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

Cera 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

Understanding the origin of novelty is a key question in evolutionary developmental biology. In arthropods, the body wall has served as a repeated source of morphological novelty. In treehoppers, an ancestrally flat part of the dorsal body wall (the pronotum) was transformed into a three-dimensional structure (the helmet), which was subsequently molded by natural selection into diverse shapes. Here, we test three hypotheses for the developmental origin of the helmet by comparing body-region transcriptomes in a treehopper and a leafhopper that retains more ancestral morphology. In leafhoppers, pronotal gene expression is most similar to that of its serial homologue, the mesonotum. By contrast, in treehoppers, helmet gene expression is most similar to that of wings, supporting the wing-patterning network co-option hypothesis for the origin of the helmet. These results suggest that serial homologues may diverge evolutionarily through replacement of, rather than tinkering with, their ancestrally shared patterning network.

One key insight of evolutionary developmental biology is that changes in the expression of a small set of regulatory genes can have major effects on the evolution of form[1–3]. This insight, derived primarily from comparative analyses of candidate genes, has yielded new models

of how new body plans and novel traits originate and diversify. One way that morphological novelty may arise is via modulation of a genetic regulatory network underlying a trait, meaning that the regulatory relationships of genes in a network do not change over evolution, but the timing, duration or intensity of expression of some of the genes may change[4]. Modulation of gene expression may account for phenotypic evolution such as changes in allometry and other large-scale shape changes[5,6]. Another model for the evolution of novelty is via co-option of existing gene regulatory networks, resulting in the expression of a set of co-regulated genes in a new developmental context[7]. Co-option has been implicated in insect novelties such as the mimicry patterns of butterflies[7,8], the hardened elytra of beetles[9], and the grasping structures of male water strider antennae[10].

In arthropods, many morphological novelties originate as outgrowths of the body wall, for example beetle horns[11–13], crustacean carapaces[14], mayfly gills[15], and perhaps even insect wings[15–17]. A particularly stunning example of a novel body wall outgrowth is found in treehoppers, sap-sucking insects of the family Membracidae (Hemiptera) and allies. These insects are distinguished from their close relatives the leafhoppers[18] (family Cicadellidae) by a body wall outgrowth referred to as a helmet (Fig. 1). Anatomically, a treehopper's helmet is composed primarily of the pronotum, the dorsal body wall of the insect's first thoracic segment[19–21]. The membracid pronotum is a bilayered evagination of the body wall projecting in three directions—anteriorly, posteriorly, and dorsally—to form a three-dimensional structure. There are more than 3,300 species of treehoppers worldwide[22], with helmet structures ranging from a simple posterior projection to architecturally complex structures sculpted to resemble hymenopterans and other shapes[23] (Fig. 1). In the ancestral condition, which is retained in leafhoppers and other hemipterans, the pronotum is a flat, shield-like part of the exoskeleton that

lies flush with the mesonotum[20,24–26] (Fig. 1e). While the pronotal identity of the treehopper helmet is clear, it remains a mystery how the dramatic transformation of flat body wall into a complex and often elaborate three-dimensional structure came about. What changed in development to give rise to the treehopper helmet?

We consider three main hypotheses for the developmental origin of the treehopper helmet. The first is that modulation of the ancestral body wall gene regulatory network led to outgrowth of the pronotal body wall, possibly via extended expression of growth-promoting pathways. The other two hypotheses, each of which is supported by some marker gene expression[27], involve co-option. The leg-network co-option hypothesis holds that the proximodistal axis of the treehopper helmet—i.e., the body wall outgrowth—evolved by redeployment of a portion of the gene regulatory network ancestrally involved in leg outgrowth, as has been observed for anatomically similar beetle horns[6,28,29]. The wing-network co-option hypothesis arises from a different interpretation of the data presented by Prud'homme and colleagues[27], who proposed that the treehopper helmet is an atavistic pair of wings that evolved by reactivation of ancestrally suppressed wing development on the normally wingless prothorax. While the identity of the helmet as a bona fide wing was refuted on morphological grounds, wing-network co-option was suggested as a plausible explanation for the similarities between wing and helmet [20,21].

Our investigation of the origin of the treehopper helmet also examines theoretical predictions about the divergence of serial homologues. Serial homologues are expected to have similar transcriptional profiles because they are built from the same developmental plan[30,31]. This expectation has been borne out in studies of flower organs[32,33] and tetrapod limbs[31], but has not been broadly tested. Additionally, the co-option of gene regulatory networks may overwrite this

transcriptional signature[7,34]. Our three hypotheses make distinct, testable predictions about transcriptional similarity between relevant body regions, which can be modeled as character trees depicting which body regions are most similar to each other[30] (Fig 2. a–c).

We tested these hypotheses by using RNA-seq to compare gene expression of different body regions in two species: a leafhopper, *Homalodisca vitripennis*, and a treehopper, *Entylia carinata*. While the pronotum clusters with its serial homologue in the leafhopper character tree, as predicted for the ancestral condition, the expression profile of the treehopper helmet is most similar to that of treehopper wings, and several genes in the canonical wing-patterning pathway are upregulated in both body regions. Thus, our results support the wing-network co-option hypothesis for the evolution of the treehopper helmet.

Results

Study design and comparative transcriptional profiling

We selected eight body regions for transcriptional profiling: eye, pronotum/helmet, mesonotum, second thoracic (T2) leg, forewing pads (T2 wings), hind wing pads (T3 wings), abdominal tergum, and ovipositor. This set of samples includes the body regions predicted to be transcriptionally most similar to the helmet according to each of our three hypotheses (mesonotum, wing, and leg), and also includes three sets of serial homologues (ovipositor/leg, pronotum/mesonotum/abdominal tergum, fore-/hind wing) and a set of conspicuously metamorphic structures (ovipositor/wings/helmet). While leafhoppers and treehoppers are hemimetabolous (i.e., have incomplete metamorphosis), and juveniles have the same body plan as adults in most respects, the metamorphic structures undergo dramatic growth and morphological change to become fully functional in the adult form. Our sampling design allowed

us to investigate transcriptional similarities due to serial homology while accounting for shared transcription due to metamorphosis. We constructed libraries from three biological replicates for two species, the treehopper *E. carinata* and the leafhopper *H. vitripennis*, using last instar nymphs (Fig. 1c,f), since the helmet is a metamorphic structure. 45 of 48 libraries passed quality control and these yielded 1.83 billion total reads. Following trimming, *de novo* assembly, annotation and clustering, a reference transcriptome was generated for each species. This produced 18,652 (19,949 isoforms; N50 = 2,719 bp) centroid protein sequences for *E. carinata* and 17,609 (19,103 isoforms; N50 = 3,196 bp) for *H. vitripennis*. A BUSCO analysis[35] using insect single-copy orthologues indicated that our assemblies are high quality, based on high completeness (97.1–97.9%), low duplication (8.0–8.4%), low fragmentation (0.4–1.2%) and few missing genes (1.7%). 3,353 transcripts were identified by DESeq2[36] as differentially expressed (significant log-fold change in pairwise tests between body regions) in *E. carinata* and 4,428 in *H. vitripennis*. (Additional assembly statistics are available in Supplemental Tables 1–2.)

Clustering analyses show altered pronotal relationships between the two species and support the wing-network co-option hypothesis for the origin of the treehopper helmet

Character trees were developed as a framework for studying the origin and divergence of morphological characters[30]. The character tree approach clusters traits hierarchically based on transcriptional similarity. Analogously to the shared ancestry depicted by phylogenetic trees, shared developmental history will lead to grouping of specific body regions. As expected, when we applied this approach to the sets of differentially expressed transcripts in the treehopper and leafhopper RNA-seq data, the samples clustered primarily by body region rather than biological replicate, with strong support (Fig 2d,e). The main exceptions to this pattern involved cases in

which serial homologues first clustered by biological replicate. According to all three of our hypotheses, the leafhopper pronotum is predicted to cluster with the mesonotum, as these structures are expected to retain the developmental signal of serial homology. This expectation was supported, with all pronotal and mesonotal leafhopper samples clustering with moderate support, as measured by multiscale bootstrap resampling; they in turn were most similar to the leg samples (Fig. 2d).

The three hypotheses for the origin of the treehopper helmet differ in the expected placement of the treehopper pronotal sample. According to the modulation hypothesis (Fig. 2a), the treehopper character tree should match the leafhopper tree, because the pronotum is patterned by a general notal developmental network in both species. By contrast, the co-option hypotheses predict that the treehopper pronotum will cluster with either its legs (Fig. 2b) or wings (Fig. 2c), with other aspects of the character tree matching the leafhopper character tree. With the exception of the pronotum, hierarchical clustering patterns are identical between the species, showing evolutionary stability of transcriptional similarity, a prerequisite for our inference approach (Fig. 2d,e). However, in the treehopper, the helmet (pronotum) clusters most closely with the wings and is only distantly related to its serial homologue the mesonotum (Fig. 2e). Multiscale bootstrap resampling support for the pronotum-wing cluster was very strong (approximately unbiased (AU) support = 100%). This indicates a major shift in gene transcription patterns, such that expression of differentially expressed genes in the treehopper pronotum most closely resembles that in treehopper wings, as predicted by the wing-network co-option hypothesis.

Character trees highlight the strongest patterns in differential gene expression, and thus are a good test of the co-option hypotheses, but they may obscure non-hierarchical relationships.

Therefore, we also used principal components analysis (PCA) to further investigate pronotal similarity to other body regions. In treehoppers, the pronotum samples were closest to wing samples (Fig. 3a,b), while in leafhoppers, the pronotum samples were closest to mesonotum and legs (Fig. 3c,d). Thus, the PCA results also match the patterns predicted under the wing-network co-option hypothesis.

In *E. carinata*, the forewing includes sclerotized regions that resemble body wall in some ways (Supplementary Fig. 1). This suggests an alternative explanation to wing-network co-option for the close helmet/wing relationship in the treehopper character tree: this clustering could instead be driven by acquisition of body-wall characteristics by the T2 wings, with the T2 wings drawing in the T3 wings based on shared wing characteristics. The T3 wing is a typical membranous wing, and thus the pronotum should only cluster with it if it shows transcriptional similarity based on wing characters. Thus, we repeated the above differential expression and character tree analyses omitting the T2 wings. The wing/helmet cluster was robust to removal of the T2 wings (AU support = 100%) (Supplementary Fig. 2), ruling out the possibility that the treehopper helmet is similar to treehopper wings as a result of co-option of exoskeleton patterning into the wings.

Most body parts, but not pronota, cluster across species, supporting an additional prediction of the wing-network co-option hypothesis.

We also sought to analyze the combined data from the two species, as an additional prediction of the co-option hypotheses is that each body region should cluster across species, with the exception of the pronotum. To do this, we repeated the character tree analyses on the set of 7,635 single-copy orthologues recovered from our annotated isoform proteomes with

OrthoFinder (version 2.33)[37]. While these transcripts represent only subsets of the transcriptome assemblies from each species, the subsets are directly comparable. The character tree resulting from the 1,420 features differentially expressed across body regions contained clusters uniting the eyes, abdominal body wall, ovipositors, and mesonota across species. However, the pronota of the two species were highly divergent. Treehopper pronota clustered with wings from both species, while leafhopper pronota formed a cluster with treehopper and leafhopper mesonota and legs (Fig. 2f). Because of a persistent 'species signal'[30], transcriptional divergence that characterizes all body regions within a species, we further sought to identify and remove species-biased transcripts using a Poisson linear discriminant analysis classifier applied to the scaled transcripts per million (TPM) values[38]. This approach identified 356 transcripts that were sufficient to reliably classify our samples by species, leaving 7,279 transcripts, of which 1,319 were differentially expressed across body regions. Overall clustering patterns were the same as those from the full orthologue set (Supplementary Fig. 3). Finally, we analyzed transcriptional similarity of differentially expressed single-copy orthologues via principal components analysis (PCA). The first three principal components collectively explained 83.0% of the variance in expression for these genes. Patterns consistent with the character trees were observed: eyes and abdominal terga were divergent from other body regions, but similar across species; a third group included the wings of both species and the treehopper pronotum; and a fourth group included mesonotum and ventral appendages of both species and the leafhopper pronotum (Fig. 3e, 3f).

Some transcript subsets support co-option while transcription factors character trees conserve the signal of serial homology

We further tested the robustness of the clustering analysis results by inferring character trees based on three transcript subsets that were chosen *a priori* based on their developmental relevance: anatomical structure development (GO:0048856), signaling (GO:0023052), and transcription factor activity (GO:0003700). Repeating our analysis workflow on transcript subsets that were annotated with these GO terms yielded 160–968 differentially expressed features. To maximize sample size, these analyses were conducted on the single-species datasets.

The structure of the character trees varied across these subsets (Fig. 4). For both the anatomical structure development and signaling subsets, the implications for the origin of the treehopper helmet were the same as for the full dataset. For these, character trees were similar or identical for the two species, with the exception of the pronotum placement. In both cases, the treehopper pronotum clustered with wings with good support (AU support = 99% for anatomical structure development genes, AU support = 76% for signaling genes) (Fig. 4b,d), while the leafhopper pronotum was divergent from wings and clustered with either the mesonotum (anatomical structure development genes, AU support = 97%) or legs and mesonotum (signaling genes, AU support > 90% accounting for sample-signal effect) (Fig. 4a,c).

Because of their roles in gene regulation, transcription factors are likely to be crucial members of character identity networks [30,39,40]. When we inferred character trees using only those transcripts annotated as encoding transcription factors with DNA binding activity (GO:0003700), the signal of serial homology was very strong. The treehopper character tree produced a perfect pattern of serial homology, with the pronotum clustering with the mesonotum and abdominal body wall, rather than with the wings, and with eyes as the most divergent body

region (Fig. 4f) (all AU supports > 80%). The leafhopper character tree (Fig. 4e) differed from the pattern predicted by serial homology in only one relationship: the legs and ovipositors did not cluster (all AU supports > 80%, except the branch subtending legs, with AU support = 68%). This result suggests that serial homologues have more strongly conserved developmental patterning at the level of transcription factors. The difference in treehopper pronotal placement between transcription factors and all other datasets also suggests that the transcriptional changes driving pronotum/wing clustering in treehoppers are largely downstream of this retained, ancestral character identity network. PCA results for the data subsets are similar (Supplementary Fig. 4–5).

What accounts for other exceptions to the predictions of clustering by serial homology?

The developmental similarity of serial homologues was most apparent in the transcription factor subset. We also found partial support for the prediction of clustering based on serial homology in the full dataset and the other subsets: in these, the T2 and T3 wings clustered in both species in most datasets, as did the pro- and mesonotum of leafhoppers. However, the ovipositor and legs did not cluster, and the abdominal body wall did not cluster with the thoracic body wall. These exceptions are interesting. In the full dataset, ovipositors of both species clustered with wings, rather than with their serial homologues the legs, though in *E. carinata* ovipositors fall outside the helmet/wing cluster. This cluster suggests a strong signal of metamorphosis or growth, as wings and ovipositors and the treehopper helmet are the traits undergoing the most extensive metamorphosis in these hemimetabolous insects. The ovipositors of both species also retain a secondary signal of serial homology evident in the PCA analyses, where they group with legs and mesonota along PC2 (Fig. 3). While the eyes were indeed

transcriptionally distant from all other body regions, in the full dataset, the abdominal body wall was even more divergent. Biological process GO term enrichment indicates that in both species' abdominal body wall, the genes that are significantly upregulated are more likely to be functionally annotated for immune system processes (GO:0002376) and responses to stimulus (GO:0050896), relative to the full transcriptome (Supplementary Fig. 6–7). Additionally, gene expression in the treehopper abdominal terga was enriched for interspecies interaction between organisms (GO:0044419) (Supplementary Fig. 6). Considering the differences in enrichment between abdominal and thoracic body wall, we suggest that the observed divergence in gene expression may be due to the presence of bacteriomes located in the abdomen near the dorsal body wall. Like other sap-sucking hemipterans, leafhoppers and treehoppers shelter endosymbiotic microorganisms in abdominal bacteriomes[41], and the abdominally-enriched GO terms appear to be related to this unique abdominal function.

