthoID %in% selectedMarkers$OrthoID))
```

Plot the *E. carinata* TPM vs. *H. vitripennis* TPM for genes identified as species markers.



Gene expression (TPM) of biased transcripts in one replicate of E. carinata versus one replicate of H. vitripennis reveals a characteristic trident pattern. Extreme outliers in expression along the axes indicate genes highly expressed in one species and not expressed in the other.

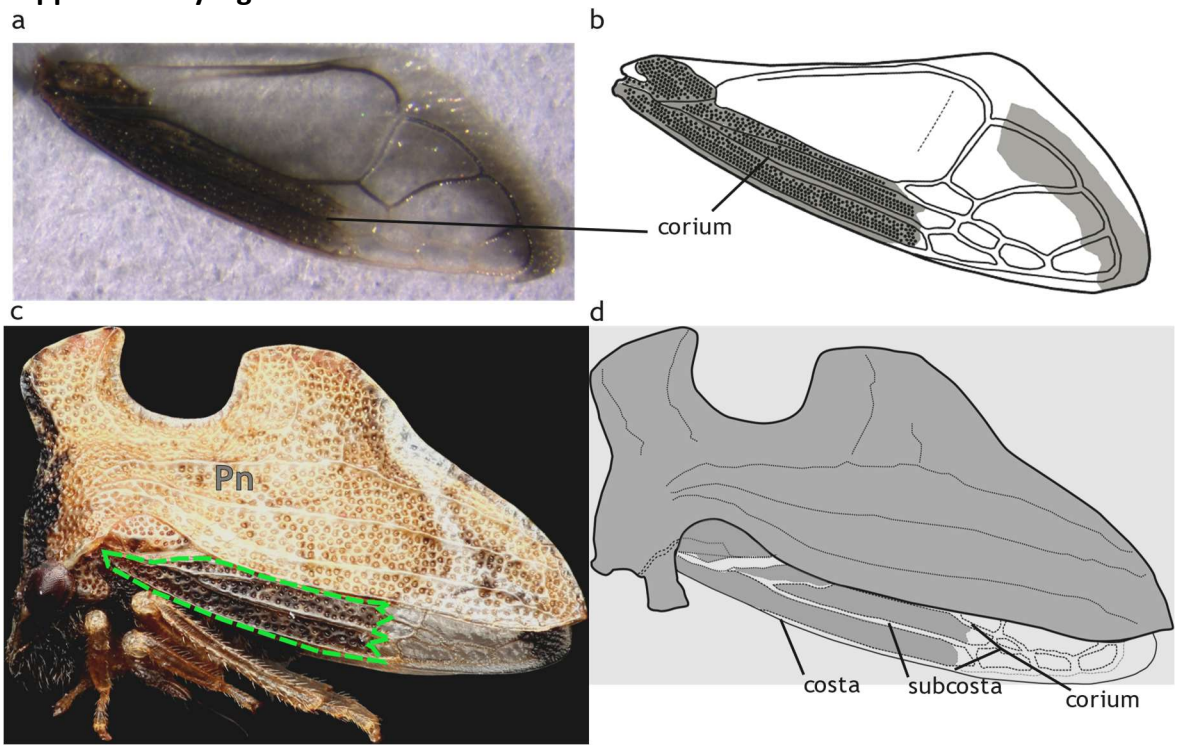
Plot *E. carinata* TPM vs. *H. vitripennis* TPM of genes not selected as species markers (unbiased genes).



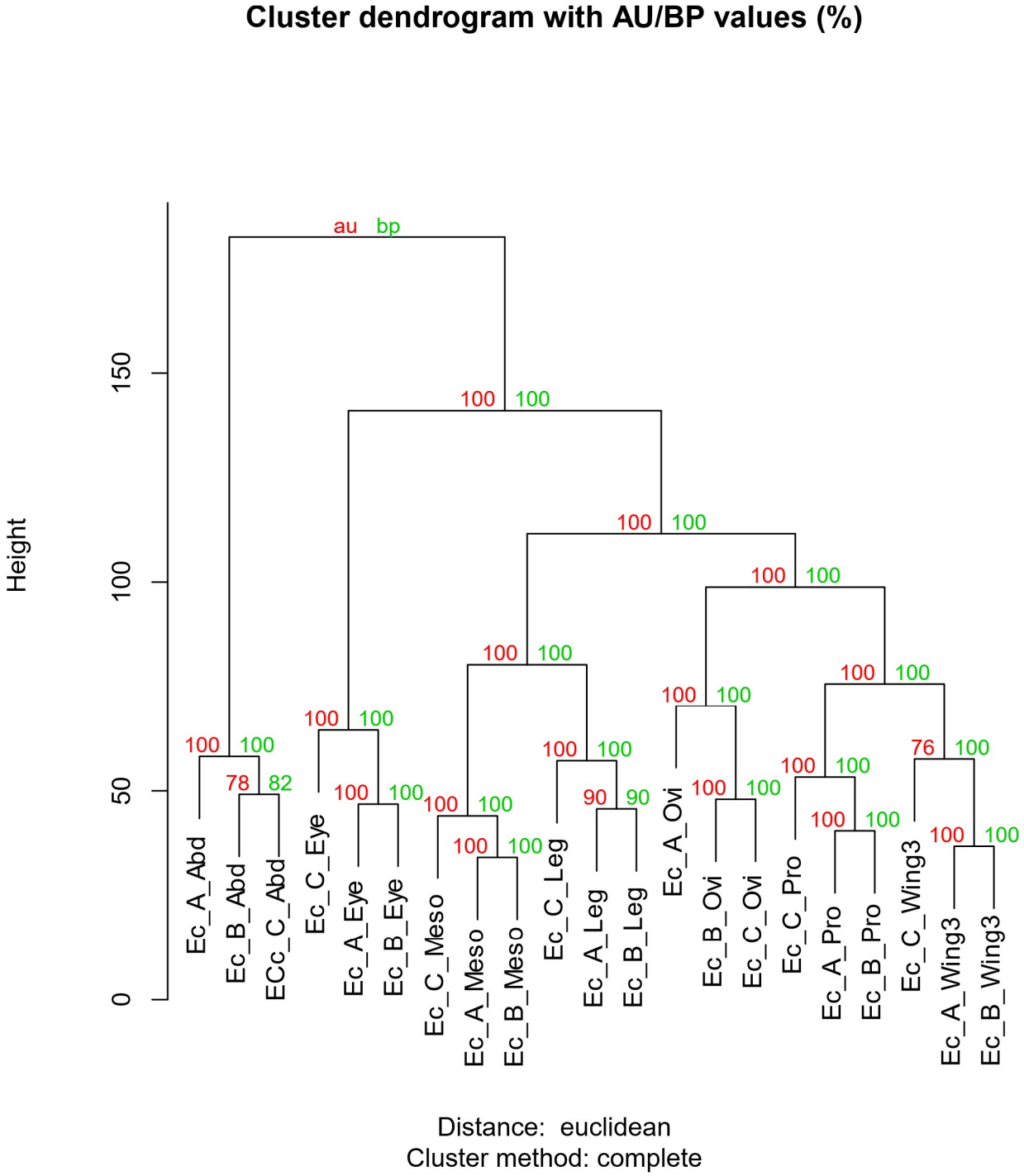
This plot does not have a trident pattern. Most genes with expression biased towards one species have been removed from the set. A countable few (10) remain with highly biased Entylia expression.

