The link between gene duplication and divergent patterns of gene expression across a complex life cycle

James G. DuBose[1*] and Jacobus C. de Roode[1]

[1]Department of Biology, Emory University

* james.g.dubose@gmail.com

Abstract

The diversification of many lineages throughout natural history has frequently been associated with evolutionary changes in life cycle complexity. However, our understanding of the processes that facilitate differentiation in the morphologies and functions expressed by organisms throughout their life cycles is limited. Theory suggests that the expression of traits is decoupled across life stages, thus allowing for their evolutionary independence. Although trait decoupling between stages is well established, explanations of how said decoupling evolves have seldom been considered. Because the different phenotypes expressed by organisms throughout their life cycles are coded for by the same genome, trait decoupling must be mediated through divergence in gene expression between stages. Gene duplication has been identified as an important mechanism that enables divergence in gene function and expression between cells and tissues. Because stage transitions across life cycles require changes in tissue types and functions, we investigated the potential link between gene duplication and expression divergence between life stages. To explore this idea, we examined the temporal changes in gene expression across the monarch butterfly (*Danaus plexippus*) metamorphosis. We found that within homologous groups, more phylogenetically diverged genes exhibited more distinct temporal expression patterns. This relationship scaled such that more phylogenetically diverse homologous groups showed more diverse patterns of gene expression. Furthermore, we found that duplicate genes showed increased stage-specificity relative to singleton genes. Overall, our findings suggest an important link between gene duplication and the evolution of complex life cycles.

Lay Summary

How do caterpillars and tadpoles turn into butterflies and frogs? It is well established that although the larval and adult stages have the same genome, larvae and adults up and down regulate the expression of different genes. However, knowing this only tells us how metamorphosis happens, not how it or other forms of complex life cycles evolved. Where did the different genes expressed by different life stages come from, and how did they evolve to generate the incredibly different morphologies seen between life stages? Theory suggests that an important mechanism that generates new genes is when an existing gene gets duplicated. The copies then evolutionarily diverge, resulting in genes with different functions and expression patterns. This theory has been predominately developed around the evolution of different cell and tissue types within an organism. Here, we apply this same concept towards explaining the changes in gene expression between different life stages across the monarch butterfly metamorphosis. Consistent with theory, we found that as genes duplicated and evolutionarily diverged, their patterns of gene expression across life stages became increasingly different. Overall, these findings link the evolution of complex life cycles to a more general understanding of how biological complexity evolves through gene duplication.

Introduction

Many groups of organisms undergo extensive morphological and ecological shifts throughout their life cycles. These shifts appear gradual in some organisms, as seen in the relatively continuous development from infant to adult in primates. However, these shifts seem more complex in many other organisms; a larva first must transition into an intermediate pupal stage before restructuring its morphology into the form of a butterfly. Changes in life cycle

complexity have been associated with the diversification of many taxa throughout natural history (Wheeler et al. 2001; Reiss 2002). Despite nearly a century of interest in the evolution of complex life cycles, we still lack a general understanding of the mechanisms that facilitate divergence in the morphologies and functions expressed by organisms throughout their lives.

This gap in our understanding can be partially attributed to the view that complex life cycles are divided into stages that are discrete, which is the central assumption made in foundational theoretical work (Istock 1967; Moran 1994). While this assumption can capture the punctual ecological and developmental dynamics exhibited by organisms that are considered to have complex life cycles, it has also generated the dichotomy that organisms either do or do not have complex life cycles (Moran 1994). However, considering life cycle complexity as a continuous spectrum has the potential to provide more basic insight into the processes that facilitate life cycle evolution. It is apparent that the transition from one life stage to the next requires continuous changes in the relative abundance, activity, or placement of different cells and tissues (Haldane 1932). Therefore, we propose that from an organismal perspective, life cycle evolution can be more fundamentally described by the body of theory concerning the evolution of cell and tissue differentiation, which focuses on describing the mechanisms that facilitate evolutionary change in gene function and patterns of expression.

The adaptive decoupling hypothesis is the most prominent explanation for the evolution of complex life cycles, and is an extension of an earlier hypothesis that different life stages adapt independently to the niches they occupy (Istock 1967; Moran 1994). The adaptive decoupling hypothesis elaborates that complex life cycles allow different stages to independently respond to natural selection by genetically decoupling the development of their traits (Moran 1994). This hypothesis predicts that genetic variation should generate phenotypic variation in certain life

93 stages but not others. Many studies have found results consistent with this prediction (Fellous
94 and Lazzaro 2011; Aguirre et al. 2014; Goedert and Calsbeek 2019; Schott et al. 2022), and more
95 recent studies have elucidated variation in gene expression between stages as the likely driver of
96 said genetic independence (Herrig et al. 2021; Collet et al. 2023). However, the predictions made
97 by the adaptive decoupling hypothesis are limited to descriptions of extant signatures of
98 decoupled traits, which fails to provide a mechanism that explains how trait decoupling evolves.
99 We propose that the evolution of trait decoupling can be explained by the mechanisms already
100 established in the evolution of cell and tissue differentiation because transitions between life
101 stages are driven by continuous turnover in cell and tissue types and functions.
102 Gene duplication is the best described mechanism that generates evolutionary change in
103 patterns of gene expression between cells and tissues (Gu et al. 2002; Huminiecki and Wolfe
104 2004; He and Zhang 2005; Li et al. 2005). However, other non-exclusive mechanisms, such as
105 regulatory network evolution, are likely at play but have been more challenging to empirically
106 study (Wagner 2001; Teichmann and Babu 2004; Zhang et al. 2004). Gene duplication can be
107 generated though unequal crossing over, retrotransposition, and chromosomal duplication and
108 provides a rich source of genetic variation that can facilitate major evolutionary change (Ohno,
109 Susumu 1970; Zhang 2003). Hypotheses concerning the evolution of duplicate genes share key
110 similarities with the adaptive decoupling hypothesis. For example, the neofunctionalization
111 hypothesis suggests that retention of the ancestral function in one copy alleviates selective
112 constraints on the other copy, allowing it to develop novel functions more efficiently (Ohno,
113 Susumu 1970). However, a role of neutral evolution in generating functional divergence between
114 duplicate genes has also been described (Force et al. 1999; He and Zhang 2005), thus offering a
115 mechanism by which traits could diverge between stages that more comprehensively accounts

for the processes that drive evolutionary change. More generally, the idea that complexity is added to a genome through gene duplication is well established (Ohno, Susumu 1970; Martin 1999; Lynch and Conery 2003), and empirical evidence for gene duplication resulting in more complex phenotypes has been documented in a variety of taxa (Tian et al. 2008; Rivera et al. 2010; Leite et al. 2018; Chen et al. 2023). Therefore, investigating the link between gene duplication and life cycle evolution has potential to broaden our understanding of how biological complexity evolves.