GO term enrichment analyses further highlight the differences between leafhopper and treehopper pronota

Given that unexpected divergences from serial homology occurred for other samples in our data sets, we asked whether the genes accounting for the similarity in treehopper helmets and wings were functionally relevant to the wing-network co-option hypothesis. We used GO term enrichment analysis[42,43] to characterize the functions of the sets of genes that were upregulated in common across sample sets. First, we sought to identify processes enriched in both T2 and T3 wings in each species. We used the differential expression results from pairwise comparisons between body regions to identify the transcripts that were upregulated in both pairs of wing relative to any other body region. These yielded very similar GO term enrichment pictures in

252 treehoppers and leafhoppers (121 genes and 116 genes, respectively) (Fig. 5, and Supplemental
253 Tables 3-4). In the biological process tree, developmental process (GO:0032502), regulation of
254 biological process (GO:0050789), and cell adhesion (GO:0007155) are significantly
255 overrepresented. The molecular function GO terms enriched in both species' wings include
256 structural constituent of cuticle (GO:0042302), binding functions, and transcription factor
257 activity. For cellular component GO terms, extracellular region-related terms are enriched in
258 both treehopper and leafhopper wings.

259 Consistent with the transcriptional similarity between treehopper wings and helmets, the
260 set of enriched GO terms for genes upregulated in both wings and helmets relative to other body
261 regions (52 genes) includes many of the terms related to wings (Fig. 5). In biological process
262 terms, developmental process and anatomical structure development are significantly enriched;
263 in molecular function, transcription factor activity and protein binding are significantly enriched;
264 and in cellular components, the terms extracellular matrix and extracellular region part are
265 significantly enriched. A similar number of genes (39) is upregulated in both leafhopper wings
266 and pronotum relative to other body regions. However, the GO term enrichment pattern is
267 dramatically different, showing little overlap with terms enriched in the wings (Fig. 5). This set
268 of GO terms is not significantly enriched for biological process or cellular components; in
269 molecular function, only structural constituent of cuticle and binding functions remain, but not
270 transcription factor activity. These results indicate that the transcriptional similarity between
271 treehopper wings and helmet is the result of shared, identifiable developmental processes that are
272 not shared between leafhopper wings and pronotum.
273

Canonical wing patterning genes are upregulated in wings of both species and the pronotum of treehoppers but not in the leafhopper pronotum

Finally, we investigated the expression of specific genes that would be expected to contribute to transcriptional similarity between the helmet and wings under the wing-network co-option hypothesis. The wing development network has been extensively characterized in the fruit fly *Drosophila melanogaster*, and available evidence suggests that this network is highly conserved across winged insects[44]. As expected, genes in this network are highly expressed (greater than two-fold upregulation) in the wings of both leafhoppers and treehoppers (Fig. 6). Many of these canonical wing patterning genes are also expressed at higher levels in the developing helmet of treehoppers relative to other body regions. These include *vestigial* (*vg*), *apterous-A* (*apA*), *four-jointed* (*fj*), *serum response factor* (*srf*), a member of the *frizzled* (*fz*) family, *wingless* (*wg*), *engrailed* (*en*), *u-shaped* (*ush*), *miniature* (*m*), two isoforms of *rotund* (*rn*), and two isoforms of *grainy head* (*grh*). In *Drosophila*, these genes are implicated in various roles in wing development, namely early patterning (*fz*[45], *wg*[46], *en*[47]), dorsoventral and proximodistal axis patterning (*apA*[48,49], *vg*[50], *fj*[51]), wing hinge/notum differentiation (*ush*[52,53], *rn*[54]) and epithelial cell morphogenesis and adhesion (*grh*[45,55], *m*[56], *srf*[57,58]). However, only a few of these genes are upregulated in the leafhopper pronotum; most notably *apA*, *fz*, *srf*, *m*, *rn*, *fj*, and one isoform of *grh* are not (Fig. 6).

In both leafhoppers and treehoppers, several candidate genes with known involvement in *Drosophila* body-wall patterning were upregulated in common in the helmet/pronotum and mesonotum[59,60].Two genes orthologous to the *Drosophila* Iroquois-C (Iro-C) locus genes *araucan* and *caupolican*[59,60] are upregulated in the mesonotum, pronotum, and wings in both leafhoppers and treehoppers (Fig. 6a,b). Another Iro-C gene, *mirror*[61], is upregulated in

297 leafhopper wings, mesonotum, and pronotum, but in the treehopper, it is dramatically
298 upregulated in the pronotum relative to all three other body regions (Fig 6b). The gene *vg*,
299 originally considered to be a marker of wing identity but known to have important roles in body-
300 wall patterning[62–64], is upregulated not only in leafhopper and treehopper wings, but also in the
301 mesonotum and pronotum of both insects. Surprisingly, we did not detect upregulation of either
302 *nubbin* or *Distal-less* in treehopper pronotum (Fig 6b), though we expected it given the antibody
303 staining results found in a closely related treehopper[27]. Other wing- and body-wall-related genes
304 that we investigated showed similar patterns of relative expression across body regions in the
305 two species; these include *ventral veins lacking* (*vvl*)[65], *spalt major* (*salm*)[66], *optomotor-blind*
306 (*omb*)[67,68], *pannier* (*pnr*)[52], and *pangolin* (*pan*)[69] (Fig. 6a,b).
307
308 **Discussion**
309 In our study, we leveraged two dimensions of comparison—between species and between
310 body regions—in order to disentangle similarity due to shared identity, as in serial homologues,
311 from similarity due to co-option. Multiple lines of evidence support the conclusion that the
312 treehopper helmet evolved by co-option of the wing-patterning network. By considering
313 comparisons in light of genes' annotated functional roles, we showed that observed patterns of
314 gene expression differ for subsets of genes in biologically relevant ways. The evolution of novel
315 morphology by co-option of regulatory networks has commonly been diagnosed based on
316 analysis of candidate genes[9,14,34,70]. In this study, we provide a roadmap for analyzing co-option
317 from the gene expression patterns of many thousands of genes at once, without *a priori*
318 knowledge of which genes are important. Our results indicate that when suites of genes are
319 redeployed in the radical transformation of a body part, the transcriptional signal may be strong

320 enough to overwrite the shared gene expression of serial homologues. Comparative RNA-seq
321 and character tree analysis of novel traits in other taxa may reveal unexpected similarities
322 between traits.

323 However, our results also indicate that while co-option can result in a very strong pattern
324 of gene expression alteration across the whole transcriptome, other informative patterns can be
325 found in finer-grained analyses based on functional annotations. In our data, the expression
326 patterns of transcription factor-encoding genes indicated that the set of regulatory genes that
327 patterns the treehopper pronotum still has many members in common with those of its serial
328 homologues the mesonotum and abdominal terga. This result reinforces the conclusion from
329 morphology that the origin of the treehopper helmet did not involve a change in identity for the
330 substrate body part, the treehopper pronotum[20,21].

331 A key question that arises is from where the wing-gene network was co-opted. The wing-
332 patterning network was initially thought to be unique to wings[71]. However, recent work has
333 documented expression of core components of this network laterally (at the tergal-pleural
334 margin) in the first thoracic segment and abdominal segments of winged insects, leading to the
335 proposal that these regions are wing serial homologues[62,72–74]. In treehoppers, the tergal-pleural
336 margin is incorporated into the helmet[19,75]. At present, the extent to which the gene regulatory
337 network has diverged between wings and regions that are serially homologous with wings
338 remains unknown. Distinguishing between these evolutionary scenarios requires identifying
339 differences in network function and membership between the wings and other body regions, and
340 comparing the functional roles of the proteins encoded by the co-opted network members in
341 developing treehoppers and close relatives with the plesiomorphic condition.
342

Materials and Methods

Study design

We quantified gene expression in eight body regions of three biological replicates of two species, one treehopper and one leafhopper. For a treehopper species, we chose *E. carinata*. This species is common in the eastern USA and Canada, and has the additional beneficial quality of being multivoltine with a short developmental cycle (about two months from when eggs are laid to adulthood). We chose *H. vitripennis*, the glassy-winged sharpshooter, as the leafhopper representative. It is an important vector of Pierce's Disease in grapes[76] and is the subject of a genome sequencing project (NCBI BioProject Accession PRJNA168119).

E. carinata was raised on *Helianthus annuus* (sunflower) in the UConn EEB Research Greenhouse. The colony was established from individuals wild-collected in Windham and Tolland counties, Connecticut, USA, and has been in continuous culture for over three years. To minimize sample variation for RNA-seq, egg clutches from single females were isolated before hatching, and 5th instar nymphs were collected into RNAlater (Invitrogen). In order to amass enough tissue for library construction, nymphs were pooled within broods. This step has the potential to cause bioinformatic complications, due to individual variation in sequence and expression[77], so the size of the pools was minimized as much as possible. Our three biological replicates represent collections of siblings (pool A n=10, pool B n=6, pool C n=9) from three different broods (raised at different times), and were collected when the majority of the brood had advanced from 4th to 5th instar. *Homalodisca vitripennis* was reared by the California Department of Food and Agriculture Pierce's Disease Control Program. 5th instar nymphs were collected into RNAlater after being pierced through the abdomen to permit saturation of tissues. After 24 hours at room temperature, RNAlater preserved specimens were frozen at -20 °C and

stored for 1 week to 4 months prior to dissection. *H. vitripennis* nymphs are larger than *E. carinata* nymphs, and so these pools were smaller: pool A n=2, pool B n=3, and pool C n=2. *H. vitripennis* broods were not reared in isolation, so the relationships among pooled individuals are unknown.

Treehoppers and leafhoppers are hemimetabolous, and like other hemimetabolous insects, their wings, genitalia, and (in the case of treehoppers) helmet are nascent in early instars and acquire their adult form rapidly in the final nymphal instar[78]. Therefore, we dissected 5th (final) instar nymphs to acquire our samples. Approximate staging within instars is possible based on internal development of tissues. This was especially important for staging *H. vitripennis*, which were not staged during rearing. Nymphs with more opaque and thicker wing pads, which are later in the 5th instar, were preferentially selected for dissection where possible. Because *E. carinata* nymphs were same-aged siblings, they varied minimally in stage.

Library construction and sequencing

Preserved nymphs were dissected under RNAlater. Dissected body regions from each pool were stored in RNAlater at -20 °C until extraction. To mitigate batch processing effects, these samples were assigned random numbers prior to RNA extraction, and this processing queue was used through all following procedures. RNA was extracted with TRIzol reagent (Invitrogen), according to the manufacturer's instructions, incorporating a modification from the RNALater manual to accommodate salt carryover. Total RNA was quantified by fluorometry and RNA integrity was assessed by gel electrophoresis. Poly(A) enrichment was performed with Sera-Mag oligo-d(T) paramagnetic beads (GE Healthcare Sciences) according to standard protocols[79], with the exception that two rounds of enrichment were performed on RNA pools in

which RNA yield from the first round (assessed by fluorometry) indicated it was necessary. For verification of ribosomal RNA depletion, fragment analysis of a small number of poly(A)-enriched samples was conducted on a TapeStation 2200 (Agilent).

Illumina-platform compatible libraries were constructed using the Stranded RNA-Seq kit (KAPA), with NEBNext adaptors for Illumina (New England Biolabs), following the manufacturers' instructions except for using half-scale reactions. Libraries were barcoded with custom unique dual indices purchased from the Vincent J. Coates Genomics Sequencing Laboratory at the University of California, Berkeley in order to detect and discard any reads subject to the index swapping that occurs on the Illumina HiSeq 4000 platform[80]. Final libraries were assessed for average fragment size and concentration using the TapeStation High Sensitivity DNA ScreenTape Analysis (Agilent), then diluted and pooled at equal molarity. The pool was subjected to size selection (350 – 800 bp) via Pippin Prep (Sage Scientific) and then sequenced across four lanes of the HiSeq 4000 at 2x100 bp (one hundred base pair paired-end reads) at the Vincent J. Coates Genomics Sequencing Laboratory at the University of California, Berkeley.

Pre-analysis data preparation

Quality trimming and adaptor removal were performed with Trimmomatic (version 0.36)[81]. Initial read quality was high, so in order to preserve length of reads (and therefore increase coverage), the MAXINFO option was applied. The ILLUMINACLIP option was used in two passes, first to find and remove adaptor and primer sequences, and then to remove poly(A/T) sequences at the beginning or end of reads, as these are expected to complicate *de novo* assembly[77]. Ribosomal reads were filtered using Bowtie2 based on a curated set of species-

specific ribosomal RNA contigs, identified by BLAST against first-draft *de novo* assemblies of *E. carinata* and *H. vitripennis* transcriptomes[82] (SRA Accession SRP152991). Tests of further filtering, such as attempting to remove contaminating reads using FastQScreen[83] showed no impact on transcriptome assembly. Decontamination was instead performed at the annotation step described below.

Assembly of species reference transcriptomes

From each species, we selected the biological replicate that had the largest absolute number of reads post-quality control. Reads for each library (i.e., a single body region) were combined across the four lanes of sequencing and assembled with Trinity v. 2.5.1[84]. The software's default parameters were used except for increasing max_pairs_distance to 800 to reflect our actual library fragment distributions and setting min_contig length to 400. Reads for each library were aligned and transcript abundances estimated using the Bowtie[85] (version 1.1.2)/RSEM[86] (version 1.3.0) pipeline script included with Trinity. We anticipated that the draft assemblies would include some spurious transcripts derived from assembly artifacts (i.e., chimeric contigs), and would likely include some transcripts derived from contaminant sources such as bacterial endosymbionts, plant material, or fungal spores. We used the EnTAP pipeline[82] for annotation of our assemblies, which includes a feature that detects and discards these unwanted contigs. Briefly, this pipeline performs expression filtering (0.5 FPKM threshold), reading-frame selection with GeneMarkS-T (beta version)[87], annotation via similarity searching (DIAMOND version 0.4.7[88]), and protein family assignment and Gene Ontology (GO) term annotation with eggNOG[89]. For similarity searching, we used the Swiss-Prot database[90] (release date 1 January 2018), a set of predicted proteins for the related species *Nilaparvata lugens*[91]

435 (Hemiptera: Delphacidae) (NCBI BioProject accession PRJNA260223), and a curated set of
436 *Drosophila* proteins from UniRef90[90] (release date 1 January 2018). Fungal, plant, and bacterial
437 contaminants were filtered out by similarity annotation and orthogroup identification.
438 The resulting eight predicted proteomes (annotated, decontaminated amino acid
439 translations) for each species were clustered at a 90% identity threshold with the USEARCH
440 cluster_fast algorithm[92]. This high threshold was selected to retain true splicing isoforms, while
441 collapsing identical proteins across all libraries. The centroid sequences were selected as the
442 most representative proteins of each cluster, and their corresponding nucleotide sequences were
443 selected to create the refined reference assembly for each species. Completeness of these refined
444 assemblies was assessed using BUSCO (version 3.0.2)[35] with the Insecta lineage dataset
445 (insecta_odb9, release date 13 February 2016).
446
447 *Single-species differential expression analyses*
448 We analyzed differential expression of body regions within each species using the Perl
449 and R scripts distributed with Trinity 2.4[84], which are designed to construct hierarchical
450 clustering diagrams. Using these scripts, trimmed and filtered reads from each library were
451 mapped to the reference assembly for their species using Bowtie. The resulting alignments were
452 processed through RSEM, which converts the total number of aligned reads to an estimated
453 number of transcripts in the library while accounting for transcript length and total number of
454 reads in the library. The count data for all libraries were analyzed for differential expression
455 using DESeq2 with a false discovery rate of 0.1 (its default)[36]. We performed pairwise
456 comparisons between each pair of body regions, using a 4 log-fold change cut-off and a p-
457 adjusted cutoff of 0.001 (the default settings for the Trinity scripts). DESeq2 fits a generalized

458 linear model for each gene, and performs a Wald test for significant log-fold change in
459 expression between conditions[36] (here, body regions). The resulting set of differentially
460 expressed transcripts was used to derive a Euclidean distance matrix for clustering analysis using
461 the hclust function and principal component analysis using the prcomp function (R Core stats v
462 3.5.1)[93]. Approximately unbiased branch supports for the resulting dendrogram were calculated
463 using the R package pvclust (version 2.0-0)[94] with 10 resampling scales (0.5–1.4 incrementing
464 by 0.1) and 1000 bootstrap replicates at each scale.