```
write.table(selectedMarkers, "MLSeq_PLDA_SelectedMarkers_991.txt")
write.table(unselected, "UnselectedTranscripts_6657.txt")
```

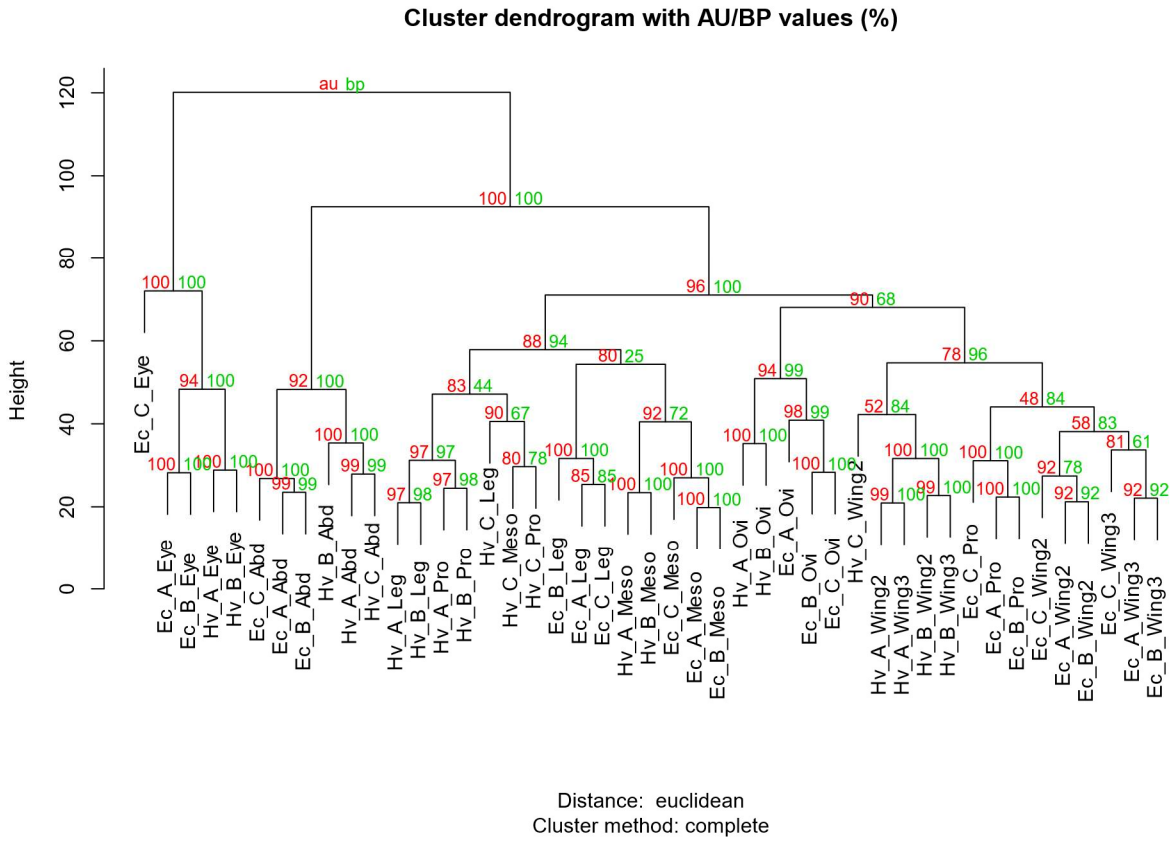
Supplementary Figures



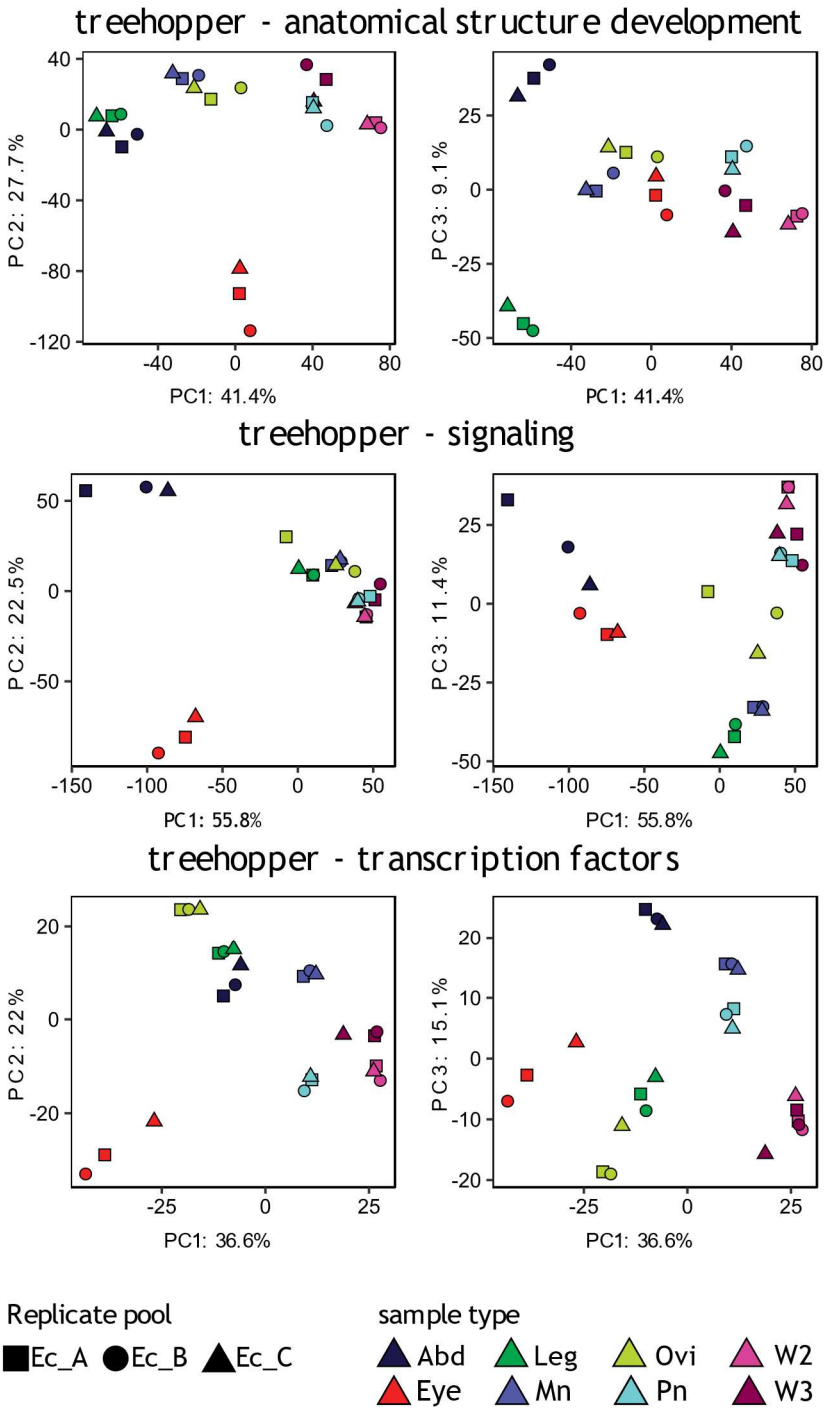
Supplementary Figure 1 The corium of the *E. carinata* forewing has very similar characteristics to the helmet. The costal-subcostal area (the corium) is sclerotized through two-thirds the length of the wing and bears the same punctate pattern as the helmet. **a**, A forewing from a wild-type male *E. carinata*, oriented as in live individuals with proximal hinge in the upper left corner and anterior edge towards the bottom, shows the corium as a darkly pigmented and sclerotized patch on the anterior edge of the wing blade. **b**, Line drawing of same wing. **c**, A female *E. carinata* shows the sclerotized, punctate character of the helmet (Pn). The corium is outlined with a dashed green line. **d**, Line drawing of the helmet and forewing; forewing is normally tucked under the helmet with the corium visible.