Although there are several nuances to predicting the relationship between sequence evolution and expression pattern evolution in duplicate genes, the general expectation is that the evolution of duplicate genes leads to more divergent and (stage) specific expression patterns (Huminiecki and Wolfe 2004; Li et al. 2005). While this insight has primarily been derived from relating duplicate gene evolution to expression pattern divergence between different mammalian tissues, we hypothesize that the same patterns will emerge when examining temporal patterns of gene expression across a complex life cycle. To test the predictions that the evolution of duplicate genes corresponds with more divergent and stage-specific expression patterns (Figure 1), we examined patterns of gene expression across the holometabolous life cycle of the monarch butterfly, *Danaus plexippus.* The *D. plexippus* life cycle is characterized by a non-dispersive caterpillar stage that is specialized for feeding on milkweed foliage, followed by a non-feeding pupal stage during which metamorphosis occurs, and a final highly dispersive imaginal (butterfly) stage that is specialized for reproduction and feeding on nectar. The extreme ontogenetic niche shifts and trait divergence between stages makes *D. plexippus* a promising model system for studying the evolutionary processes that generate morphological and functional divergence throughout life cycles.

Methods

Experimental design and D. plexippus rearing

To quantify changes in gene expression across the holometabolous development of *D. plexippus*, we sequenced mRNA extracted from third instars, fifth instars, early pupae (one day after pupation), late pupae (6-8 days after pupation), and adults (several hours after eclosion). A previous study has suggested that feeding on more toxic milkweed induces changes in gene expression during the second instar (Tan et al. 2019). Therefore, we reared larvae on both *Asclepias incarnata* (less toxic) and *Asclepias curassavica* (more toxic) to ensure that our findings are robust to a major source of environmental variation. We collected five individuals at each stage and from each plant for mRNA quantification.

Parental (P) *D. plexippus* butterflies were caught in St. Marks, Florida, U.S.A. (30°09′33″N 84°12′26″W) during October of 2022. Butterflies were overwintered in a 14°C incubator (to maintain a state of diapause) and were fed approximately 10%-20% honey water every ten days. During March of 2023, butterflies were mated to establish an F1 generation. F1 caterpillars were reared on *Asclepias curassavica* and after maturation and mating, the F2 caterpillars used in this experiment were reared on either *A. curassavica* or *A. incarnata.* To reach the necessary sample size, we used F2 caterpillars from two different lineages that did not share P or F1 ancestors. Treatments of plant species and development stage were randomly distributed to caterpillars from both lineages to minimize confounding due to genetic background. All individuals sampled in this study were reared at the same time and in the same conditions (See Appendix section 1.1-1.3 for details).

Sample collection and preparation

To minimize possible effects of sample handling, all caterpillars, pupae, and adults were snap frozen in liquid nitrogen before being stored at -80°C. Third instars, fifth instars, early pupae, and late pupae were all frozen in sterile centrifuge tubes, and adults were frozen in glassine envelopes several hours after eclosion (after their wings had finished expanding). For each day freezing took place, samples were stored in a polystyrene foam cooler full of dry ice until all flash freezing for that day was completed. This process took approximately one hour or less on any given day, so no sample was on dry ice for more than an hour before being transferred to the -80°C freezer. All freezing took place in the same greenhouse room that the caterpillars were reared in, and no individual left said room before being frozen throughout the duration of the experiment.

Because we were interested in global gene expression patterns, we collected samples by homogenizing whole bodies using a sterile porcelain mortar and pestle. Each sample for a given round of homogenization was placed in a cooler filled with dry ice. Samples were individually placed in a mortar and liquid nitrogen was constantly added throughout the homogenization to prevent samples from thawing. After a given sample was completely homogenized, homogenate was quickly collected using a sterile polypropylene spatula and stored in a fresh centrifuge tube. Twenty samples were randomly selected for each round of homogenization.

RNA extraction and sequencing

We used a Promega SV Total Isolation System kit to extract total RNA from *D. plexippus* homogenate. While our workflow generally followed the manufacturer's suggested protocol, we made several alterations to obtain higher quality RNA. Briefly, we doubled the recommended RNA lysis buffer, increased the relative centrifugal force in all centrifugation steps, and included an additional centrifugation step to better clear organic contaminants and improve final extract purity. The full RNA extraction protocol used can be found in Appendix Section 1.5. After each extraction, we used a NanoDrop to quantify the purity and concentration of the RNA. Samples with an A260/A280 or an A260/A230 of less than 1.95 were discarded and re-extracted. After all extractions were completed, purified RNA was packaged in dry ice and sent to Novogene (Sacramento, CA) for library preparation and sequencing. Briefly, Novogene used an Agilent 5400 Fragment Analyzer System to confirm that all samples had adequate purity levels, concentrations, and volumes, as well as acceptable RNA integrity numbers (minimum = 7.9). Libraries were then prepared via poly-A tail selection and sequenced using a 150bp paired-end approach on a NovaSeq 6000 sequencing system, thus ensuring at least 20 million reads were obtained for each sample.