465 This analysis workflow was repeated for each species on three subsets of transcripts that
466 were selected based on functional annotation: transcription factor activity (GO:0003700),
467 anatomical structure development (GO:0048856), and signaling (GO:0023052). Additionally, it
468 was repeated on a set of seven treehopper body regions, excluding the T2 wings.

469

470 *Multispecies differential expression analyses*

471 Multispecies transcriptome comparisons are challenging due to a confounding species
472 signal that can result in more similar gene expression between the divergent body regions within
473 a species than between the homologous body regions of different species[30,31,95]. To address this
474 challenge, we restricted our analyses to single-copy orthologues shared by our leafhopper and
475 treehopper species, a common approach[96–98]. We curated this set with OrthoFinder[37] using the
476 isoform-level proteomes predicted by EnTAP. We filtered the TPM-normalized matrices down
477 to only these 7,635 single-copy orthologues using custom R scripts. We calculated a TPM
478 scaling factor to account for the different number of mapped transcripts for each species as
479 described in reference 30; further details and the R code are provided in the supplemental
480 materials. To produce the multispecies analyses, we followed the same pipeline as above for

single-species analysis (including hierarchical clustering, bootstrap resampling, and PCA), with one exception. Instead of using the raw count matrix as input to DESeq2, we used the scaled TPM matrix. Our sample design matrix modeled each sample as a replicate of the body region, regardless of species. This allowed for a pairwise differential expression analysis that emphasized difference between body regions rather than differences between species. Given our evidence that different data subsets can have different clustering patterns, we also confirmed that the single-species clustering patterns for the set of the single-copy orthologues matched those from the full datasets by repeating the differential expression pipeline for each species individually using the scaled TPM matrix of the single-copy orthologues as input (Supplementary Fig. 8–9).

The set of orthologues included some with a strong species signal. To winnow out the transcripts that primarily differed in expression between species rather than between body regions, we used a classification strategy employing a Poisson log-linear discriminant analysis (PLDA)[38,99] in the R package MLSeq (version 1.20.3)[100]. This approach was developed to find biological markers in RNA-seq data for cancer or other diseases, but here we applied it to determine which transcripts were the best species-specific markers and then exclude them from further analysis. Our PLDA model identified 356 species markers (discrete control parameters: tuneLength=30, method=repeatedcv, number=30, repeats=10000). We used the same pipeline described above to identify differentially expressed genes from the TPM matrix for the 7,279 remaining orthologues. We performed hierarchical clustering, bootstrap resampling, and principal components analysis on the differential expression results.

503 *Gene identity and ontology analysis*

504 Sets of genes upregulated in wings (fore- and hind wings), pronotum, wings plus
505 pronotum, and abdominal terga were selected for Gene Ontology (GO) enrichment analysis. We
506 identified sets of genes by parsing the results of within-species pairwise comparisons. For each
507 body region (e.g., hind wings) we identified all of the genes that were significantly upregulated
508 in it relative to any other body region (p-adjusted < 0.001). We used the union function of the R
509 package dplyr[101] (version 0.7.8) to find all transcripts upregulated in one body region relative to
510 any other body region, and then used the intersect function to find the set of transcripts
511 upregulated in common across two or more body regions.

512 GO term enrichment analysis was performed using the R package GoSeq (version
513 1.32.0)[102] from Bioconductor. We curated a background for each of our species based on the GO
514 terms assigned by eggNOG during the EnTAP annotation process. Scripts and steps used for this
515 process are posted to https://github.com/fishercera/TreehopperSeq. Briefly, having identified a
516 set of transcripts that related to some body region or set of body regions (as described above), we
517 fitted the transcript length data to a probability weighting function. This accounts for selection
518 biases arising from gene length. We then used the goseq function to calculate over- or
519 underrepresentation for each GO term annotated in our transcript set relative to that term's
520 representation in the background. Because we found that goseq can be sensitive to small
521 transcript-set sizes, we used simulations to identify an appropriate p-value cut-off. Based on
522 these, we selected a p-value cut-off of 0.005.

523

Wing-related candidate gene expression survey

To determine whether genes of known relevance to wing development were among those upregulated in the treehopper helmet, we selected a set of wing-related candidate genes and assessed their expression in other body regions. We read the annotations for all of the transcripts differentially upregulated in the wings of treehoppers or leafhoppers, and selected those that were annotated with terms or descriptions related to wing development. We further refined the list by reviewing phenotype studies in *Drosophila*, keeping genes that had known roles in wing blade, vein, hinge, or notum formation (Fig. 6). Initial orthology assignments were based on the EnTAP annotations. Orthology to the candidate *Drosophila* genes was established first by reciprocal best BLAST hit [103,104]. Where reciprocal BLAST was not determinative, OrthoFinder[37] was used in a 5-species comparison with the input being the *Drosophila* proteome from UniProt[90], the annotated proteins of *Tribolium castaneum* from RefSeq (BioProject accession number PRJNA15718), the official gene set for *Oncopeltus fasciatus* from the USDA NAL i5k project[105,106], and the predicted proteomes from *E. carinata* and *H. vitripennis*. Orthogroup membership, alignment of conserved protein motifs, and gene trees were all used as evidence in determining orthology. See Supplementary Table 5 for the list of gene names and transcript IDs for each species.

Expression profiles for these genes were compared across body regions in each species. DESeq2 was run on counts to estimate size normalization factors and common dispersion. Because not all of the candidate wing genes are differentially expressed in both species, a variance stabilizing transformation, rather than a centered log2 transformation, was used to normalize counts for visualization.

Literature cited

1. Wagner, G. P. *Homology, Genes, and Evolutionary Innovation*. (Princeton University Press, 2014).
2. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008).
3. Davidson, E. H. & Erwin, D. H. Gene regulatory networks and the evolution of animal body plans. *Science* **311**, 796–800 (2006).
4. Gellon, G. & McGinnis, W. Shaping animal body plans in development and evolution by modulation of *Hox* expression patterns. *BioEssays* **20**, 116–125 (1998).
5. Refki, P. N., Armisén, D., Crumière, A. J. J., Viala, S. & Khila, A. Emergence of tissue sensitivity to Hox protein levels underlies the evolution of an adaptive morphological trait. *Developmental Biology* **392**, 441–453 (2014).
6. Emlen, D. J., Szafran, Q., Corley, L. S. & Dworkin, I. Insulin signaling and limb-patterning: candidate pathways for the origin and evolutionary diversification of beetle 'horns'. *Heredity* **97**, 179–191 (2006).
7. Monteiro, A. Gene regulatory networks reused to build novel traits. *BioEssays* **34**, 181–6 (2011).
8. Martin, A. *et al.* Multiple recent co-options of Optix associated with novel traits in adaptive butterfly wing radiations. *EvoDevo* **5**, 1–9 (2014).
9. Tomoyasu, Y., Arakane, Y., Kramer, K. J. & Denell, R. E. Repeated Co-options of Exoskeleton Formation during Wing-to-Elytron Evolution in Beetles. *Current Biology* **19**, 2057–2065 (2009).

10. Khila, A., Abouheif, E. & Rowe, L. Function, Developmental Genetics, and Fitness Consequences of a Sexually Antagonistic Trait. *Science* **336**, 585–589 (2012).
11. Emlen, D. J., Corley Lavine, L. & Ewen-Campen, B. On the origin and evolutionary diversification of beetle horns. *Proceedings of the National Academy of Sciences* **104**, 8661–8668 (2007).
12. Warren, I. A. *et al.* Insights into the Development and Evolution of Exaggerated Traits Using *De Novo* Transcriptomes of Two Species of Horned Scarab Beetles. *PLoS ONE* **9**, (2014).
13. Moczek, A. P. Pupal remodeling and the development and evolution of sexual dimorphism in horned beetles. *The American Naturalist* **168**, 711–729 (2006).
14. Shiga, Y. *et al.* Repeated co-option of a conserved gene regulatory module underpins the evolution of the crustacean carapace, insect wings and other flat outgrowths. *bioRxiv* **4026**, 160010 (2017).
15. Niwa, N. *et al.* Evolutionary origin of the insect wing via integration of two developmental modules. *Evolution and Development* **12**, 168–176 (2010).
16. Crampton, G. The phylogenetic origin and the nature of the wings of insects according to the paranotal theory. *Journal of the New York Entomological Society* **24**, 1–39 (1916).
17. Tomoyasu, Y. Evo-devo: the double identity of insect wings. *Current Biology* **28**, R75–R77 (2018).
18. Dietrich, C. H. *et al.* Anchored hybrid enrichment-based phylogenomics of leafhoppers and treehoppers (Hemiptera: Cicadomorpha: Membracoidea). *Insect Systematics and Diversity* **1**, 57–72 (2017).

19. Stegmann, U. E. An exaggerated trait in insects: The prothoracic skeleton of *Stictocephala bisonia* (Homoptera: Membracidae). *Journal of Morphology* **238**, 157–178 (1998).

20. Mikó, I. *et al.* On dorsal prothoracic appendages in treehoppers (Hemiptera: Membracidae) and the nature of morphological evidence. *PLoS ONE* **7**, e30137 (2012).

21. Yoshizawa, K. The treehopper's helmet is not homologous with wings (Hemiptera: Membracidae). *Systematic Entomology* **37**, 2–6 (2012).

22. McKamey, S. H. *Taxonomic Catalogue of the Membracoidea (exclusive of Leafhoppers): Second Supplement to Fascicle 1, Membracidae, of the General Catalogue of the Hemiptera*. (American Entomological Institute, 1998).

23. Evangelista, O., Sakakibara, A. M., Cryan, J. R. & Urban, J. M. A phylogeny of the treehopper subfamily Heteronotinae reveals convergent pronotal traits (Hemiptera: Auchenorrhyncha: Membracidae). *Systematic Entomology* **42**, 410–428 (2017).

24. Matsuda, R. Morphology and evolution of the insect thorax. *Memoirs of the Entomological Society of Canada* **102**, 5–431 (1970).

25. Snodgrass, R. E. *Morphology and Mechanism of the Insect Thorax*. *Smithsonian Miscellaneous Collections* vol. 80 (The Smithsonian Institution, 1927).

26. Govind, C. K. & Dandy, J. W. T. The thoracic mechanism of the milkweed bug, *Oncopeltus fasciatus* (Heteroptera: Lygaeidae). *The Canadian Entomologist* **102**, 1057–1074 (1970).

27. Prud'Homme, B. *et al.* Body plan innovation in treehoppers through the evolution of an extra wing-like appendage. *Nature* **473**, 83–86 (2011).

28. Moczek, A. P. & Nagy, L. M. Diverse developmental mechanisms contribute to different levels of diversity in horned beetles. *Evolution and Development* **7**, 175–185 (2005).

29. Moczek, A. P. & Rose, D. J. Differential recruitment of limb patterning genes during development and diversification of beetle horns. *Proceedings of the National Academy of Sciences* **106**, 8992–8997 (2009).

30. 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).

31. Liang, C., Musser, J. M., Cloutier, A., Prum, R. O. & Wagner, G. P. Pervasive correlated evolution in gene expression shapes cell and tissue type transcriptomes. *Genome Biology and Evolution* **10**, 538–552 (2018).

32. Chanderbali, A. S. *et al.* Conservation and canalization of gene expression during angiosperm diversification accompany the origin and evolution of the flower. *Proceedings of the National Academy of Sciences* **107**, 22570–22575 (2010).

33. Chanderbali, A. S. *et al.* Transcriptional signatures of ancient floral developmental genetics in avocado (*Persea americana*; Lauraceae). *Proceedings of the National Academy of Sciences of the United States of America* **106**, 8929–8934 (2009).

34. Glassford, W. J. *et al.* Co-option of an ancestral Hox-regulated network underlies a recently evolved morphological novelty. *Developmental Cell* **34**, 520–531 (2015).

35. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

36. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 1–21 (2014).

636 37. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome
637 comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**,
638 157 (2015).
639 38. Witten, D. M. Classification and clustering of sequencing data using a poisson model.
640 *Annals of Applied Statistics* **5**, 2493–2518 (2011).
641 39. McCune, A. R. & Schimenti, J. C. Using genetic networks and homology to understand the
642 evolution of phenotypic traits. *Current Genomics* **13**, 74–84 (2012).
643 40. Wagner, G. P. The developmental genetics of homology. *Nature Reviews Genetics* **8**, 191–
644 206 (2007).
645 41. Mao, M., Yang, X., Poff, K. & Bennett, G. Comparative Genomics of the Dual-Obligate
646 Symbionts from the Treehopper, *Entylia carinata* (Hemiptera: Membracidae), Provide
647 Insight into the Origins and Evolution of an Ancient Symbiosis. *Genome Biology and*
648 *Evolution* **9**, 1803–1815 (2017).
649 42. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics*
650 **25**, 25–29 (2000).
651 43. The Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and
652 resources: The gene ontology consortium. *Nucleic Acids Research* **45**, D331–D338 (2017).
653 44. Jockusch, E. L. & Smith, F. W. Hexapoda: Comparative Aspects of Later Embryogenesis
654 and Metamorphosis. in *Evolutionary Developmental Biology of Invertebrates 5* 111–208
655 (Springer Vienna, 2015). doi:10.1007/978-3-7091-1868-9_3.
656 45. Lee, H. & Adler, P. N. The *grainy head* transcription factor is essential for the function of
657 the *frizzled* pathway in the *Drosophila* wing. *Mechanisms of Development* **121**, 37–49
658 (2004).