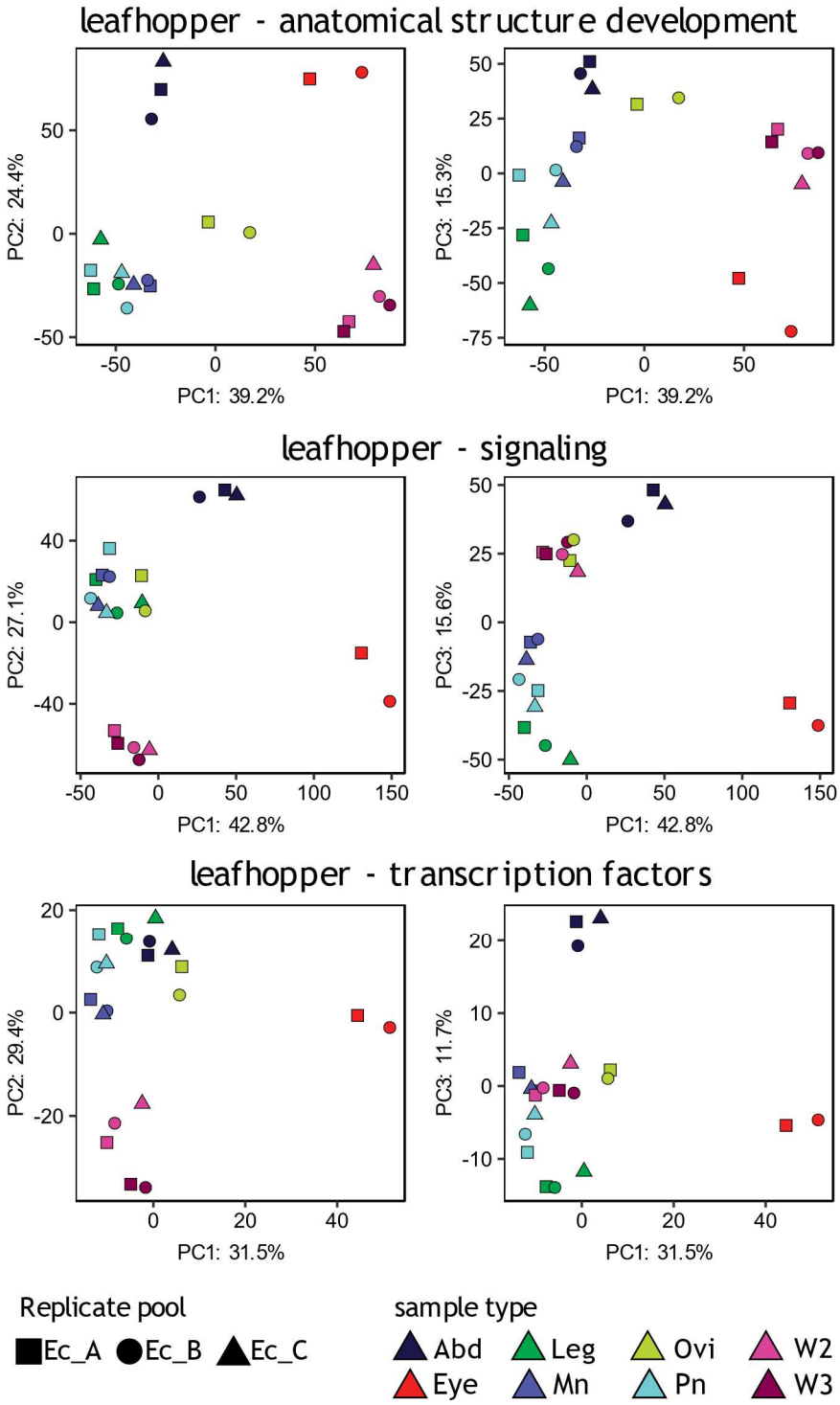
Supplementary Figure 2 Treehopper character tree with forewings excluded. Hierarchical clustering based on all genes differentially expressed across seven body regions (excluding forewings) in *Entylia carinata*. The helmet clusters with the hind wings. Support values (approximately unbiased values for multiscale bootstrap analysis in red and bootstrap in green) are percent of 1,000 replicates. A, B and C indicate the three different sibling pools. Ec – *E. carinata*; Abd – abdominal tergites; Meso – mesonotum; Pro – pronotum (=helmet); Wing3 – hind wing



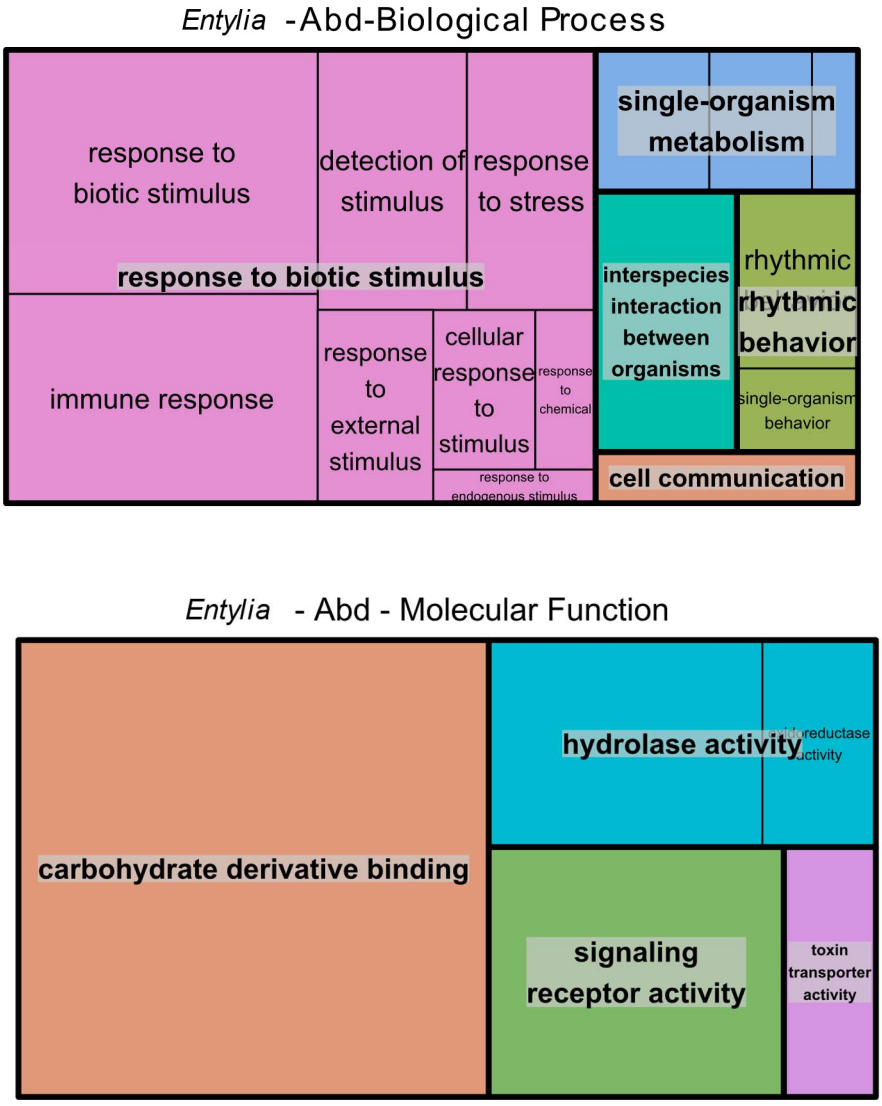
Supplementary Figure 3 Multispecies character tree excluding genes with a strong species signal. Hierarchical clustering of unbiased single copy orthologs results in