Sequence processing and gene expression quantification

Initial quality control of raw sequences was performed by Novogene, where adapter sequences, reads with ambiguous base calls in greater than 10% of the read, and reads with a phred score of less than or equal to 5 in 50% of the read were removed. After receiving the sequences from Novogene, we used FASTQC to generate additional quality reports for each

sample (Andrews 2010). This showed that the median phred score did not drop below 30 at any position for any sample. Therefore, no additional sequence quality control was performed.

To quantify transcript abundances for each gene, we used kallisto (v.0.46.2) to pseudo-align reads to the coding sequences of the *D. plexippus* reference genome (v.Dpv3, GenBank Assembly = GCA_000235995.2) (Zhan et al. 2011). Downstream analyses were performed using transcript per million normalized read counts (automatically generated by kallisto) to minimize biases due to unequal gene lengths and varying library sizes (Wagner et al. 2012; Abrams et al. 2019). Prior to analyses that involved phylogenetic-gene expression comparisons and expression specificity, transcript/million values were log transformed.

Quantifying gene expression divergence between stages

Our objective was to quantify the overall transcriptional dissimilarity between stages. We used Manhattan distances to quantify this dissimilarity because our data were high dimensional and because we wanted to consider the magnitudes of transcriptional changes. We first computed the Manhattan distance between each sample using the *dist* R function (R Core Team 2022). We then used the *adonis2* function from the *vegan* R package (v.2.6-4) (Oksanen et al. 2022) to perform a permutational multivariate analysis of variance (PERMANOVA) with 999 permutations, where developmental stage and plant were initially considered as factors. We then performed a PERMANOVA on each set of adjacent stages, as well as between each larval stage and the adult stage. To visualize global expression divergence between stages, we performed principal coordinate analysis using the *prcomp* R function (R Core Team 2022).

Quantifying the relationship between gene phylogenetic divergence and expression pattern divergence within homologous groups

To infer homology between genes, we first used PSI-BLAST (BLAST 2.5.0+) (Altschul 1997) with five iterations to align all *D. plexippus* protein sequences to each other. Genes were then inferred to be homologous if the query sequence showed at least 30% similarity across the length of the target sequence, as well as an E-value of at least $1x10^{-10}$. To examine how including more distant homologs could impact our analysis, we performed an additional analysis where homology was inferred based on at least 20% similarity across 70% of the target sequence and an E-value of less at least $1x10^{-5}$. These less stringent similarly cutoffs for homology inference showed consistent results with our primary analysis (Appendix section 3.2). Homologous pairs were assembled into sets of two-node subgraphs, and subgraphs were then merged based on common node identity to assemble homologous groups.

To quantify the phylogenetic distance between members of inferred homologous groups, we first used MUSCLE (v.5.1) to create a multiple sequence alignment for each group (Edgar 2004). We then used IQ-TREE2 (v.2.1.4) to identify the best fit sequence evolution model and infer maximum likelihood phylogenies for each multiple sequence alignment (Kalyaanamoorthy et al. 2017; Minh et al. 2020). We then used the *cophenetic.phylo* function from the *ape* R package (v. 5.7-1) (Paradis and Schliep 2019) to calculate pairwise phylogenetic distances from each homologous group tree, which we note are based on sequence divergence and not inferred divergence time. To calculate pairwise expression pattern distances we mean centered and standardized the median transcript/million value for each gene across stages by dividing the difference between the transcript/million value and the mean value for each gene by the standard

deviation of transcript/million values across stages. This allowed us to better capture temporal trends in expression by minimizing similarities due to expression magnitudes. We then calculated the pairwise Euclidian distance between each gene expression pattern within a given homologous group using the *dist* R function (R Core Team 2022). Finally, we used Mantel tests to calculate the correlation between phylogenetic and expression pattern distance matrices for each homologous group, which were implemented via the *mantel* function in the *vegan* R package (v.2.6-4) (Oksanen et al. 2022). We then used a t-test to test if the distribution of correlation coefficients was positively shifted from 0, which was implemented using the *t.test* R function (R Core Team 2022).

Quantifying the relationship between phylogenetic diversity and expression pattern diversity across homologous groups

The diversity (D) of each previously described phylogenetic tree was calculated as the sum of branch lengths: $D = \sum_{i=1}^{n} l_i$, where n represents the number of branches and l_i represents the length of the ith branch. To quantify expression pattern diversity, we first used the Ward method to created hierarchical clustering graphs of the temporal expression patterns for each gene. Prior to clustering, the transcripts/million values for each gene were mean centered and standardized because hierarchical clustering will group expression patterns that show distinct temporal trends but have more similar average relative abundances across time points. For each hierarchical clustering graph, diversity was calculated as previously described for phylogenetic diversity. All hierarchical clustering graphs were constructed using the *hclust* R function and all linear models were fit using the *lm* R function (R Core Team 2022). Our data was non-linearly

related and both phylogenetic diversity (Shapiro-Wilk test, W = 0.717, p = 2.909e-16) and expression pattern diversity (W = 0.570, p < 2.2e-16) were non-normally distributed. Therefore, we tested that expression pattern diversity monotonically increases with phylogenetic diversity using Spearman's rank correlations, which was implemented using the *cor.test* R function (R Core Team 2022). We examined correlations across all homologous groups, as well as within homologous group sizes that had five or more groups to discern the effects of gene addition and phylogenetic diversification within groups.

Expression specificity calculation and analysis

Stage-specificity for each gene was calculated using the tissue specificity index τ (Yanai et al. 2005), which ranges from 0 (equal expression across stages) to 1 (expression in a single stage): $\tau = \frac{\sum_{i=1}^{N}(1-x_i)}{N-1}$, where N is the number of stages (for our purposes) and x_i is the expression level normalized to the maximum expression value across stages. Although τ was developed for assessing tissue specificity, it has been used to gain insight into temporal specificity as well (Cardoso-Moreira et al. 2019). We then performed a Kolmogorov–Smirnov test using the *ks.test* R function (R Core Team 2022) to assess if the distribution of τ values was shifted in duplicated genes relative to singleton genes.

Results

The extent of transcriptional divergence between D. plexippus larvae and pupae is comparable to the divergence between larvae and adults.