659 46. Zecca, M. & Struhl, G. Subdivision of the *Drosophila* wing imaginal disc by EGFR-
660 mediated signaling. *Development* **129**, 1357–1368 (2002).
661 47. Hidalgo, A. Three distinct roles for the *engrailed* gene in *Drosophila* wing development.
662 *Current Biology* **4**, 1087–1098 (1994).
663 48. O'Keefe, D. D. & Thomas, J. B. *Drosophila* wing development in the absence of dorsal
664 identity. *Development* **128**, 703–710 (2001).
665 49. Cohen, B., McGuffin, M. E., Pfeifle, C., Segal, D. & Cohen, S. M. *apterous*, a gene
666 required for imaginal disc development in *Drosophila* encodes a member of the LIM family
667 of developmental regulatory proteins. *Genes & Development* **6**, 715–729 (1992).
668 50. Halder, G. *et al.* The Vestigial and Scalloped proteins act together to directly regulate wing-
669 specific gene expression in *Drosophila*. *Genes and Development* **12**, 3900–3909 (1998).
670 51. Cho, E. & Irvine, K. D. Action of *fat*, *four-jointed*, *dachsous* and *dachs* in distal-to-
671 proximal wing signaling. *Development* **131**, 4489–4500 (2004).
672 52. Sato, M. & Saigo, K. Involvement of *pannier* and *u-shaped* in regulation of
673 Decapentaplegic-dependent *wingless* expression in developing *Drosophila* notum.
674 *Mechanisms of Development* **93**, 127–138 (2000).
675 53. Tomoyasu, Y., Ueno, N. & Nakamura, M. The Decapentaplegic morphogen gradient
676 regulates the notal *wingless* expression through induction of *pannier* and *u-shaped* in
677 Drosophila. *Mech Dev* **96**, 37–49 (2000).
678 54. St. Pierre, S. E., Galindo, M. I., Couso, J. P. & Thor, S. Control of *Drosophila* imaginal
679 disc development by *rotund* and *roughened eye*: differentially expressed transcripts of the
680 same gene encoding functionally distinct zinc finger proteins. *Development* **129**, 1273–
681 1281 (2002).

55. Furriols, M. & Bray, S. A model Notch response element detects Suppressor of Hairless-dependent molecular switch. *Current Biology* **11**, 60–64 (2001).

56. Roch, F., Alonso, C. R. & Akam, M. *Drosophila miniature* and *dusky* encode ZP proteins required for cytoskeletal reorganisation during wing morphogenesis. *Journal of Cell Science* **116**, 1199–1207 (2003).

57. Fristrom, D. *et al.* Blistered: a gene required for vein/intervein formation in wings of Drosophila. *Development (Cambridge, England)* **120**, 2661–71 (1994).

58. Montagne, J. *et al.* The *Drosophila* Serum Response Factor gene is required for the formation of interveing tissue of the wing and is allelic to *blistered*. *Development* **122**, 2589–2597 (1996).

59. Diez del Corral, R., Aroca, P., Gómez-Skarmeta, J. L., Cavodeassi, F. & Modolell, J. The iroquois homeodomain proteins are required to specify body wall identity in *Drosophila*. *Genes and Development* **13**, 1754–1761 (1999).

60. Gömez-Skarmeta, J. L., Diez del Corral, R., de la Calle-Mustienes, E., Ferrés-Marcó, D. & Modolell, J. *araucan* and *caupolican*, two members of the novel iroquois complex, encode homeoproteins that control proneural and vein-forming genes. *Cell* **85**, 95–105 (1996).

61. Ikmi, A., Netter, S. & Coen, D. Prepatterning the Drosophila notum: The three genes of the iroquois complex play intrinsically distinct roles. *Developmental Biology* **317**, 634–648 (2008).

62. Clark-Hachtel, C. M., Linz, D. M. & Tomoyasu, Y. Insights into insect wing origin provided by functional analysis of *vestigial* in the red flour beetle, *Tribolium castaneum*. *Proceedings of the National Academy of Sciences* **110**, 16951–16956 (2013).

63. Medved, V. *et al.* Origin and diversification of wings: Insights from a neopteran insect. *Proceedings of the National Academy of Sciences* **112**, 15946–15951 (2015).

64. Elias-Neto, M. & Belles, X. Tergal and pleural structures contribute to the formation of ectopic prothoracic wings in cockroaches. *Royal Society Open Science* **3**, 160347 (2016).

65. Yao, L. *et al.* Genome-wide identification of Grainy head targets in *Drosophila* reveals regulatory interactions with the POU domain transcription factor Vvl. *Development (Cambridge, England)* **144**, 3145–3155 (2017).

66. Tomoyasu, Y., Wheeler, S. R. & Denell, R. E. *Ultrabithorax* is required for membranous wing identity in the beetle *Tribolium castaneum*. *Nature* **433**, 643–647 (2005).

67. Grimm, S. & Pflugfelder, G. O. Control of the gene *optomotor-blind* in *Drosophila* wing development by decapentaplegic and wingless. *Science* **271**, 1601–1604 (1996).

68. Kopp, A. & Duncan, I. Anteroposterior patterning in adult abdominal segments of Drosophila. *Developmental Biology* **242**, 15–30 (2002).

69. Brunner, E., Peter, O., Schweizer, L. & Basler, K. *pangolin* encodes a Lef-1 homologue that acts downstream of Armadillo to transduce the Wingless signal in *Drosophila*. *Nature* **385**, 829–833 (1997).

70. Hu, Y. *et al.* A morphological novelty evolved by co-option of a reduced gene regulatory network and gene recruitment in a beetle. *Proceedings of the Royal Society B: Biological Sciences* **285**, 20181373 (2018).

71. Brook, W. J., Diaz-Benjumea, F. J. & Cohen, S. M. ORGANIZING SPATIAL PATTERN IN LIMB DEVELOPMENT. *Annual Review of Cell and Developmental Biology* **12**, 161–80 (1996).

72. Clark-Hachtel, C. M. & Tomoyasu, Y. Exploring the origin of insect wings from an evo-devo perspective. *Current Opinion in Insect Science* **13**, 77–85 (2016).

73. Ohde, T., Yaginuma, T. & Niimi, T. Insect morphological diversification through the modification of wing serial homologs. *Science* **340**, 495–498 (2013).

74. Tomoyasu, Y., Ohde, T. & Clark-Hachtel, C. M. What serial homologs can tell us about the origin of insect wings. *F1000Research* **6**, 268 (2017).

75. Stegmann, U. E. Revaluation of the prothoracic pleuron of the Membracidae (Homoptera): The presence of an epimeron and a subdivided episternum in *Stictocephala bisonia* Kopp and Yonke, *Oxyrchachis taranda* (Fabr.), and *Centrotus cornutus* (L.). *International Journal of Insect Morphology and Embryology* **26**, 35–42 (1997).

76. Kyrkou, I., Pusa, T., Ellegaard-Jensen, L., Sagot, M. & Hansen, L. H. Pierce's disease of grapevines: a review of control strategies and an outline of an epidemiological model. *Frontiers in Microbiology* **9**, 2141 (2018).

77. Wolf, J. B. W. Principles of transcriptome analysis and gene expression quantification: An RNA-seq tutorial. *Molecular Ecology Resources* **13**, 559–572 (2013).

78. Mito, T., Nakamura, T. & Noji, S. Evolution of insect development: To the hemimetabolous paradigm. *Current Opinion in Genetics and Development* **20**, 355–361 (2010).

79. Lambert, K. N. & Williamson, V. M. cDNA library construction from small amounts of RNA using paramagnetic beads and PCR. in *cDNA Library Protocols* (eds. Cowell, I. G. & Austin, C. A.) vol. 69 1–12 (Springer Humana Press, 1997).

80. Costello, M. *et al.* Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* **19**, 332 (2018).

81. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

82. Hart, A. J. *et al.* EnTAP: Bringing Faster and Smarter Functional Annotation to Non-Model Eukaryotic Transcriptomes. *Mol Ecol Resour* 1755–0998.13106 (2019) doi:10.1111/1755-0998.13106.

83. Wingett, S. W. & Andrews, S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research* **7**, 1338 (2018).

84. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644–652 (2011).

85. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25 (2009).

86. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

87. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Research* **43**, e78 (2015).

88. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59–60 (2015).

89. Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research* **44**, D286–D293 (2016).

90. The Uniprot Consortium. UniProt: The universal protein knowledgebase. *Nucleic Acids Research* **45**, D158–D169 (2017).

91. Li, K.-Y. *et al.* Wing patterning genes of *Nilaparvata lugens* identification by transcriptome analysis, and their differential expression profile in wing pads between brachypterous and macropterous morphs. *Journal of Integrative Agriculture* **14**, 1796–1807 (2015).

92. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

93. R Core Team. R: A Language and Environment for Statistical Computing. (2018).

94. Suzuki, R. & Shimodaira, H. pvclust: Hierarchical clustering with p-values via multiscale bootstrap resampling. (2015).

95. Sudmant, P. H., Alexis, M. S. & Burge, C. B. Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biology* **16**, 1–11 (2015).

96. Plaza, D. F., Lin, C. W., van der Velden, N. S. J., Aebi, M. & Künzler, M. Comparative transcriptomics of the model mushroom *Coprinopsis cinerea* reveals tissue-specific armories and a conserved circuitry for sexual development. *BMC Genomics* **15**, 1–17 (2014).

97. Zhou, Z. C. *et al.* Transcriptome sequencing of sea cucumber (*Apostichopus japonicus*) and the identification of gene-associated markers. *Molecular Ecology Resources* **14**, 127–138 (2014).

98. Connahs, H., Rhen, T. & Simmons, R. B. Transcriptome analysis of the painted lady butterfly, *Vanessa cardui* during wing color pattern development. *BMC Genomics* **17**, 270 (2016).

792 99. Zararsız, G. *et al.* A comprehensive simulation study on classification of RNA-Seq data.
793 *Plos One* **12**, e0182507 (2017).
794 100. Goksuluk, D. *et al. MLSeq: Machine learning interface to RNA-seq data*. (Bioconductor,
795 2018).
796 101. Wickham, H., François, R., Henry, L. & Müller, K. dplyr: A Grammar of Data
797 Manipulation. (2018).
798 102. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for
799 RNA-seq: accounting for selection bias. *Genome Biology* **11**, R14 (2010).
800 103. Ward, N. & Moreno-Hagelsieb, G. Quickly Finding Orthologs as Reciprocal Best Hits with
801 BLAT, LAST, and UBLAST: How Much Do We Miss? *PLOS ONE* **9**, e101850 (2014).
802 104. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A Genomic Perspective on Protein Families.
803 *Science* **278**, 631–637 (1997).
804 105. Panfilio, K. A. *et al.* Molecular evolutionary trends and feeding ecology diversification in
805 the Hemiptera, anchored by the milkweed bug genome. *Genome Biology* **20**, 64 (2019).
806 106. Vargas Jentzsch, I. M. *et al.* Oncopeltus fasciatus Official Gene set v1.1. (2015)
807 doi:10.15482/usda.adc/1173142.
808 107. Flynn, Dawn. Review of the Genus *Cladonota* Stål (Hemiptera: Membracidae:
809 Membracinae: Hypsoprorini) with Keys, Illustrations of Adults, and Known Nymphs, and
810 Description of a New Species from Costa Rica. I. Introduction and Subgenus *Falculifera*
811 McKamey. *Proceedings of the Entomological Society of Washington* **120**, 725–747 (2018).
812 108. Gross, L. Bacterial Symbionts May Prove a Double-Edged Sword for the Sharpshooter.
813 *PLOS Biology* **4**, e218 (2006).
814

815

Acknowledgments

Financial support for this work was provided by NSF IOS 1656572 to ELJ; CRF was supported in part by an Outstanding Scholar Fellowship from the University of Connecticut. Seed funding for pilot studies was provided to CRF by the Sigma Xi Grants in Aid of Research program, the Society for the Study of Evolution Rosemary Grant Award program, and the Connecticut Museum of Natural History. Analyses were run on the UConn Health Center High Performance Computing (HPC) cluster administered by the Computational Biology Core within the Institute for Systems Genomics. We appreciate assistance with treehopper colony maintenance provided by Matt Opel and other UConn EEB greenhouse staff, Adam Chiu and Michelle Deering. Support for library quantification and fragment analysis was provided by Bo Reese and Lu Li at the UConn Center for Genome Innovation. Valuable comments on the manuscript were provided by Charlie Henry, Carl Schlichting, Frank Smith, Yaowu Yuan. We wish to acknowledge the kind assistance of Youngsoo Son, of the California Department of Food and Agriculture Pierce's Disease Control Program, for providing *H. vitripennis* specimens. We thank Patrick Coin, Pavel Kirillov, Michael Schmidt, and Kelly Swing for granting permission to use their photographs.

Author contributions

CRF and ELJ conceived the study and designed the sampling. All authors contributed to the data analysis design. CRF reared specimens, prepared libraries, analyzed the data in consultation with ELJ and JLW, and wrote the first draft of the manuscript. CRF, ELJ and JLW revised the manuscript.

839 **Data and Code availability**

840 The raw sequencing reads and refined transcriptome assemblies that form the basis for
841 this study have been deposited in the NCBI Sequence Reads Archive (SRA) and Transcriptome
842 Shotgun Assembly (TSA) database, respectively. All data are components of NCBI BioProject
843 PRJNA415641. The transcriptome assembly for *H. vitripennis* is under accession code
844 GHXA00000000 in the TSA. The transcriptome assembly for *E. carinata* is under accession
845 code GHWZ00000000. Raw reads are attached to SRA study SRP152991 with accession codes
846 SRR9942929 through SRR9942973. Analytical code, along with test data, have been posted at
847 www.github.com/fishercera/TreehopperSeq.
848

849 **Competing interests**

850 The authors declare no competing interests.

851 **Figure legends**

852

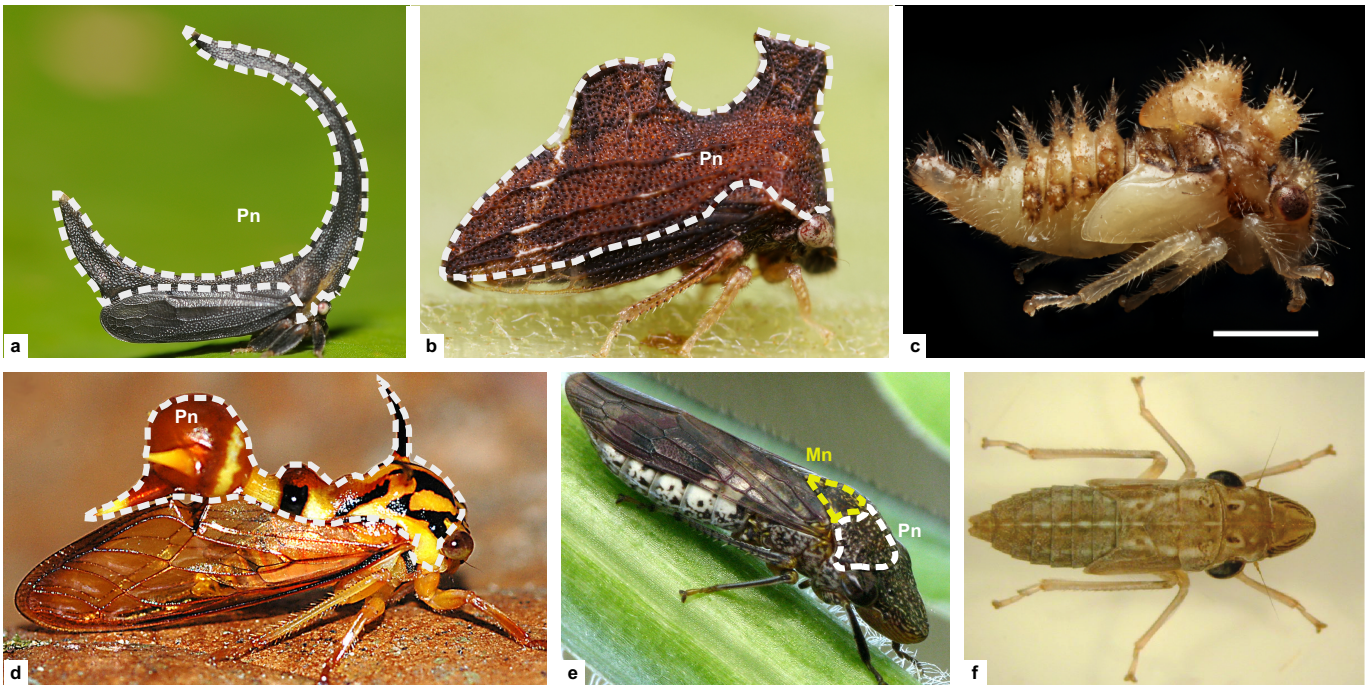
853 **Fig. 1 Pronotal morphology in treehoppers and leafhoppers.** Treehopper helmets are 3D
854 projections of the pronotum that are thought to aid in predator defense. **a,** *Cladonota apicalis,* a
855 treehopper. Typical length for this species is 6–9 mm[107]. **b**, Adult *Entylia carinata,* the
856 treehopper species used in this study. Typical length for this species is 4–5 mm. **c,** Fifth instar
857 nymphal *E. carinata*. Scale bar denotes 1 mm. **d,** *Heteronotus* sp., a treehopper with a helmet
858 that mimics a wasp. Typical length for members of this genus range from 5–10 mm[23]. **e,** Adult
859 *Homalodisca vitripennis,* the leafhopper species used in this study. Typical length for this
860 species is 10–14 mm. **f,** Fifth instar nymphal *H. vitripennis*. White dashes outline the pronotum
861 and yellow dashes outline its serial homologue the mesonotum; the pronotum entirely covers the