nearly the same topology as the hierarchical clustering of all single copy orthologs (Fig. **2f**). Treehopper helmets (Ec_Pro samples) cluster with treehopper and leafhopper wings. Support values are percent of 1,000 replicates. Sample codes indicate species, pool, and body region. Ec – *Entylia carinata* (treehopper), Hv – *Homalodisca vitripennis* (leafhopper); A, B and C indicate the three pools in each species; Abd – abdominal tergites; Meso – mesonotum; Ovi – ovipositor; Pro – pronotum (=helmet); Wing2 = fore wing; Wing3 – hind wing.



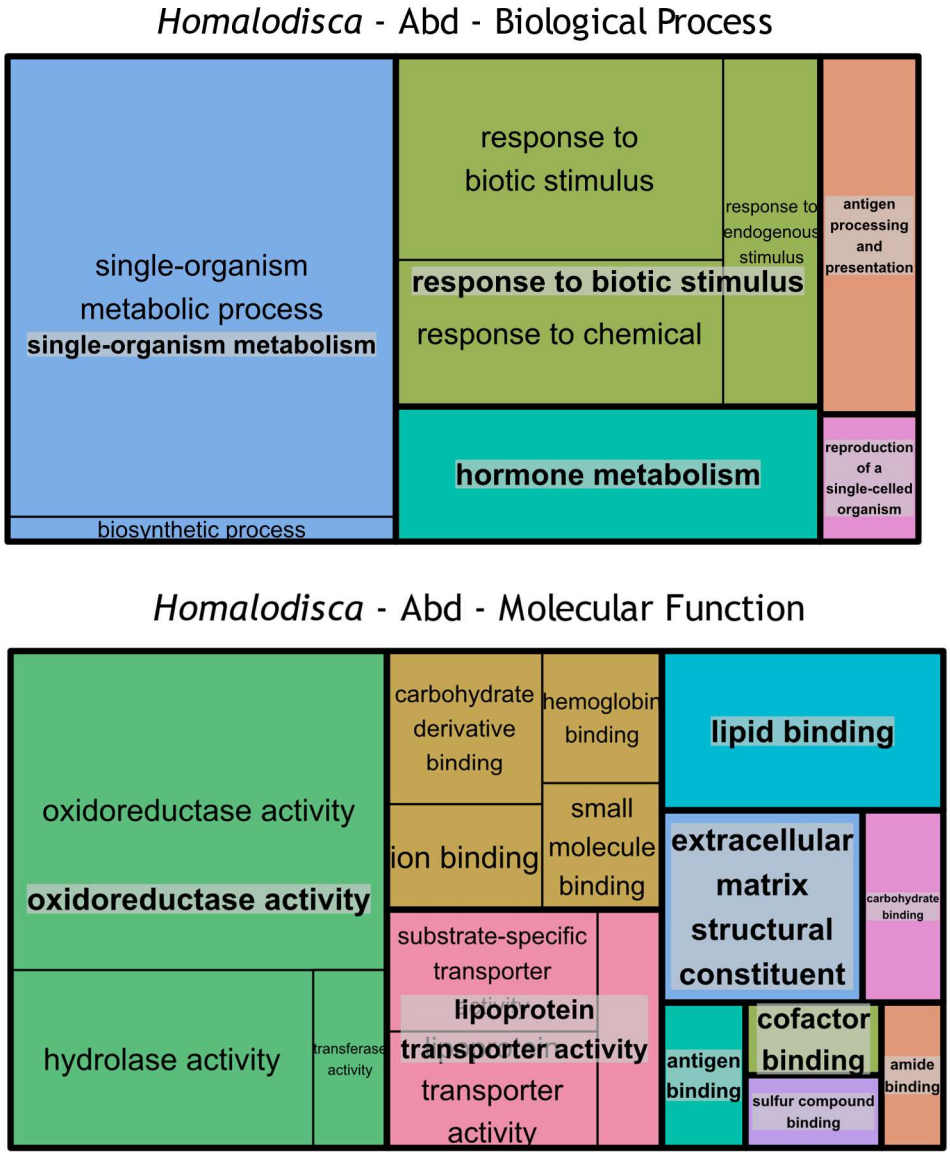
Supplementary Figure 4 Principal components analysis of differentially expressed gene subsets in the treehopper *Entylia carinata*. Genes annotated with (top row) anatomical structure development (GO:0048856), (middle row) signaling (GO:0023052), and (bottom row) transcription factor activity (GO:0003700). The pronotum (helmet) samples (cyan) are