Because all distinct phenotypes expressed throughout a complex life cycle are coded by the same genome, trait decoupling must be mediated through variation in gene expression across stages. Therefore, we were first interested in the extent that gene expression has diverged between stages throughout the *D. plexippus* metamorphosis.

Overall, we found that gene expression significantly varied by developmental stage ($F = 61.36$, $p < 0.001$) but not plant host ($F = 0.88$, $p = 0.47$) (Figure 2). We then performed pairwise comparisons to test for differences between subsequent stages, as well as between larvae and adults. Following *D. plexippus* throughout metamorphosis: the transition from third instar to fifth instar involves some, but relative few changes in gene expression (distance = 6.97×10^5, $F = 18.67$, $p < 0.001$). Then a substantial change in gene expression occurs during the transition from fifth instar to early pupa (distance = 1.22×10^6, $F = 68.06$, $p < 0.001$), followed by a slightly smaller but comparable change from early pupa to late pupa (distance = 1.20×10^6, $F = 62.25$, $p < 0.001$). Finally, the transition from late pupa to adult involves a modest change in gene expression (distances = 9.37×10^5, $F = 35.43$, $p < 0.001$), but said change is notably less than the changes involved in the previous two transitions. It's interesting to note that the extent of divergence in gene expression between fifth instars and early pupae is comparable to the divergence between both larval stages and adults (third instar: distance = 1.16×10^6, $F = 108.08$, $p < 0.001$; fifth instar: distance = 1.24×10^6, $F = 71.65$, $p < 0.001$). This distinction in early pupae appears to involve a decrease in metabolic investment and an increase in immune investment (Appendix Figure S7). More broadly, the transcriptional changes across stages appear to be mostly driven by differential investment in metabolism and genetic information processing, consistent with niche shifting and developmental requirements (Appendix Figure S7).

Phylogenetic divergence between homologs generally corresponds with increased divergence in temporal expression pattern.

As previously described, the general hypothesized outcome of evolutionary divergence between homologs, which we measured using phylogenetic distances based on sequence divergence, is increased divergence in their expression patterns. Consistent with this hypothesis, we generally found a positive relationship between phylogenetic distance and expression pattern distance within homologous groups (Figure 3). Specifically, a positive association was observed in approximately 72% of groups. However, we note that there is variation in the both the strength and direction of said correlations, with the remaining 28% of groups showing null or negative correlations. Nonetheless, the distribution of correlation coefficients is shifted positively from 0 (mean = 0.19, $t = 6.29$, $p = 5.38 \times 10^{-9}$).

Diversity in the temporal expression patterns exhibited by homologous groups increases with their phylogenetic diversity.

If expression pattern diverges with phylogenetic divergence between genes within a homologous group, the predicted emergent pattern is that as a homologous groups diversifies (in both size and sequence divergence), the group as a whole should accumulate more different patterns of gene expression. This would result in increased overall expression pattern diversity in homologous groups that are more phylogenetically diverse, which we measured using sequence divergence. Consistent with this hypothesis, we found a positive relationship between phylogenetic diversity and expression pattern diversity ($\rho = 0.8345$, $p < 2.2e\text{-}16$) (Figure 4).

However, we also found that the increase in expression pattern diversity started to saturate at higher phylogenetic diversities, and that this relationship was better described by a quadratic model than a linear model (linear model SSE = 826.78, quadratic model SSE = 823.52).

The positive relationship between phylogenetic and expression pattern diversity could have been driven by the addition of genes to homologous groups, as opposed to phylogenetic diversification within the group. Therefore, we examined the relationship within each homologous group with five or more replicates. This analysis revealed positive correlations between phylogenetic and expression pattern diversification for each of the six smaller homologous group sizes (mean $\rho = 0.30$, 95% CI = [0.15, 0.45]), where we had statistical power to detect this positive relationship in two out of six homologous group sizes ($p \leq 0.0185$). However, in the two larger group sizes, we found no associations between phylogenetic and expression pattern diversification (mean $\rho = -0.52$, p range = [0.825, 0.8792]). Full results for each group size can be found in Appendix Table S4. Overall, these results support a positive but saturating relationship between phylogenetic and expression pattern diversity.

Genes within duplicated genes tend to show more stage-specific expression patterns than singleton genes.

Another key prediction regarding expression pattern divergence between duplicate genes is that copies will show increased stage specificity. Consistent with this prediction, we found that genes within homologous groups tended to show increased stage specificity relative to singleton genes ($D = 0.193$, $p < 2.2 \times 10^{-16}$) (Figure 5).

Discussion

Although many studies have investigated the genetic decoupling of traits between life stages, the evolutionary causes and consequences of trait decoupling remain less understood. Therefore, we investigated the link between gene duplication and transcriptional divergence between stages across the *D. plexippus* metamorphosis. By examining how temporal gene expression patterns changed with phylogenetic divergence between duplicate genes, we found that more distantly related genes tended to show more diverged patterns of gene expression (Figure 3). Although the use of pairwise comparisons has been criticized for assessing the relationship between sequence and expression divergence across species (Dunn et al. 2018), as this was not a comparative study across species, distinguishing patterns of divergence between orthologs and paralogs was not central to our goals (duplications that occurred in an ancestral species or more recently could both contribute to trait decoupling between stages). We also found that more phylogenetically diverse groups generally exhibited more diverse patterns of gene expression (Figure 4) and that genes within homologous groups showed increased stage-specificity relative to singleton genes (Figure 5). As predicted, these results are consistent with studies that have examined the role of gene duplication in facilitating expression divergence between different cells and tissues (Gu et al. 2002; Huminiecki and Wolfe 2004; He and Zhang 2005; Li et al. 2005; Yanai et al. 2005; Cardoso-Moreira et al. 2019). This consistency suggests that theories of evolution by gene duplication can be applied more generally towards understanding functional differentiation between stages at the organismal level.