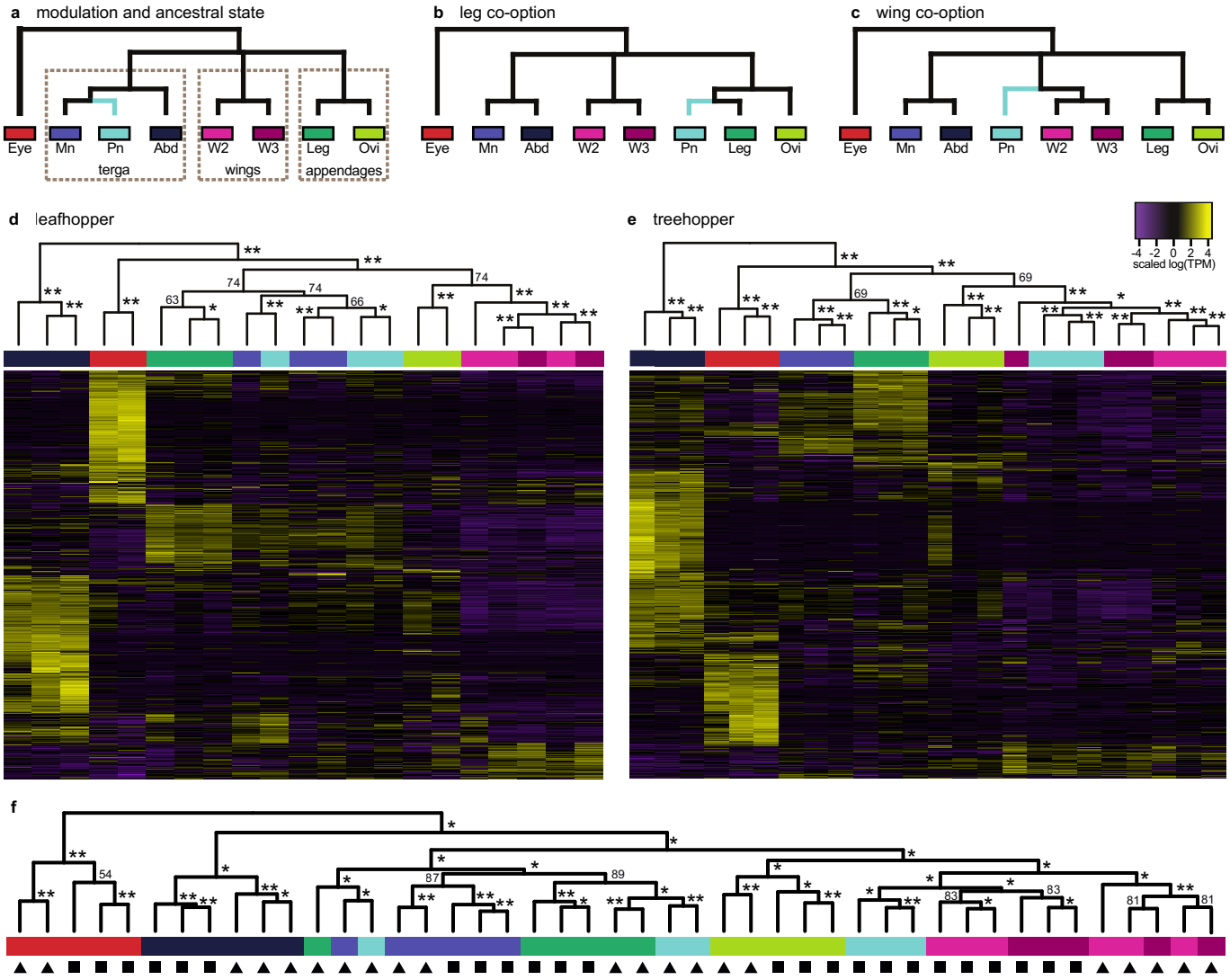
862 mesonotum in treehoppers. Mn – mesonotum; Pn – pronotum. Pavel Kirillov (**a**); Patrick Coin
863 (**b**); Kelly Swing (**d**); Michael Schmidt (**e**); Liza Gross [108] .
864
865 **Fig. 2 Character trees showing predictions and results of hierarchical clustering based on**
866 **differential gene expression. a–c,** Models of the predicted patterns of hierarchical clustering
867 based on differential gene expression for the modulation (**a**), leg-network co-option (**b**) and
868 wing-network co-option (**c**) hypotheses for the origin of the treehopper helmet; **a** is also the
869 predicted ancestral state, expected to be retained in leafhoppers. The position of the pronotum
870 differs across hypotheses, but the trees are otherwise the same. Dashed boxes in **a** group serial
871 homologues. **d,** Hierarchical clustering of samples and expression heatmap of differentially
872 expressed genes (4,428 of 17,609 features differentially expressed) in the leafhopper *H.*
873 *vitripennis*. Pronotum and mesonotum cluster. **e**, Hierarchical clustering of samples and
874 expression heatmap of differentially expressed genes (3,353 of 18,652 features differentially
875 expressed) in the treehopper *E. carinata*. Pronotum and wings cluster. **f**, Hierarchical clustering
876 of samples from both species in a combined analysis of the 1,420 differentially expressed single-
877 copy orthologues. Black squares and triangles indicate treehoppers and leafhoppers respectively.
878 In **d–f**, numbers at nodes are approximately unbiased multiscale bootstrap support values. A
879 single asterisk indicates support > 90, and two asterisks indicate support of 100. Color coding
880 and abbreviations: red, eye; dark blue, abdominal tergum (Abd); dark green, leg; blue,
881 mesonotum (Mn); cyan, pronotum (Pn); light green, ovipositor (Ovi); bright purple, forewing
882 pad (W2); maroon, hind wing page (W3).
883
884 **Fig. 3 PCA of genes differentially expressed across body regions. a**,**b**, PCA plots for PC2 (**a**)

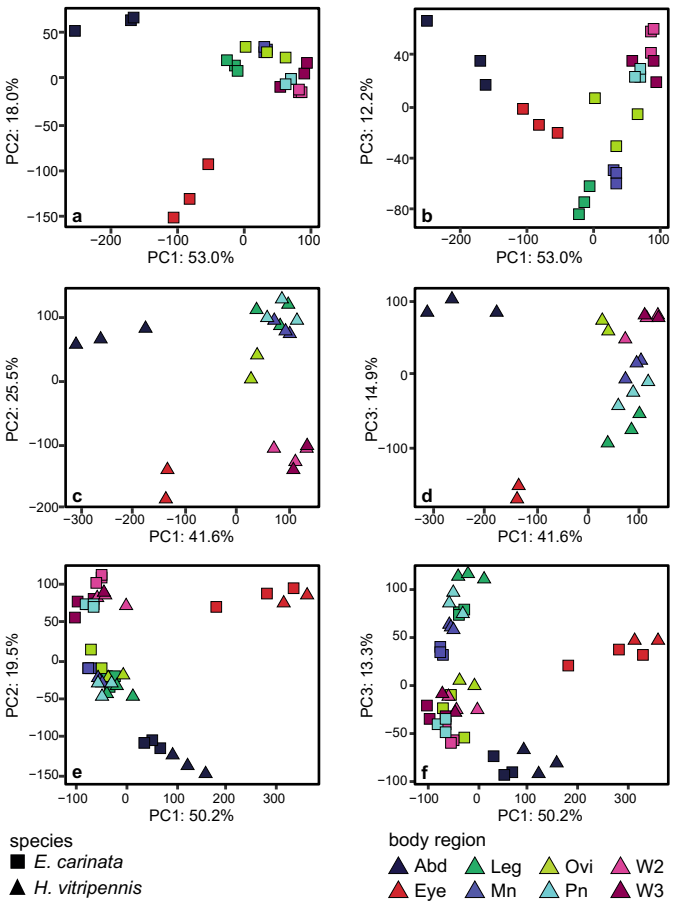
and PC3 (**b**) for treehopper (*E. carinata*) samples. Pronotum samples group with wing samples, and are distinct from mesonotum, leg, and ovipositor samples. **c**,**d**, PCA plots for PC2 (**c**) and PC3 (**d**) for leafhopper (*H. vitripennis*) samples. Pronotal samples are closest to mesonotal and leg samples, and widely separated from wing samples. **e**,**f**, PCA plots for PC2 (**e**) and PC3 (**f**) versus PC1 for the combined analysis of treehopper and leafhopper samples based on the set of single-copy orthologues. Samples from each body region, except the pronotum, cluster across species. Treehopper pronotum samples group with leafhopper and treehopper wings, while leafhopper pronotum samples group with treehopper and leafhopper mesonota, legs, and ovipositors. Data point shapes denote species: squares indicate treehoppers; triangles indicate leafhoppers. Color coding and abbreviations are used as in Fig. 2.

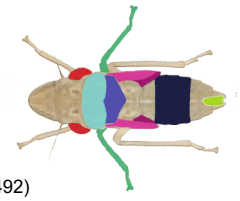
Fig. 4 Character trees based on hierarchical clustering of annotated subsets of differentially expressed genes. a,**b,** Anatomical structure development genes (GO:0048856). Leafhopper (*H. vitripennis*) mesonotum and pronotum samples are most similar to each other (**a**), while treehopper (*E. carinata*) helmets are most similar to wings (**b**). **c**,**d,** Signaling genes (GO:0023052). Leafhopper pronotum and mesonotum samples cluster with legs (**c**), while treehopper helmets are most similar to wings (**d**). **e**,**f**, Transcription factor activity genes (GO:0003700). Clustering patterns in both leafhoppers (**e**) and treehoppers (**f**) match the predictions of serial homology (Fig. 2a), with the exception that legs and ovipositors did not form a separate cluster in the leafhopper. Support values, body regions, and abbreviations are as in Fig. 2. Numbers beside labels indicate the number of features differentially expressed (DE) across body regions and the total number of features in each subset. Body regions studied are false-colored on the leafhopper (left) and treehopper (right) nymphs at top. A single asterisk

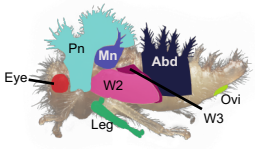
908 indicates support >90, and two asterisks indicate support of 100. Color coding and abbreviations
909 are used as in Fig. 2.
910
911 **Fig. 5 GO term enrichment for differentially expressed genes upregulated in the wings and**
912 **pronotum.** The functional annotation supports the wing-patterning network co-option
913 hypothesis. All terms enriched at a p-value < 0.005 in the indicated set of samples are included.
914 Leafhopper (*H. vitripennis*) and treehopper (*E. carinata*) wings have highly overlapping sets of
915 significantly enriched GO terms, and many of these terms are also enriched in the set of
916 transcripts upregulated in the treehopper pronotum. Enrichment patterns in the leafhopper
917 pronotum are highly divergent from enrichment patterns in the wings of either species or
918 pronotum of treehoppers. P-values for over-representedness are shown; the darkest blue values
919 have the highest significance.
920
921 **Fig. 6 Expression of candidate genes related to wing development across body regions**. **a**,**b**
922 Heatmaps depict variance stabilization-transformed counts averaged across body regions within
923 each species for a selected set of transcripts that are orthologous to genes with known wing-
924 patterning roles in *Drosophila*. Leafhopper (*H. vitripennis*) (**a**) and treehopper (*E. carinata*) (**b**)
925 expression values. Many genes that are more highly expressed in the wings of both species are
926 also more highly expressed in the pronotum of treehoppers, but not in the pronotum of
927 leafhoppers. Different color scales are used for the two species because normalization was done
928 on each species individually, and therefore expression levels cannot be directly compared across
929 species. Color coding and abbreviations are used as in Fig. 2.