closets to the wings (dark and light purple) except in the case of transcription factors, where they appear intermediate between wings and their serial homologue the mesonotum. Abd – abdominal tergites; Meso – mesonotum; Ovi – ovipositor; Pro – pronotum (=helmet); W2 = forewings; W3 – hind wings



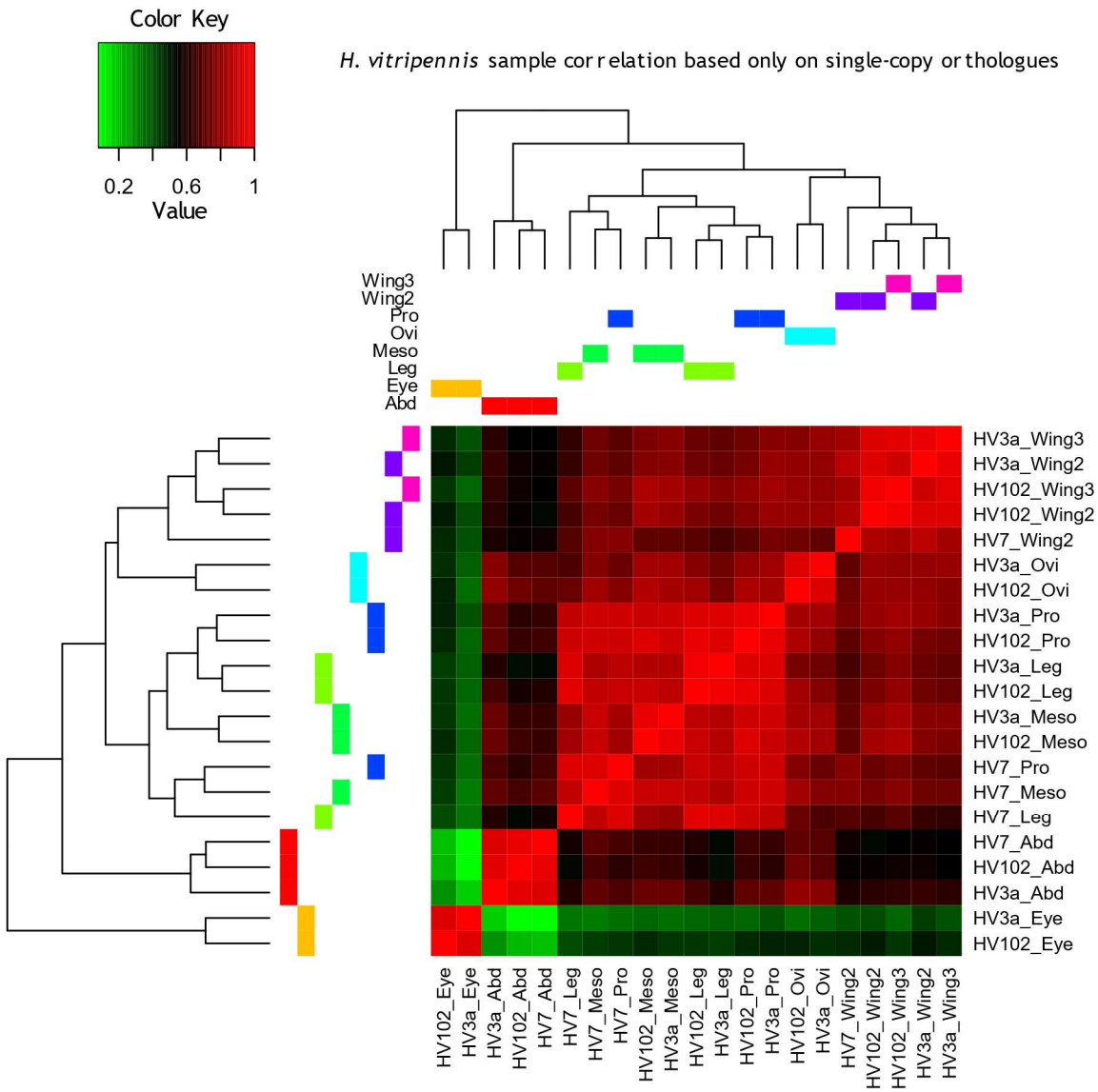
Supplementary Figure 5 Principal components analysis of differentially expressed gene subsets in the leafhopper *Homalodisca vitripennis*. Genes annotated with (top row) anatomical structure development (GO:0048856, top row), signaling (GO:0023052, middle row), and transcription factor activity (GO:0003700, bottom row). Color indicates body region and shape indicates sample pool. In all cases, the pronotum is closest to the mesonotum and legs, with the wings forming a separate cluster. Abd – abdominal tergites; Meso – mesonotum; Ovi – ovipositor; Pro – pronotum; W2 = forewings; W3 – hind wings.