Our findings significantly expand on previous findings that duplicate genes were more likely to vary in expression between larvae and pre-pupae in several *Drosophila* species than singleton genes. (Gu et al. 2004). A more nuanced pattern that we observed was a saturating

390 relationship between phylogenetic diversity and expression pattern diversity. This pattern was
391 recapitulated across homologous group sizes, where the positive relationship between expression
392 pattern diversity and phylogenetic diversity disappeared at larger and more diverse groups
393 (Figure 4). Similar patterns have been documented in humans, mice, and yeasts, with expression
394 divergence occurring more rapidly at shorter evolutionary time scales before plateauing at longer
395 time-scales (Gu et al. 2002; Makova and Li 2003; He and Zhang 2005). Possible explanations for
396 this pattern include decoupled rates of evolution in coding sequence and regulatory elements,
397 dosage sensitivities/balancing, and additional complexities related to neo/sub-functionalization
398 dynamics (Wagner 2000; Wagner 2001; Papp et al. 2003; Qian and Zhang 2008). Regardless of
399 the specific mechanisms, which are beyond the focus of this study, finding this consistency
400 provides stronger evidence that our results recapitulate more fundamental work on duplicate
401 gene evolution. However, deviations from the predicted relationship between phylogenetic
402 divergence and expression pattern divergence were also found. These deviations could likely be
403 explained by the historical context in which specific homologous groups originated and evolved.
404 For example, whether or not the duplication event was lineage-specific or occurred ancestrally,
405 whether or not duplicates arose from a small-scale duplication event or a chromosomal
406 duplication event, and relative importance of selective and neutral processes in generating
407 sequence divergence, are all expected influence duplicate functionalization and expression
408 divergence (Makova and Li 2003; Huminiecki and Wolfe 2004; He and Zhang 2005).
409 We interpret our results as evidence for an important link between gene duplication and
410 life cycle evolution. However, it is important to emphasize that we do not suggest that expansion
411 of the specific homologous groups identified in our analyses were directly involved in the origin
412 of holometabolous development; the origin of holometabolous development was not the focus of

this study. Rather, our aim was to search for a general process by which traits become temporally decoupled, which would result in greater life cycle complexity when said traits are accumulated over time. Previous studies have documented the patterns that emerge from temporal trait decoupling. Genetic independence of traits expressed by different stages has been well described (see: (Cheverud et al. 1982; Aguirre et al. 2014; Goedert and Calsbeek 2019; Medina et al. 2020) for examples and (Collet and Fellous 2019) for a detailed review), and more recent studies have elucidated variation in gene expression between stages as the likely cause of said independence (Critchlow et al. 2019; Herrig et al. 2021; Schott et al. 2022; Collet et al. 2023). Our findings are consistent with this interpretation as well. However, a common theme across previous studies is that decoupling is variable and not universal to all traits or genes. Therefore, a more mechanistic understanding of how decoupling evolves is needed to understand life cycle evolution more comprehensively. Our findings suggest a role of gene duplication in the decoupling of traits and more generally in facilitating divergence in temporal gene expression patterns across stages.

Although our findings suggest an important link between gene duplication and life cycle evolution, we are not able to make causal inferences because gene duplication is not the only mechanism that facilitates evolutionary change in gene expression patterns. Genes are expressed through regulatory networks, and evolutionary changes to said regulatory elements may be facilitated by, but do not require gene duplication. (Wagner 2001; Zhang et al. 2004). It is possible that decoupling of traits between life stages is predominately driven by regulatory evolution. Under this hypothesis, the associations we described between sequence divergence and expression divergence could be attributed to regulatory divergence between homologous genes, as opposed to their differential functionalization. Likewise, the regulatory environment of a given stage can shape patterns of stage specificity in gene expression, which has the potential

to influence how duplicate genes evolve. For example, if a gene is expressed ubiquitously across stages, duplication could lead to broad deleterious effects through dosage sensitivity (Papp et al. 2003). Therefore, it is possible that duplicates of genes with stage-specific expression patterns are more likely to be retained, which could explain our observation that duplicate genes show more stage-specific expression patterns. These alternative hypotheses do not necessarily exclude a role of gene duplication in facilitating life cycle evolution, and future studies that aim to quantify their relative importance will lend key insight into the evolution of life cycle complexity. One approach would be to understand how regulatory elements (such as transcription factors) and duplicate genes have evolved across lineages with varying degrees of life cycle complexity.

Because our samples consisted of whole bodies, the variation in gene expression observed between stages likely represents shifts in the relative abundance or activity of different cell and tissue types throughout the *D. plexippus* post-embryonic development. This, paired with the consistency of our findings with work on the role of gene duplication in generating functional differentiation between cells and tissues suggests that life cycle evolution in multicellular organisms can be more fundamentally understood through evolutionary shifts in the timing at which different cell and tissue types and functions are expressed. This echoes Haldane's earlier ideas that changes in the timings in which genes act is an important aspect of evolutionary change (Haldane 1932). From this perspective, the continuous transition from infant to adult in primates could be mechanistically linked to the extreme transition from larva to butterfly in lepidopterans.

Acknowledgements

We thank Christopher P. Catano and Mackenzie Hoogshagen for helpful comments and discussion on this work. We thank Erik Edwards for growing the plants used for *D. plexippus* rearing, and the members of the de Roode lab for help in managing monarch mating. This work was supported by National Science Foundation grant IOS-1922720 to J.C.dR.

Author Contributions

J.G.D and J.C.dR designed and performed research. J.G.D analyzed the data. J.G.D and J.C.dR wrote the paper.

Competing interests

The authors declare no competing interests.

Data and code availability

All sequences and count matrices generated for this project have been deposited in the NCBI GEO database and can be accessed with the accession number GSE253389 or the BioProject accession number PRJNA1065445. All code written for data analysis can be accessed at https://github.com/gabe-dubose/mtstp.

References

Abrams ZB, Johnson TS, Huang K, Payne PRO, Coombes K. 2019. A protocol to evaluate RNA sequencing normalization methods. *BMC Bioinformatics* 20:679.