Pn
a
Pn
b
c
Pn
d
Mn
Pn
e
f

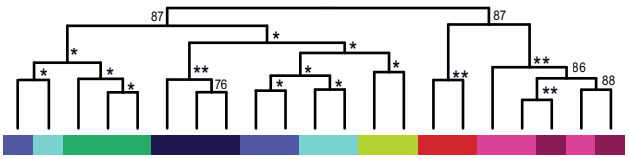

a modulation and ancestral state
Eye
Mn
Pn
Abd
W2
W3
Leg
Ovi
terga
wings
appendages
b leg co-option
c wing co-option
d leafhopper
e treehopper
-4 -2 0 2 4
scaled log(TPM)
f


PC2: 18.0%
PC1: 53.0%
a
PC3: 12.2%
PC1: 53.0%
b
PC2: 25.5%
PC1: 41.6%
c
PC3: 14.9%
PC1: 41.6%
d
PC2: 19.5%
PC1: 50.2%
e
PC3: 13.3%
PC1: 50.2%
f
species
E. carinata
H. vitripennis
body region
Abd
Leg
Ovi
W2
Eye
Mn
Pn
W3

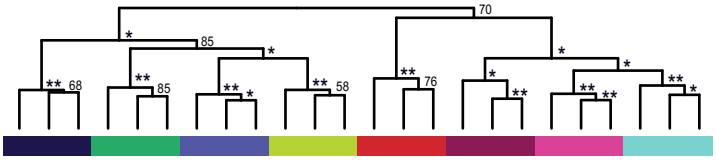




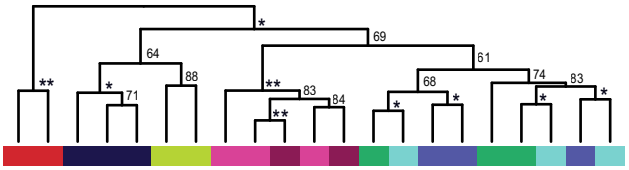
a 968 DE (of 3,492)



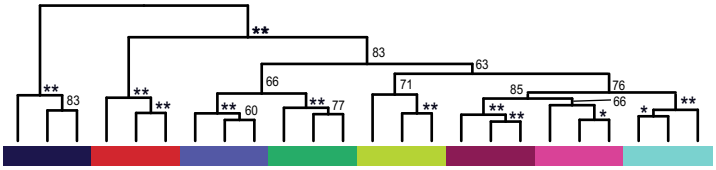
b 615 DE (of 3,496)



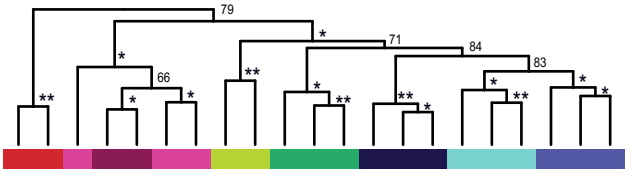
c 730 DE (of 2,228)



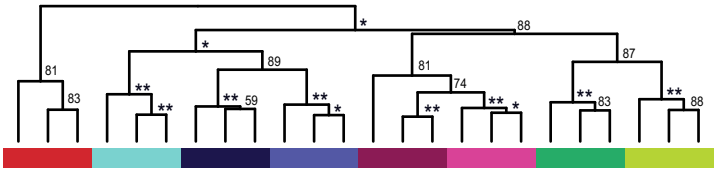
d 528 DE (of 2,300)

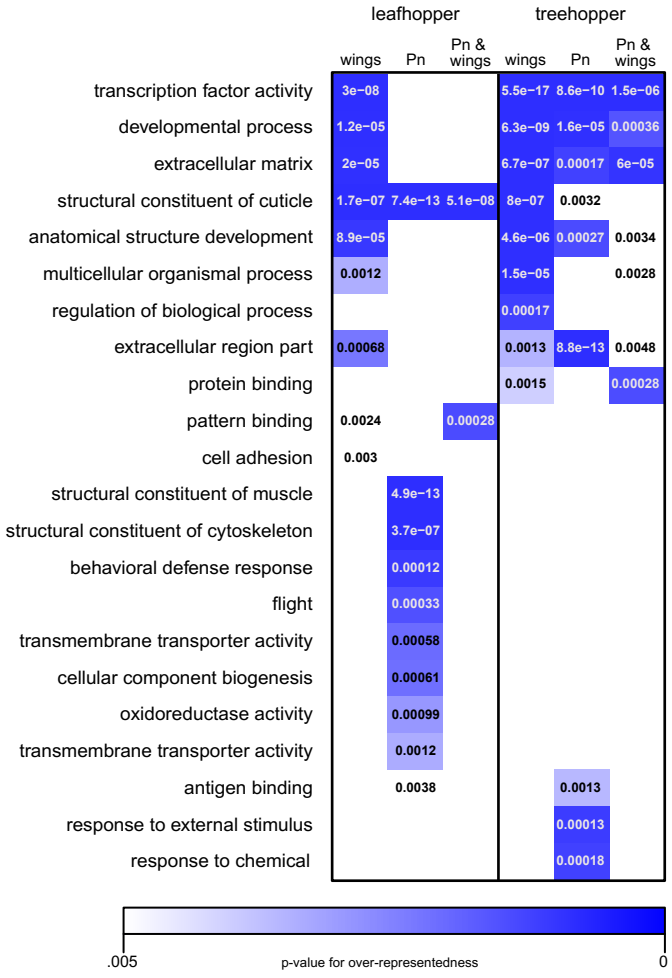


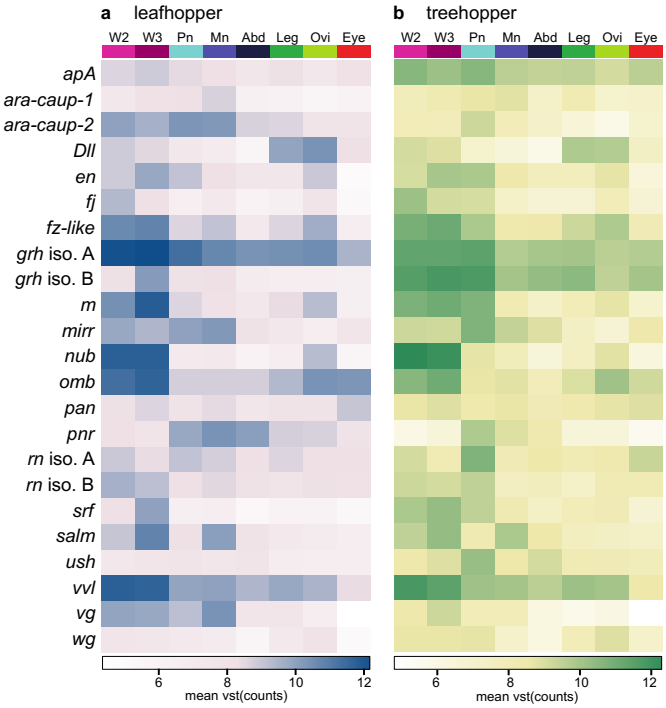
e 180 DE (of 713)



f 160 DE (of 791)




leafhopper
treehopper
wings
Pn
Pn & wings
wings
Pn
Pn & wings
transcription factor activity
developmental process
extracellular matrix
structural constituent of cuticle
anatomical structure development
multicellular organismal process
regulation of biological process
extracellular region part
protein binding
pattern binding
cell adhesion
structural constituent of muscle
structural constituent of cytoskeleton
behavioral defense response
flight
transmembrane transporter activity
cellular component biogenesis
oxidoreductase activity
transmembrane transporter activity
antigen binding
response to external stimulus
response to chemical
3e−08
5.5e−17
8.6e−10
1.5e−06
1.2e−05
6.3e−09
1.6e−05
0.00036
2e−05
6.7e−07
0.00017
6e−05
1.7e−07
7.4e−13
5.1e−08
8e−07
0.0032
8.9e−05
4.6e−06
0.00027
0.0034
0.0012
1.5e−05
0.0028
0.00017
0.00068
0.0013
8.8e−13
0.0048
0.0015
0.00028
0.0024
0.00028
0.003
4.9e−13
3.7e−07
0.00012
0.00033
0.00058
0.00061
0.00099
0.0012
0.0038
0.0013
0.00013
0.00018
.005
p-value for over-representedness
0


a leafhopper
b treehopper
W2 W3 Pn Mn Abd Leg Ovi Eye
apA
ara-caup-1
ara-caup-2
Dll
en
fj
fz-like
grh iso. A
grh iso. B
m
mirr
nub
omb
pan
pnr
rn iso. A
rn iso. B
srf
salm
ush
vvl
vg
wg
6 8 10 12
mean vst(counts)

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

R (version 3.6.1) code used to create PLDA classifier and separate species markers from other genes.

```
#~~~~ 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")
 ind <- as.vector(GoodInd)
 ind <- ind[,1]
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$OrthoID)
mcols(cts.trainS4) <- DataFrame(mcols(cts.trainS4), featureData)
mcols(cts.trainS4) <- DataFrame(mcols(cts.trainS4), data.frame(HVid=MergedCounts$HVid))
mcols(cts.trainS4) <- DataFrame(mcols(cts.trainS4), data.frame(ECid=MergedCounts$ECid))
```

```
# Set the parameters for the classifier
ctrl.PLDA <- discreteControl(method="repeatedcv", number=30, tuneLength=100, repeats=100
000, 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_tuneLength1
00.txt")
 ## Filter out the ones that are not species biased.
 UnselectedGenes <- MergedCounts[(which(!(MergedCounts$OrthoID %in% MergedCounts[Mar
kers, ]$OrthoID))), ]
 write.table(UnselectedGenes, "New_SCOs_TunedMergedCounts_UnselectedGenes_tuneLength1
00.txt", sep="\t", quote=FALSE)
```

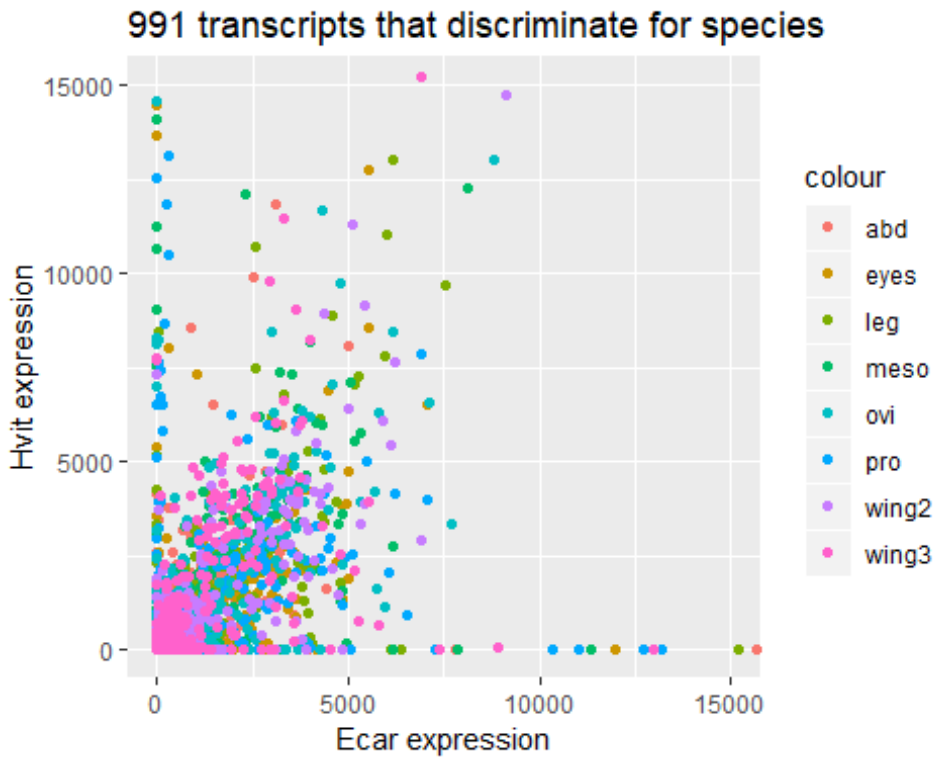
The outcome of the PLDA classifier can be seen by plotting expression levels of "marker" genes and "unbiased" genes.

Comparing selected species markers to unbiased genes

Read in selected markers, filter the counts matrices.

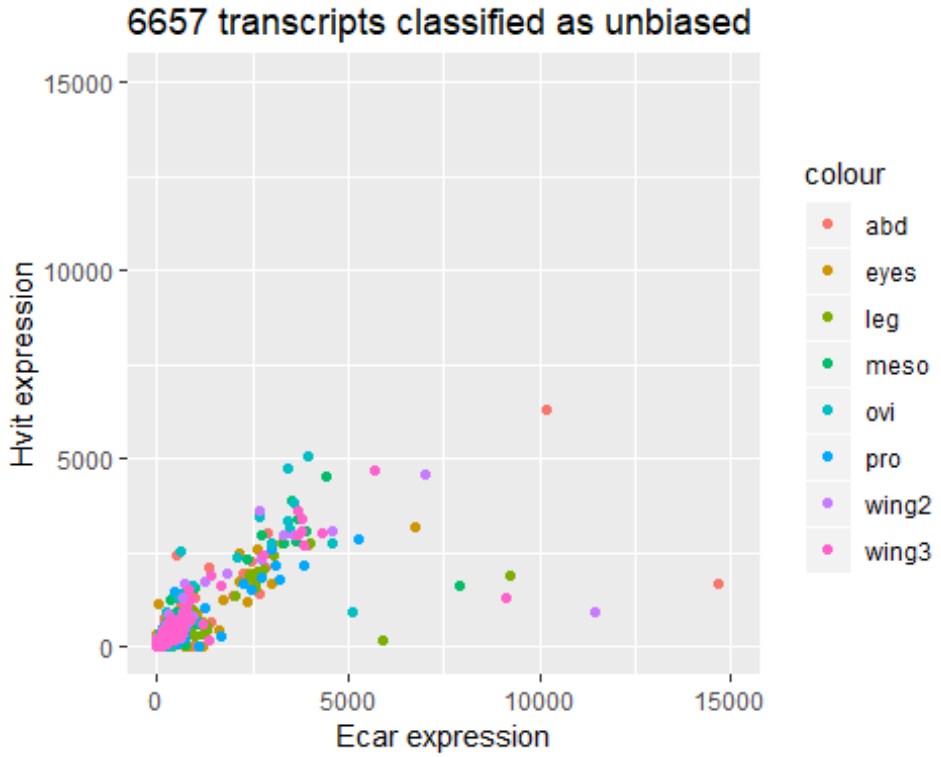
```
selectedMarkers <- read.table("1_Tuned_MergedCounts_SelectedMarkers_10000iter
.txt")
Merged.Counts <- read.table("HV_EC_NewSamples.SingleCopyOrthogroups.MergedTPM
.txt")
unselected <- filter(Merged.Counts,
                     !(Merged.Counts$OrthoID %in% selectedMarkers$OrthoID))
```

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.

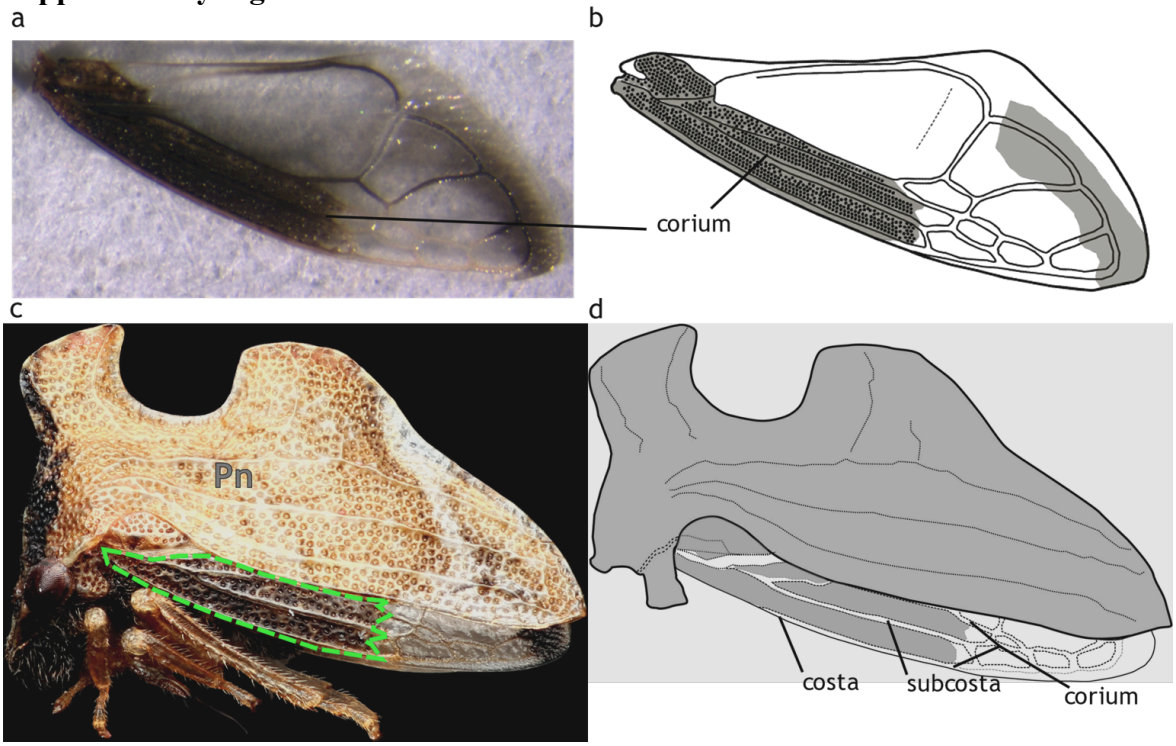
Plot *E. carinata* TPM vs. *H. vitripennis* TPM of genes not selected as species markers (unbiased genes).