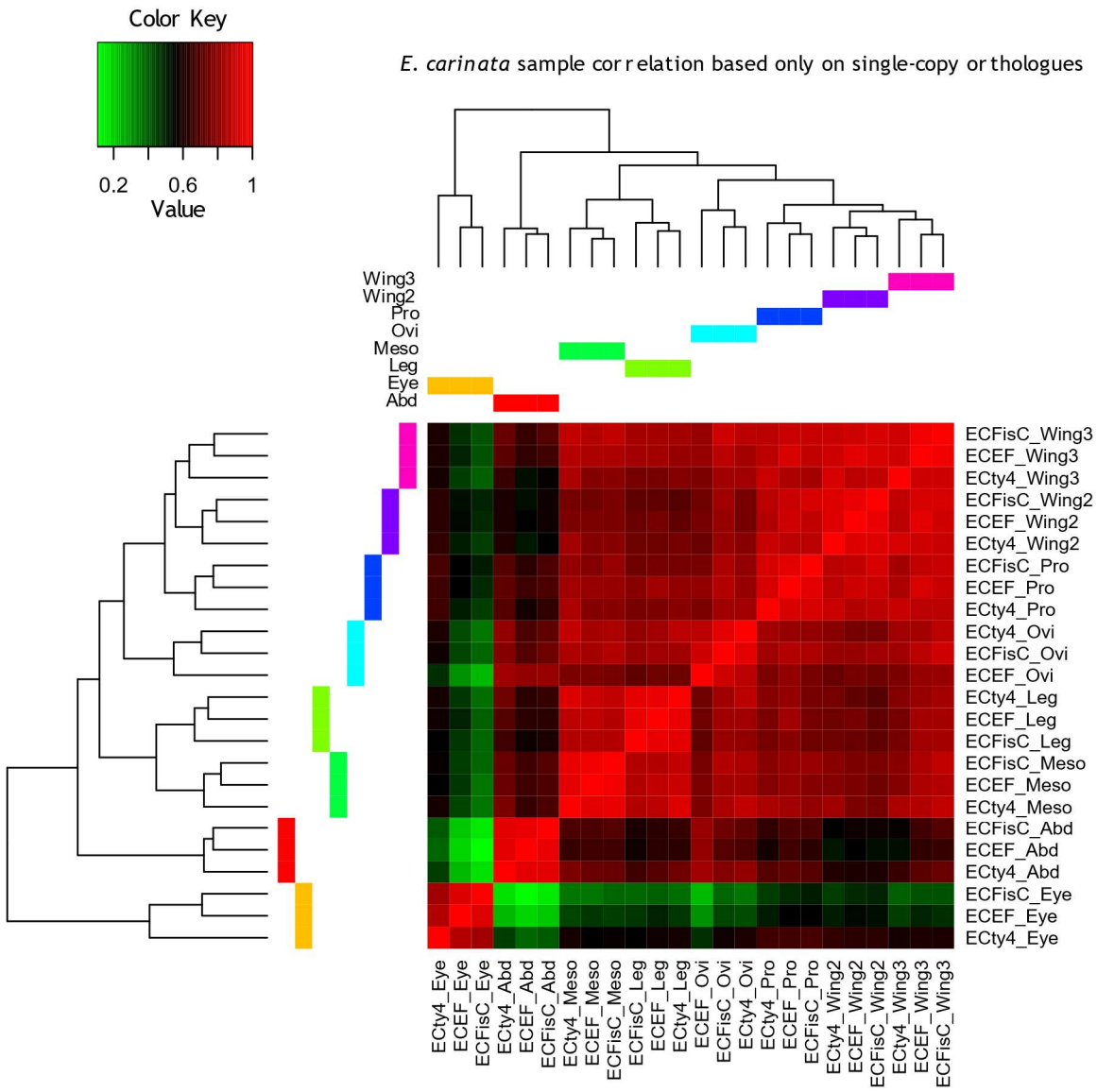
Supplementary Figure 6 Space-filling tree map of enriched GO terms for *Entylia carinata* abdominal tergite samples. The size of the box is inversely proportional to p-value for over-representedness (larger box = more significant). Semantically similar terms are colored with the same color. Produced using REVIGO[4].



Supplementary Figure 7 Space-filling tree map of enriched GO terms for *Homalodisca vitripennis* abdominal tergite samples. The size of the box is inversely proportional to p-value for over-representedness (larger box = more significant). Semantically similar terms are colored with the same color. Produced using REVIGO[4].



Supplementary Figure 8. *H. vitripennis* sample correlation based on single-copy orthologues. The topology of the hierarchical clustering dendrogram for *H. vitripennis* based on the single-copy orthologues is substantially identical to the clustering of *H. vitripennis* samples in Fig. **2f**. Hierarchical clustering was derived from a Euclidean distance matrix; heatmap values are pearson correlation.



Supplementary Figure 9. *E. carinata* sample correlation based on single-copy orthologues. As in Fig. S8, the topology of the hierarchical clustering dendrogram is identical to the clustering of *E. carinata* samples in Fig. **2f**. Hierarchical clustering derived from Euclidean distance matrix; heatmap values are pearson correlation.

Supplementary Tables

Supplementary Table 1. Summary of quality trimming and ribosomal RNA removal for the RNA-seq libraries. The number of raw reads, reads retained after quality control, percent of reads from ribosomal RNA, number of reads retained after removal of ribosomal RNA reads, and percent of the raw reads retained in analyses are shown. Libraries are named using the following convention: Ec (*Entylia carinata*, treehopper) and Hv (*Homalodisca vitripennis*, leafhopper) designate species; the following code (A, B, or C) designates sample pool; the last code designates body region (Abd – abdominal tergites; Eye – eyes; Leg – legs of 2nd thoracic segment; Meso – mesonotum; Ovi – ovipositor; Pro – pronotum (=helmet in treehopper); Wing2 = forewings; Wing3 – hind wings.)