Aguirre JD, Blows MW, Marshall DJ. 2014. The genetic covariance between life cycle stages separated by metamorphosis. *Proc. R. Soc. B.* 281:20141091.

Altschul S. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389–3402.

Andrews S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Cardoso-Moreira M, Halbert J, Valloton D, Velten B, Chen C, Shao Y, Liechti A, Ascenção K, Rummel C, Ovchinnikova S, et al. 2019. Gene expression across mammalian organ development. *Nature* 571:505–509.

Chen H, Guo M, Dong S, Wu X, Zhang G, He L, Jiao Y, Chen S, Li L, Luo H. 2023. A chromosome-scale genome assembly of *Artemisia argyi* reveals unbiased subgenome evolution and key contributions of gene duplication to volatile terpenoid diversity. *Plant Communications* 4:100516.

Cheverud JM, Rutledge JJ, Atchley W R. 1982. Quantitative genetics of development: genetic correlations among age-specific trait values and the evolution of ontogeny. *Evol* 37:895–905.

Collet J, Fellous S. 2019. Do traits separated by metamorphosis evolve independently? Concepts and methods. *Proc. R. Soc. B.* 286:20190445.

Collet JM, Nidelet S, Fellous S. 2023. Genetic independence between traits separated by metamorphosis is widespread but varies with biological function. *Proc. R. Soc. B.* 290:20231784.

Critchlow JT, Norris A, Tate AT. 2019. The legacy of larval infection on immunological dynamics over metamorphosis. *Phil. Trans. R. Soc. B* 374:20190066.

Dunn CW, Zapata F, Munro C, Siebert S, Hejnol A. 2018. Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proc. Natl. Acad. Sci. U.S.A.* [Internet] 115. Available from: https://pnas.org/doi/full/10.1073/pnas.1707515115

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32:1792–1797.

Fellous S, Lazzaro BP. 2011. Potential for evolutionary coupling and decoupling of larval and adult immune gene expression. *Molecular Ecology* 20:1558–1567.

Force A, Lynch M, Pickett FB, Amores A, Yan Y, Postlethwait J. 1999. Preservation of Duplicate Genes by Complementary, Degenerative Mutations. *Genetics* 151:1531–1545.

Goedert D, Calsbeek R. 2019. Experimental Evidence That Metamorphosis Alleviates Genomic Conflict. *The American Naturalist* 194:356–366.

Gu Z, Nicolae D, Lu HH-S, Li W-H. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends in Genetics* 18:609–613.

Gu Z, Rifkin SA, White KP, Li W-H. 2004. Duplicate genes increase gene expression diversity within and between species. *Nat Genet* 36:577–579.

Haldane JBS. 1932. The Time of Action of Genes, and Its Bearing on some Evolutionary Problems. *The American Naturalist* 66:5–24.

He X, Zhang J. 2005. Rapid Subfunctionalization Accompanied by Prolonged and Substantial Neofunctionalization in Duplicate Gene Evolution. *Genetics* 169:1157–1164.

Herrig DK, Vertacnik KL, Kohrs AR, Linnen CR. 2021. Support for the adaptive decoupling hypothesis from whole-transcriptome profiles of a hypermetamorphic and sexually dimorphic insect, *Neodiprion lecontei*. *Molecular Ecology* 30:4551–4566.

Huminiecki L, Wolfe KH. 2004. Divergence of Spatial Gene Expression Profiles Following Species-Specific Gene Duplications in Human and Mouse. *Genome Res.* 14:1870–1879.

Istock CA. 1967. The evolution of complex life cycle phenomena: An ecological perspective. *Evolution* 21:592–605.

Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 14:587–589.

Leite DJ, Baudouin-Gonzalez L, Iwasaki-Yokozawa S, Lozano-Fernandez J, Turetzek N, Akiyama-Oda Y, Prpic N-M, Pisani D, Oda H, Sharma PP, et al. 2018. Homeobox Gene Duplication and Divergence in Arachnids.O'Connell MJ, editor. *Molecular Biology and Evolution* 35:2240–2253.

Li W-H, Yang J, Gu X. 2005. Expression divergence between duplicate genes. *Trends in Genetics* 21:602–607.

Lynch M, Conery JS. 2003. The Origins of Genome Complexity. *Science* 302:1401–1404.

Makova KD, Li W-H. 2003. Divergence in the Spatial Pattern of Gene Expression Between Human Duplicate Genes. *Genome Res.* 13:1638–1645.

Martin AP. 1999. Increasing Genomic Complexity by Gene Duplication and the Origin of Vertebrates. *The American Naturalist* 154:111–128.

Medina I, Vega-Trejo R, Wallenius T, Symonds MRE, Stuart-Fox D. 2020. From cryptic to colorful: Evolutionary decoupling of larval and adult color in butterflies. *Evolution Letters* 4:34–43.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.Teeling E, editor. *Molecular Biology and Evolution* 37:1530–1534.

Moran NA. 1994. Adaptation and constraints in the complex life cycles of animals. *Annu. Rev. Ecol. Syst.* 25:573–600.

Ohno, Susumu. 1970. Evolution by Gene Duplication. New York: Springer

Oksanen J, Simpson GL, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Solymos P, Stevens MHH, Szoecs E, et al. 2022. vegan: Community Ecology Package. Available from: https://CRAN.R-project.org/package=vegan

Papp B, Pál C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424:194–197.

Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R.Schwartz R, editor. *Bioinformatics* 35:526–528.

Qian W, Zhang J. 2008. Gene Dosage and Gene Duplicability. *Genetics* 179:2319–2324.

R Core Team. 2022. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing Available from: https://www.R-project.org/

Reiss JO. 2002. The phylogeny of amphibian metamorphosis. *Zoology* 105:85–96.

Rivera AS, Pankey MS, Plachetzki DC, Villacorta C, Syme AE, Serb JM, Omilian AR, Oakley TH. 2010. Gene duplication and the origins of morphological complexity in pancrustacean eyes, a genomic approach. *BMC Evol Biol* 10:123.