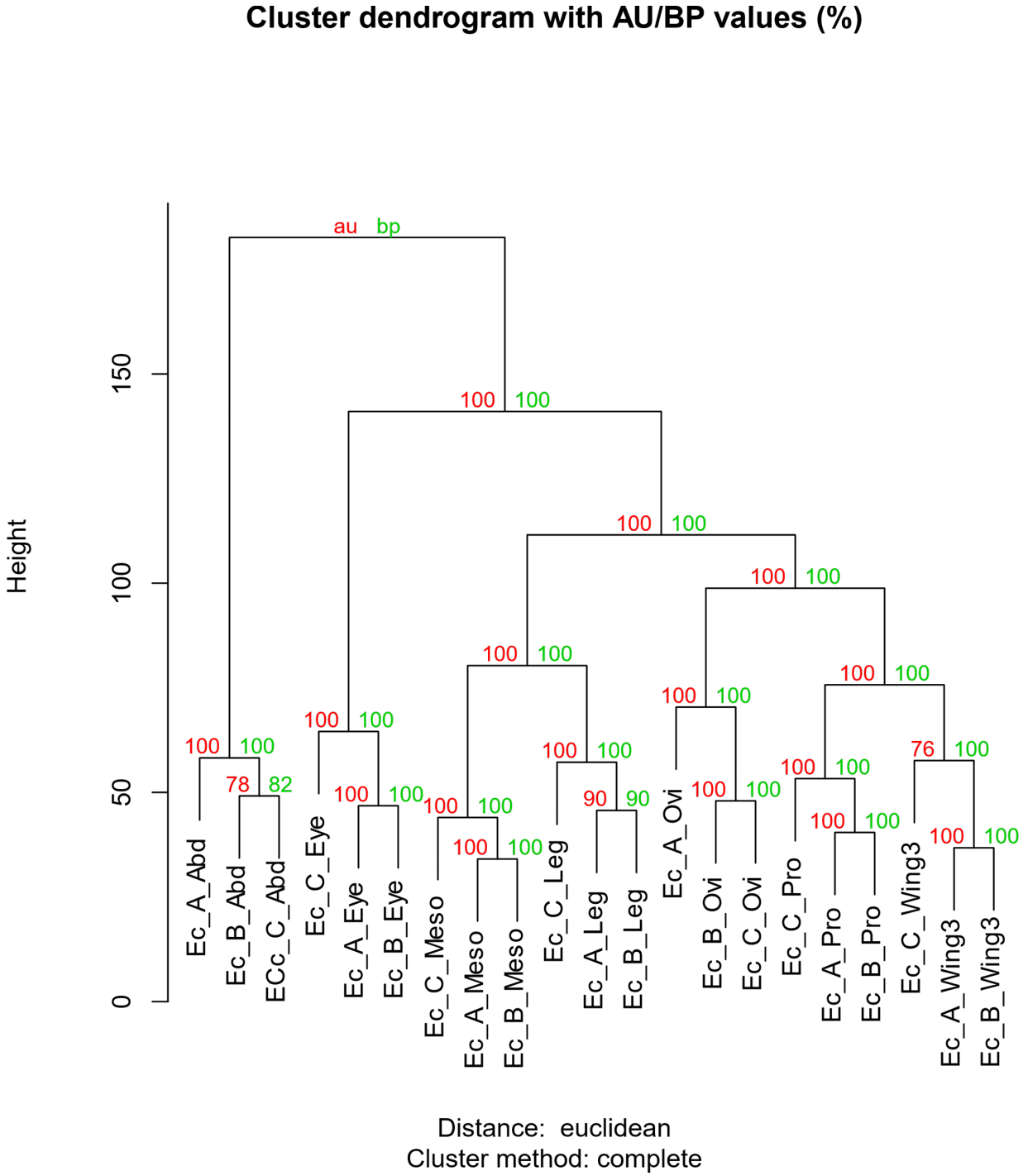
This plot does not have a trident pattern. Most genes with expression biased towards one species have been removed from the set. A countable few (10) remain with highly biased *Entylia* expression.

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

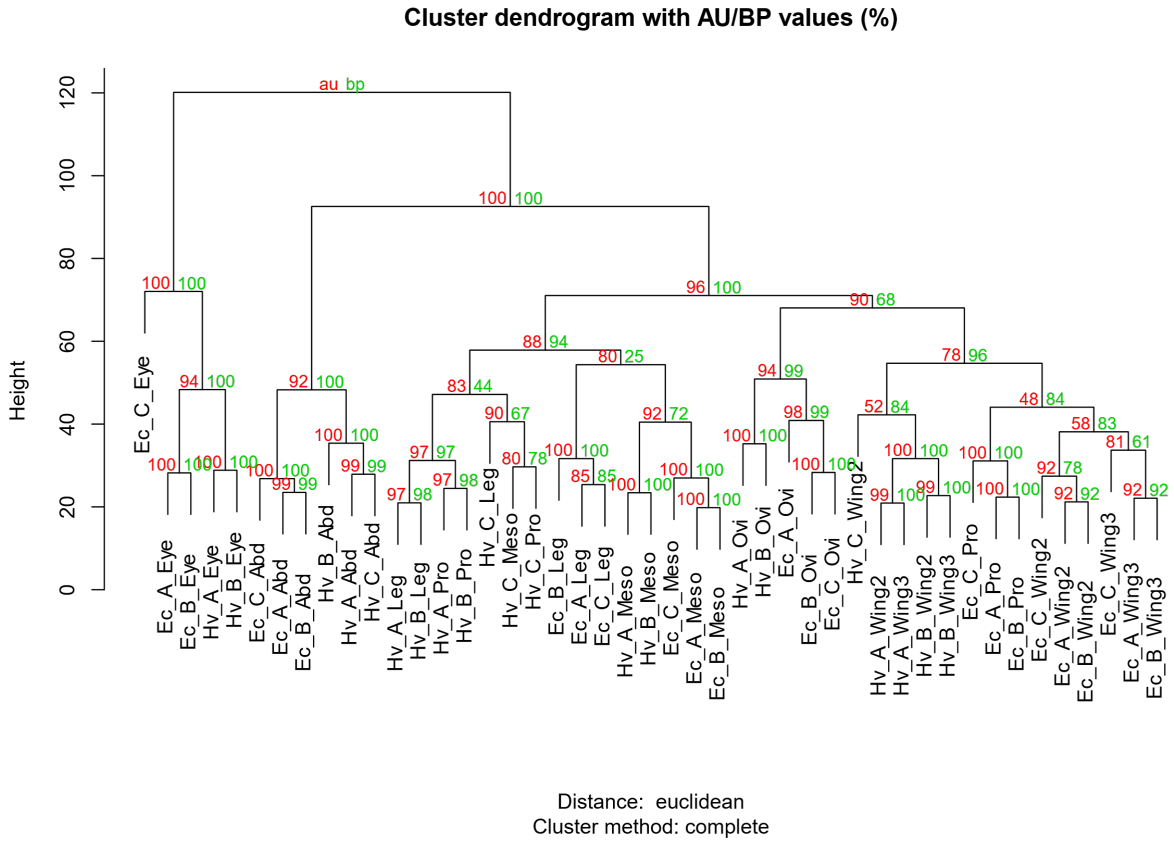
Supplementary Figures



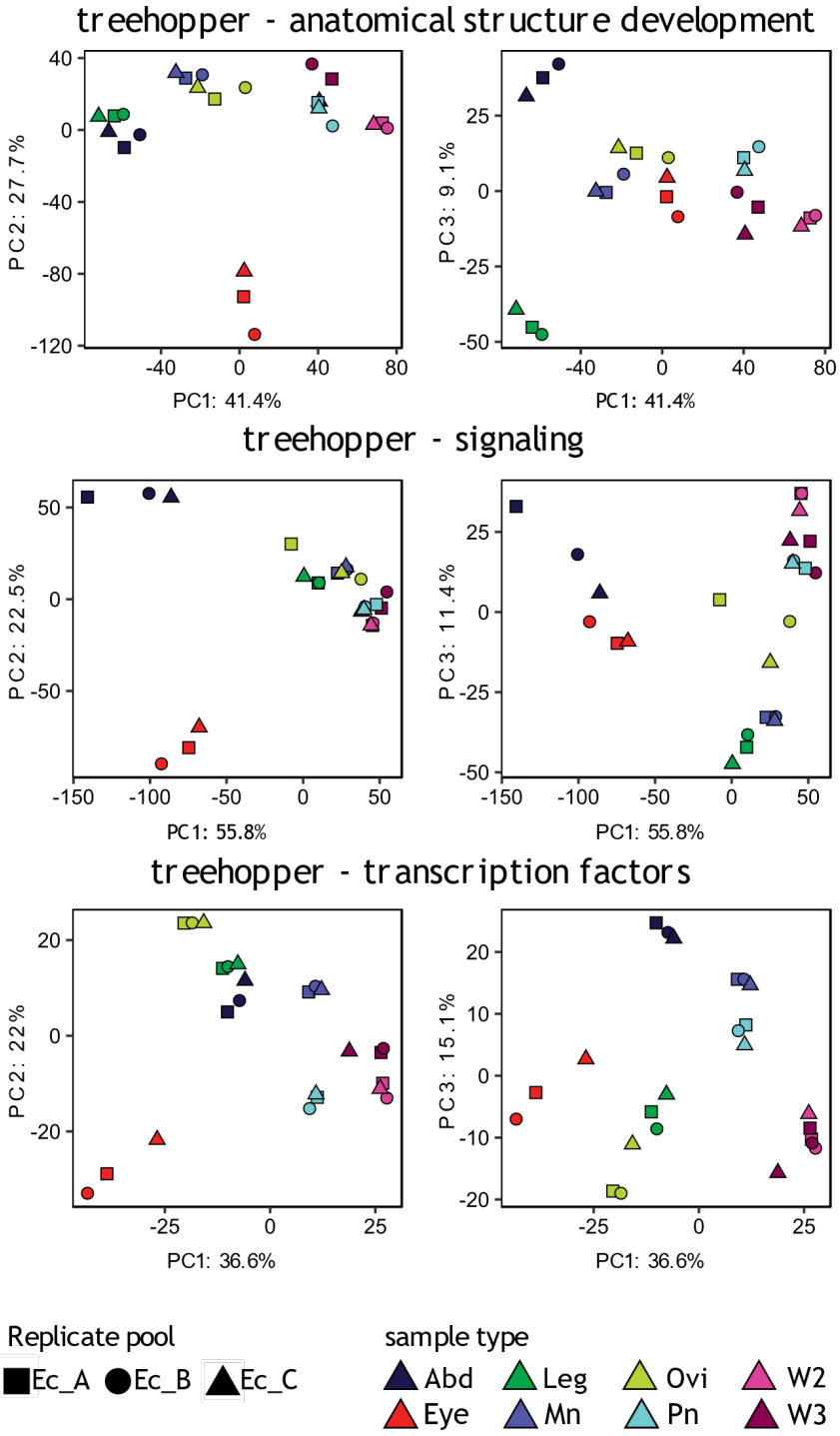
Supplementary Figure 1 The corium of the *E. carinata* forewing has very similar characteristics to the helmet. The costal-subcostal area (the corium) is sclerotized through two-thirds the length of the wing and bears the same punctate pattern as the helmet. **a,** A forewing from a wild-type male *E. carinata*, oriented as in live individuals with proximal in the upper left corner and anterior 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.



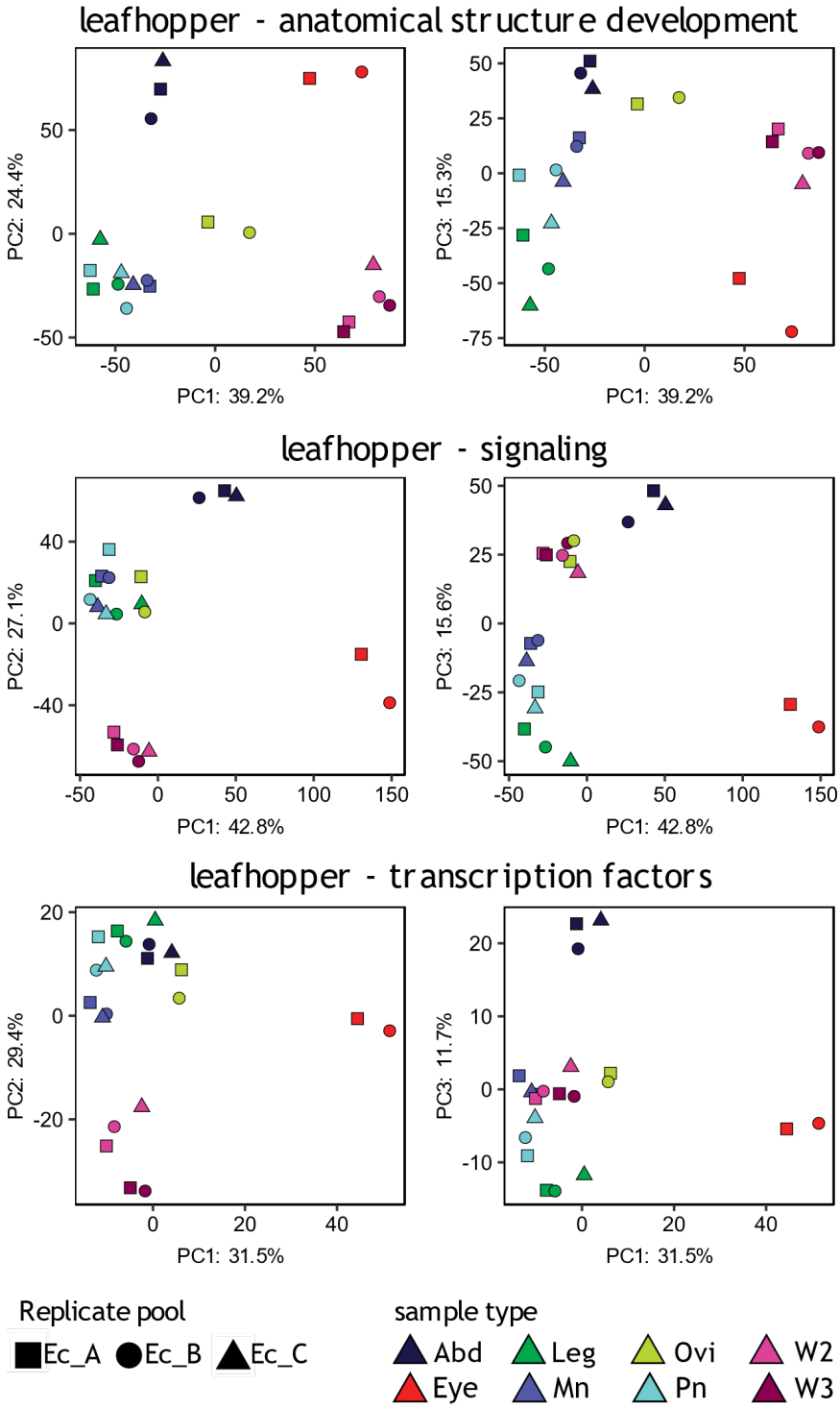
Supplementary Figure 2 Treehopper character tree with forewings excluded. Hierarchical clustering based on all genes differentially expressed across seven body regions (excluding forewings) in *Entylia carinata*. The helmet clusters with the hind wings. Support values (approximately unbiased values for 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



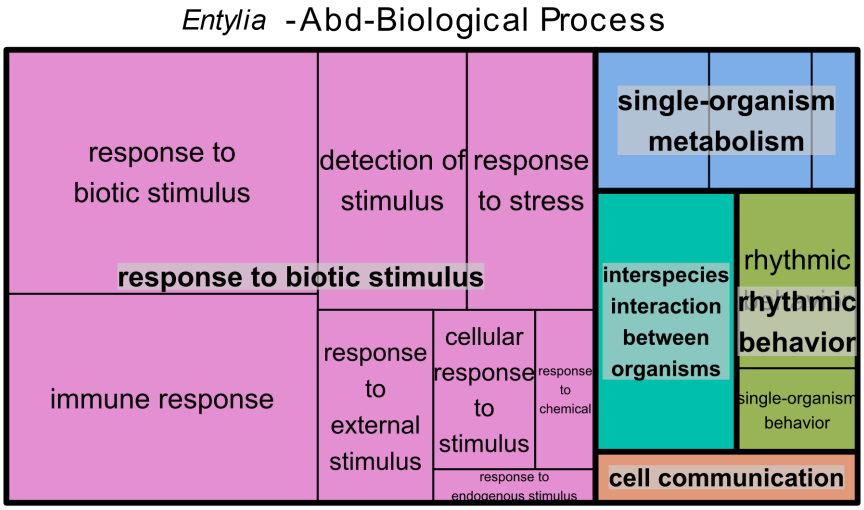
Supplementary Figure 3 Multispecies character tree excluding genes with a strong species signal. Hierarchical clustering of unbiased single copy orthologs results in nearly the same topology as the hierarchical clustering of all single copy orthologs (Fig. 2f). Treehopper helmets (Ec_Pro samples) cluster 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.