Library	raw reads	post-trimming reads	% rRNA	final reads	net reads
Ec_A_Abd	23,227,569	23,003,993	5.15	21,889,131	94%
Ec_A_Eye	14,994,889	14,853,134	3.85	14,299,939	95%
Ec_A_Leg	24,848,721	24,687,463	6.65	23,120,630	93%
Ec_A_Meso	23,233,788	22,848,595	6.81	21,365,045	92%
Ec_A_Ovi	14,457,588	14,249,614	37.25	9,230,892	64%
Ec_A_Pro	13,985,336	13,897,289	13.92	12,055,253	86%
Ec_A_Wing2	20,969,693	20,857,757	6.35	19,589,285	93%
Ec_A_Wing3	25,991,260	25,857,311	36.87	16,979,394	65%
Ec_B_Abd	25,953,896	25,826,880	9.53	23,452,799	90%
Ec_B_Eye	25,497,617	25,328,462	6.56	23,759,195	93%
Ec_B_Leg	29,336,816	29,090,847	7.94	26,927,762	92%
Ec_B_Meso	25,438,630	25,198,786	2.64	24,564,652	97%
Ec_B_Ovi	20,200,399	20,039,181	7.29	18,618,932	92%
Ec_B_Pro	23,239,495	23,093,393	20.25	18,416,517	79%
Ec_B_Wing2	23,239,495	23,093,393	5.61	21,798,685	93%
Ec_B_Wing3	29,919,674	29,711,738	5.56	28,059,790	93%
Ec_C_Abd	24,343,022	24,121,921	3.49	23,324,342	96%
Ec_C_Eye	24,604,413	24,343,546	10.89	21,842,271	89%
Ec_C_Meso	29,132,415	28,903,710	4.49	27,677,741	95%

Library	raw reads	post-trimming reads	% rRNA	final reads	net reads
Ec_C_Ovi	28,566,102	28,340,173	2.99	27,557,138	96%
Ec_C_Pro	4,509,102	4,478,157	7.66	4,146,460	92%
Ec_C_Wing2	30,816,101	30,586,373	6.30	28,756,824	93%
Ec_C_Wing3	9,786,714	9,274,273	52.28	4,717,893	48%
Hv_A_Abd	24,693,802	24,532,330	4.35	23,495,596	95%
Hv_A_Eye	23,436,539	23,261,615	22.77	18,094,539	77%
Hv_A_Leg	35,092,710	34,830,393	1.86	34,200,771	97%
Hv_A_Meso	27,087,280	26,869,934	2.96	26,216,036	97%
Hv_A_Ovi	29,624,512	29,291,924	3.91	28,178,005	95%
Hv_A_Pro	25,103,114	24,646,791	2.21	24,120,203	96%
Hv_A_Wing2	29,392,885	29,026,425	25.29	21,972,857	75%
Hv_A_Wing3	34,172,192	33,847,851	24.79	25,786,602	75%
Hv_B_Abd	30,971,951	30,489,720	1.50	30,047,911	97%
Hv_B_Eye	33,676,632	33,463,799	24.70	25,517,945	76%
Hv_B_Leg	26,043,049	25,852,170	21.59	20,473,536	79%
Hv_B_Meso	27,732,956	27,256,565	32.54	18,722,058	68%
Hv_B_Ovi	19,528,212	19,354,241	10.60	17,358,784	89%
Hv_B_Pro	22,471,128	22,266,942	43.96	12,838,883	57%
Hv_B_Wing2	14,112,100	13,952,378	1.79	13,712,923	97%
Hv_B_Wing3	31,372,769	31,083,967	29.13	22,367,163	71%
Hv_C_Abd	29,231,825	29,003,230	3.60	27,989,467	96%
Hv_C_Leg	20,685,457	20,377,174	4.30	19,516,233	94%
Hv_C_Meso	13,816,701	13,715,799	58.58	6,106,913	44%
Hv_C_Pro	22,912,613	22,426,612	48.77	11,830,188	52%
Hv_C_Wing2	16,470,406	16,082,786	36.96	10,284,694	62%

Supplementary Table 2 Assembly statistics for reference transcriptomes

Reference assemblies for *E. carinata* and *H. vitripennis* were derived from *de novo* assembly of reads from each body region library. Below are initial statistics for each body region assembly from the two biological replicates with the highest total read count ("ECEF" = Ec_A = *E. carinata*; "HV102" = Hv_A = *H. vitripennis*) and the statistics for the final reference assemblies. Note that while total number of genes, transcripts, and assembled bases are lower in the reference assemblies than the body region assemblies, N50 (a measure of overall contig length) is higher for the reference assemblies than the body region assemblies.

library	genes	transcripts	N50	assembled bases
ECEF Abd	27,212	42,964	1,838	57,194,491
ECEF Eye	31,635	56,580	2,430	90,356,240
ECEF Leg	31,886	58,165	2,402	91,991,239
ECEF Meso	35,381	64,334	2,293	95,854,108
ECEF Ovi	24,734	36,845	1,874	48,954,549
ECEF Pro	27,339	43,990	1,979	60,004,120
ECEF Wing2	31,000	54,767	2,017	76,074,529
ECEF Wing3	30,996	53,560	1,990	73,511,702
HV102 Abd	28,174	43,926	2,241	67,314,338
HV102 Eye	28,230	46,629	2,811	83,542,484
HV102 Leg	26,955	47,919	2,613	83,907,025
HV102 Meso	26,606	46,535	2,893	86,195,261
HV102 Ovi	27,567	48,294	2,677	83,494,954
HV102 Pro	23,977	38,665	2,600	65,881,279
HV102 Wing2	32,650	56,845	2,791	99,052,314
HV102 Wing3	35,598	63,490	2,555	103,965,286
ECEF Reference	18,675	19,975	2,718	38,648,445
HV102 Reference	17,630	19,126	3,193	43,074,103

Supplementary Table 3 GO term enrichment for *H. vitripennis* for the set of transcripts upregulated in the wings and pronotum, and the set of transcripts upregulated in the wings only. Frequency means the frequency with which the term occurs in the set relative to the background. Log10 p-value measures how significantly over-represented the term is in the transcript set relative to the whole transcriptome. Ontology codes identify which aspect of gene function the GO term applies to: CC = cellular component, MF = molecular function, BP = biological process.

GO term ID	description	frequency	log10 p-value	ontology
	***H. vitripennis* wings & pronotum**			
GO:0001871	pattern binding	0.13%	-3.8834	MF
GO:0042302	structural constituent of cuticle	1.26%	-7.2426	MF
GO:0030246	carbohydrate binding	1.17%	-2.3682	MF
	***H. vitripennis* wings**			
GO:0031012	extracellular matrix	1.87%	-4.7754	CC
GO:0044421	extracellular region part	6.65%	-2.8517	CC
GO:0007155	cell adhesion	2.01%	-2.8386	BP
GO:0044767	single-organism developmental process	30.72%	-6.5387	BP
GO:0050789	regulation of biological process	34.90%	-3.0748	BP
GO:0044707	single-multicellular organism process	29.71%	-3.9876	BP
GO:0048856	anatomical structure development	29.52%	-5.4388	BP
GO:0001871	pattern binding	0.13%	-2.7133	MF
GO:0003700	transcription factor activity	3.36%	-7.9222	MF
GO:0042302	structural constituent of cuticle	1.26%	-5.5851	MF
	H. vitripennis pronotum			
GO:0008307	structural constituent of muscle	50.00%	-28.342	MF
GO:0042302	structural constituent of cuticle	34.0%	-27.929	MF
GO:0005200	structural constituent of cytoskeleton	28.57%	-14.807	MF
GO:0002209	behavioral defense response	100%	-9.010	BP
GO:0060361	flight	33.33%	-8.028	BP
GO:0044085	cellular component biogenesis	6.90%	-7.397	BP
GO:0016491	oxidoreductase activity	6.90%	-6.920	MF
GO:0022857	transmembrane transporter activity	6.45%	-6.733	MF
GO:0003823	antigen binding	33.33%	-5.580	MF

Supplementary Table 4 GO term enrichment for *E. carinata* for the set of transcripts upregulated in the wings and pronotum, and the set upregulated in wings alone. Columns are as in Table S2.