Schott RK, Bell RC, Loew ER, Thomas KN, Gower DJ, Streicher JW, Fujita MK. 2022. Transcriptomic evidence for visual adaptation during the aquatic to terrestrial metamorphosis in leopard frogs. *BMC Biol* 20:138.

Tan W, Acevedo T, Harris EV, Alcaide TY, Walters JR, Hunter MD, Gerardo NM, De Roode JC. 2019. Transcriptomics of monarch butterflies (*Danaus plexippus*) reveals that toxic host plants alter expression of detoxification genes and down-regulate a small number of immune genes. *Molecular Ecology* 28:4845–4863.

Teichmann SA, Babu MM. 2004. Gene regulatory network growth by duplication. *Nat Genet* 36:492–496.

Tian C, Gao B, Rodriguez MDC, Lanz-Mendoza H, Ma B, Zhu S. 2008. Gene expression, antiparasitic activity, and functional evolution of the drosomycin family. *Molecular Immunology* 45:3909–3916.

Wagner A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: Implications for the neutralist-selectionist debate. *Proc. Natl. Acad. Sci. U.S.A.* 97:6579–6584.

Wagner A. 2001. The Yeast Protein Interaction Network Evolves Rapidly and Contains Few Redundant Duplicate Genes. *Molecular Biology and Evolution* 18:1283–1292.

Wagner GP, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* 131:281–285.

585 Wheeler WC, Whiting M, Wheeler QD, Carpenter JM. 2001. The Phylogeny of the Extant
586 Hexapod Orders. *Cladistics* 17:113–169.

587 Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban
588 S, Safran M, Domany E, et al. 2005. Genome-wide midrange transcription profiles reveal
589 expression level relationships in human tissue specification. *Bioinformatics* 21:650–659.

590 Zhan S, Merlin C, Boore JL, Reppert SM. 2011. The Monarch Butterfly Genome Yields Insights
591 into Long-Distance Migration. *Cell* 147:1171–1185.

592 Zhang J. 2003. Evolution by gene duplication: an update. *Trends in Ecology & Evolution*
593 18:292–298.

594 Zhang Z, Gu J, Gu X. 2004. How much expression divergence after yeast gene duplication could
595 be explained by regulatory motif evolution? *Trends in Genetics* 20:403–407.

596

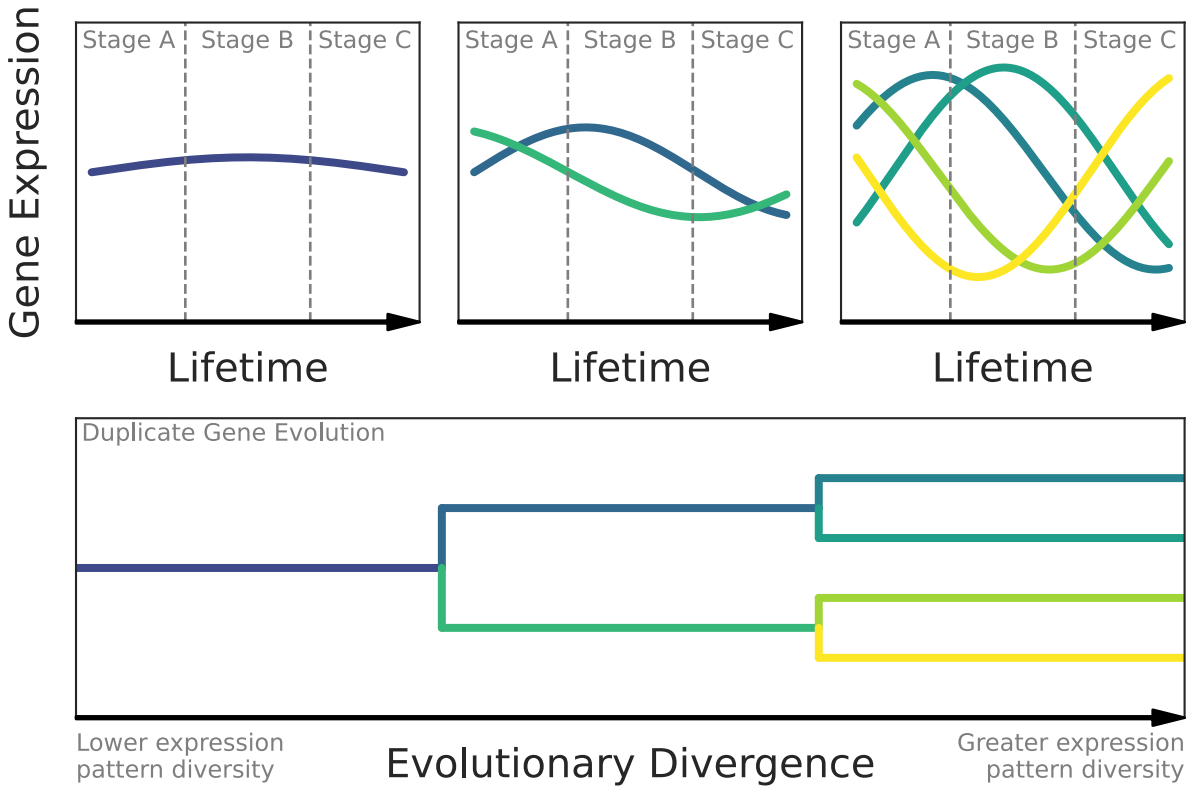

Figure 1. A conceptual diagram showing the hypothesized mechanism of how duplicate gene evolution could lead to divergence in gene expression (and consequently phenotypes) between perceived stages in a complex life cycle. Initially, a given gene has an expression pattern that is relatively uniform throughout an organism's lifetime. After duplication, the expression patterns of each copy tend to diverge and become more stage specific. After additional duplication and divergence, expression tends to diverge and specify even more between copies. This makes the expression at each stage substantially more distinct from other stages, which would result in greater phenotypic divergence between stages if the duplicates functionally diverged as well. This diagram does not show all possible fates of duplicate genes.