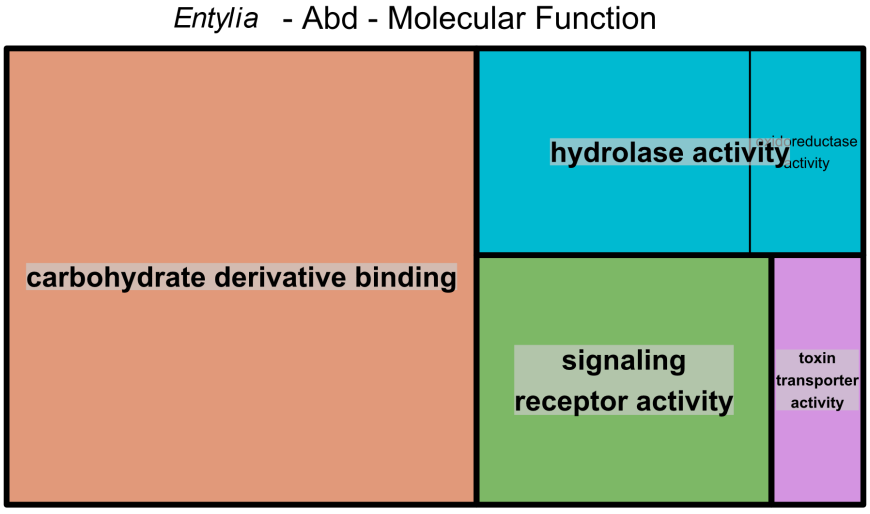


Supplementary Figure 4 Principal components analysis of differentially expressed gene subsets in the treehopper *Entylia carinata*. Genes annotated with (top row) anatomical structure development (GO:0048856), (middle row) signaling (GO:0023052), and (bottom row) transcription factor activity (GO:0003700). Color indicates body region and shape indicates sibling pool. 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

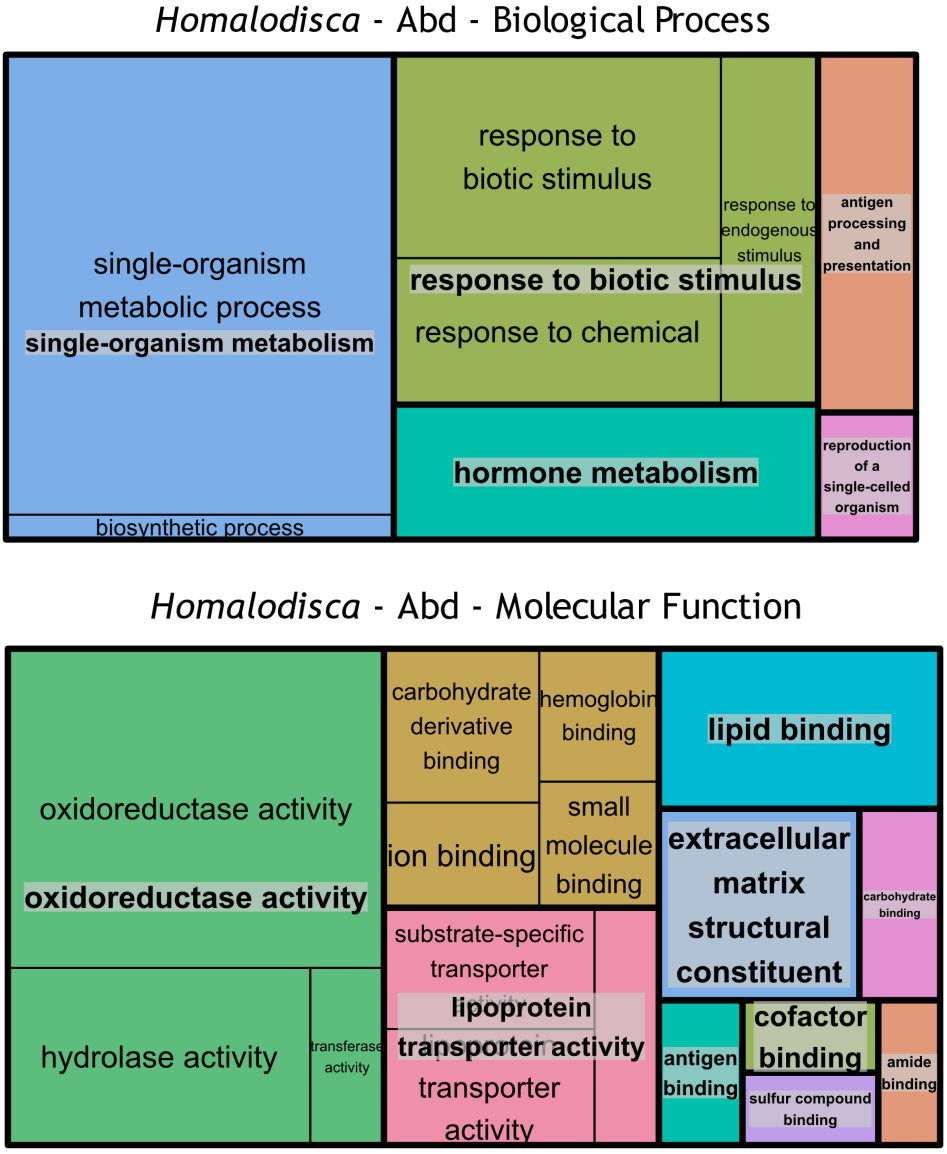


Supplementary Figure 5 Principal components analysis of differentially expressed gene subsets in the leafhopper Homalodisca vitripennis. Genes annotated with (top row) anatomical structure development (GO:0048856, top row), signaling (GO:0023052, middle row), and transcription factor activity (GO:0003700, 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.

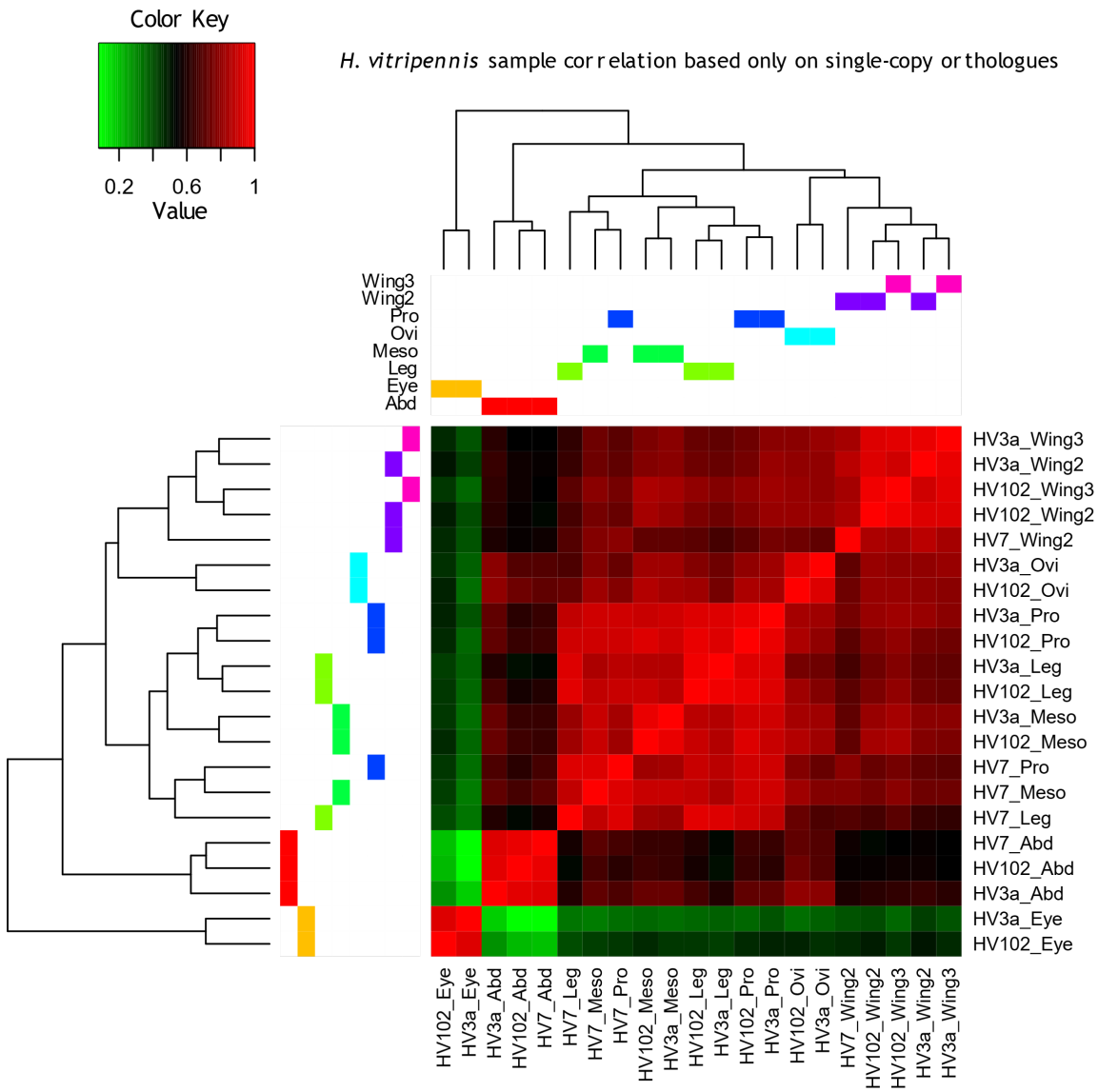




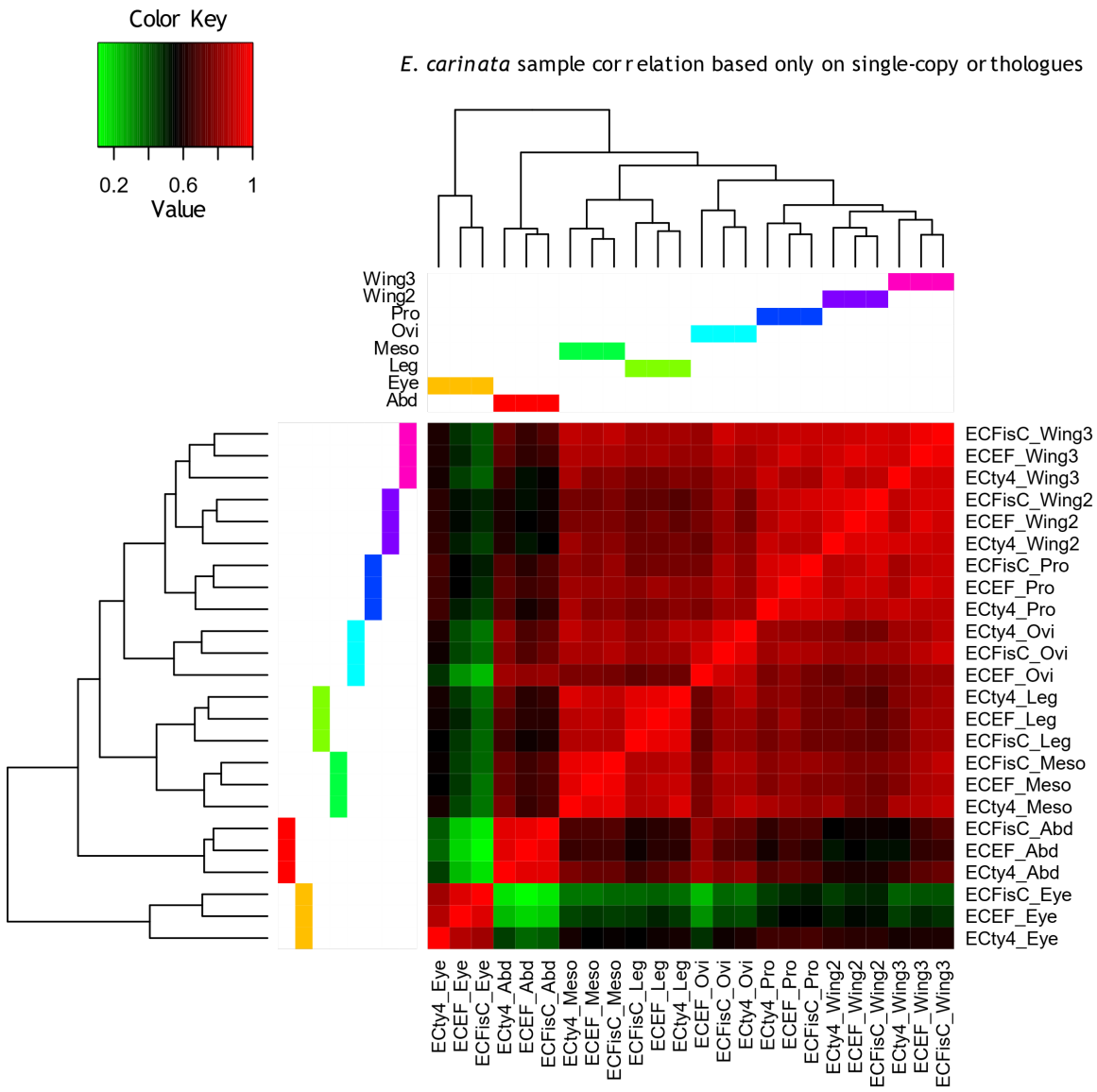
Supplementary Figure 6 Space-filling tree map of enriched GO terms for *Entylia carinata* abdominal tergite samples. The size of the box is inversely proportional to p-value for over-representedness (larger box = more significant). Semantically similar terms are colored with the same color. Produced using REVIGO[2].



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[2].



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)[3]. 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. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**, e21800 (2011).

3. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**, 157 (2015).