GO term ID	description	frequency	log10 p-value	ontology
	***E. carinata* wings & pronotum**			
GO:0031012	extracellular matrix	1.87%	-4.7105	CC
GO:0044421	extracellular region part	6.65%	-2.3320	CC
GO:0044767	single-organism developmental process	30.72%	-4.4225	BP
GO:0050789	regulation of biological process	34.90%	-2.2416	BP
GO:0044707	single-multicellular organism process	29.71%	-3.3088	BP
GO:0048856	anatomical structure development	29.52%	-3.3373	BP
GO:0003700	transcription factor activity	3.36%	-6.0669	MF
GO:0005515	protein binding	20.27%	-4.3922	MF
	***E. carinata* wings**			
GO:0031012	extracellular matrix	1.87%	-8.5437	CC
GO:0043227	membrane-bounded organelle	38.98%	-4.4874	CC
GO:0044464	cell part	64.71%	-3.9552	CC
GO:0044421	extracellular region part	6.65%	-4.1917	CC
GO:0007155	cell adhesion	2.01%	-3.0871	BP
GO:0009605	response to external stimulus	8.60%	-2.3755	BP
GO:0048589	developmental growth	3.25%	-3.1008	BP
GO:0050789	regulation of biological process	34.90%	-11.0504	BP
GO:0051674	localization of cell	2.95%	-3.7315	BP
GO:0044707	single-multicellular organism process	29.71%	-13.1404	BP
GO:0044763	single-organism cellular process	42.86%	-6.0246	BP
GO:0044767	single-organism developmental process	30.72%	-17.4772	BP
GO:0003006	developmental process involved in reproduction	7.70%	-3.3401	BP
GO:0048856	anatomical structure development	29.52%	-13.9632	BP
GO:0003700	transcription factor activity, sequence-specific DNA binding	3.36%	-21.59	MF
GO:0005515	protein binding	20.27%	-7.7896	MF
GO:0042302	structural constituent of cuticle	1.26%	-5.5229	MF
GO:1901363	heterocyclic compound binding	25.11%	-4.3025	MF
	***E. carinata* pronotum**			
GO:0044421	extracellular region part	9.29%	-27.763	CC
GO:0003700	transcription factor activity	5.81%	-20.869	MF
GO:0044767	single-organism developmental process	2.62%	-11.039	BP
GO:0009605	response to external stimulus	3.33%	-8.948	BP
GO:0042221	response to chemical	3.13%	-8.646	BP
GO:0048856	anatomical structure development	2.40%	-8.207	BP
GO:0003823	antigen binding	40.00%	-6.638	MF
GO:0042302	structural constituent of cuticle	6.45%	-5.748	MF

Supplementary Table 5. **Candidate wing genes and corresponding transcript ids.** Orthology of treehopper and leafhopper genes to selected *Drosophila* genes was established by reciprocal best BLAST hit and by OrthoFinder (version 2.3.3)[5]. One-to-one orthology was established between *E. carinata* and *H. vitripennis* transcipts by reciprocal best blast hit.

Gene name	*Entylia carinata*	*Homalodisca vitripennis*
apterous a	ECEF_Abd_TRINITY_DN21391_c1_g2	HV102_Abd_TRINITY_DN15582_c0_g1
ara-caup-1	ECEF_Leg_TRINITY_DN20415_c0_g1	HV102_Wing2_TRINITY_DN30205_c2_g3
ara-caup-2	ECEF_Pro_TRINITY_DN23458_c0_g1	HV102_Abd_TRINITY_DN21849_c0_g1
Distal-less	ECEF_Leg_TRINITY_DN18062_c0_g1	HV102_Leg_TRINITY_DN17912_c0_g1
engrailed	ECEF_Pro_TRINITY_DN23915_c2_g1	HV102_Meso_TRINITY_DN15783_c0_g1
four-jointed	ECEF_Pro_TRINITY_DN22154_c1_g1	HV102_Ovi_TRINITY_DN13942_c0_g1
frizzled-like	ECEF_Abd_TRINITY_DN18125_c0_g1	HV102_Meso_TRINITY_DN18120_c0_g1
grainy head iso. A	ECEF_Pro_TRINITY_DN23201_c0_g1	HV102_Meso_TRINITY_DN20168_c0_g1
grainy head iso. B	ECEF_Abd_TRINITY_DN22039_c0_g1	HV102_Leg_TRINITY_DN20894_c0_g1
miniature	ECEF_Wing3_TRINITY_DN27385_c0_g1	HV102_Wing3_TRINITY_DN28832_c2_g2
mirror	ECEF_Abd_TRINITY_DN20001_c0_g1	HV102_Meso_TRINITY_DN19892_c1_g2
nubbin	ECEF_Pro_TRINITY_DN18827_c0_g1	HV102_Wing3_TRINITY_DN33568_c0_g1
optomotor-blind	ECEF_Leg_TRINITY_DN19449_c0_g1	HV102_Ovi_TRINITY_DN16153_c1_g1
pangolin	ECEF_Eye_TRINITY_DN21520_c0_g2	HV102_Leg_TRINITY_DN13373_c0_g1
pannier	ECEF_Abd_TRINITY_DN20906_c1_g3	HV102_Abd_TRINITY_DN18890_c0_g1
rotund iso. A	ECEF_Eye_TRINITY_DN23855_c0_g1	HV102_Abd_TRINITY_DN20180_c0_g1
rotund iso. B	ECEF_Wing2_TRINITY_DN25108_c0_g2	HV102_Ovi_TRINITY_DN19368_c0_g1
serum response factor	ECEF_Leg_TRINITY_DN20191_c0_g1	HV102_Meso_TRINITY_DN17951_c0_g1
spalt major	ECEF_Leg_TRINITY_DN16942_c0_g1	HV102_Meso_TRINITY_DN15588_c0_g1
u-shaped	ECEF_Leg_TRINITY_DN21668_c0_g1	HV102_Wing2_TRINITY_DN26420_c1_g2
ventral veins lacking	ECEF_Abd_TRINITY_DN20071_c0_g1	HV102_Abd_TRINITY_DN21560_c0_g1
vestigial	ECEF_Ovi_TRINITY_DN5341_c0_g1	HV102_Wing3_TRINITY_DN32678_c1_g2
wingless	ECEF_Eye_TRINITY_DN17804_c0_g1	HV102_Leg_TRINITY_DN16063_c0_g1

References

1. Musser, J. M. & Wagner, G. P. Character trees from transcriptome data: Origin and individuation of morphological characters and the so-called 'species signal'. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* **324**, 588–604 (2015).
2. Goksuluk, D. *et al. MLSeq: Machine learning interface to RNA-seq data*. (Bioconductor, 2018).
3. Witten, D. M. Classification and clustering of sequencing data using a poisson model. *Annals of Applied Statistics* **5**, 2493–2518 (2011).
4. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**, e21800 (2011).
5. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**, 157 (2015).