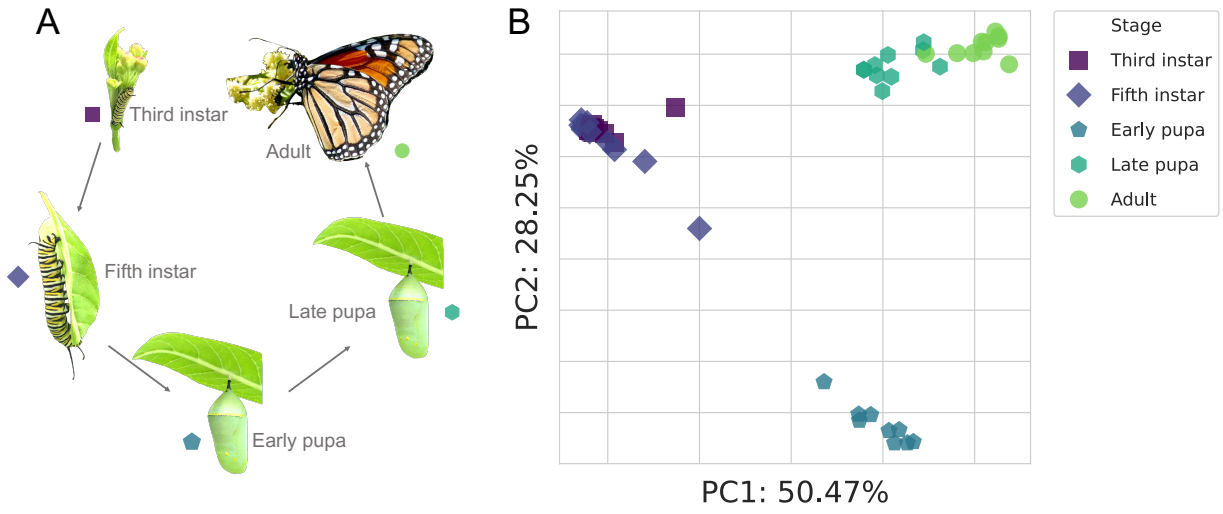

Figure 2. A depiction of how morphology and transcription changes across the *D. plexippus* lifecycle. A) Images of each life stage sampled in this study showing. B) A principal coordinate analysis plot showing substantial transcriptional divergence between life stages. Each point represents the global gene expression profile of an individual, and closer points indicate more similar gene expression profiles. Axis labels indicate principal coordinate rank and the proportion of variance explained.

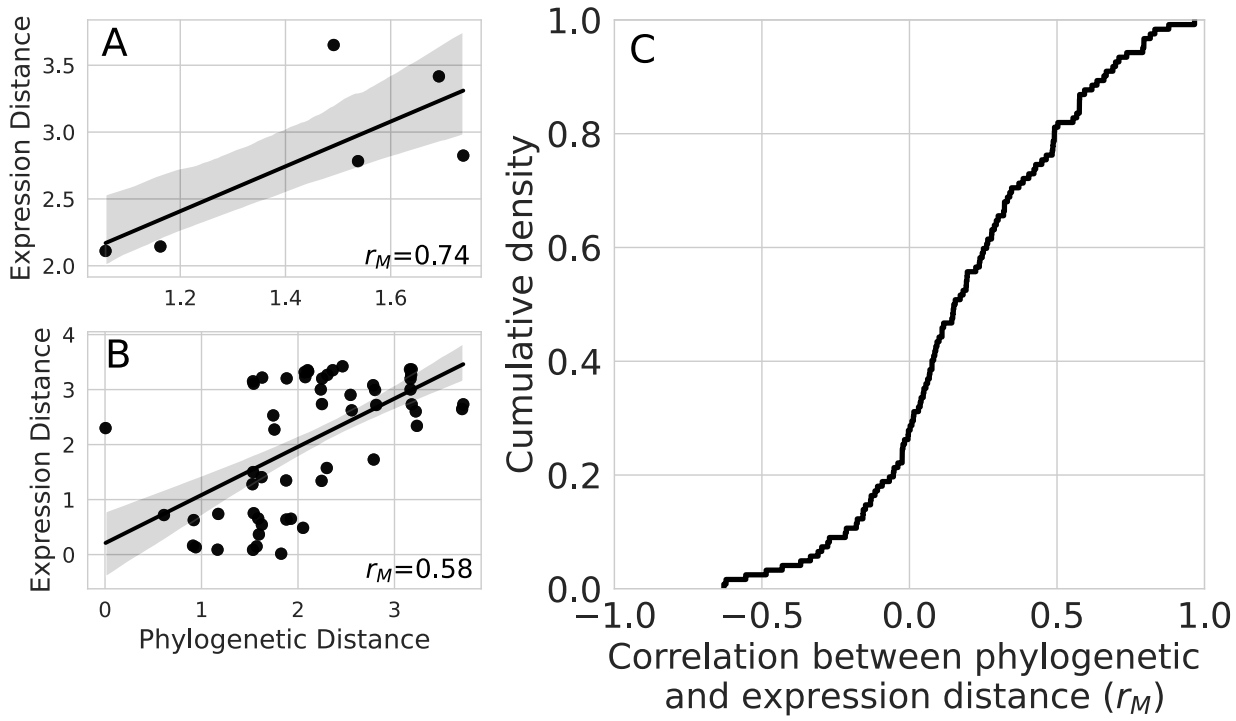

Figure 3. Phylogenetic distance positively correlates with expression pattern distance in most homologous gene groups. The correlation between phylogenetic and expression distances in a set of A) Hox homologs and B) Osiris homologs. In A and B, each point indicates the phylogenetic and expression distance for a pair of genes within the homologous groups. A and B are meant to contextualize the broader analysis, not to lend interpretations about the specific homologous groups used for demonstration purposes. C) The empirical cumulative density function of correlation coefficients between phylogenetic distance and expression pattern distance across all homologous groups. Values greater than 0 indicate a positive correlation and greater values indicate stronger correlations. Overall, the majority of the distribution (approximately 72%) consists of positive correlations.

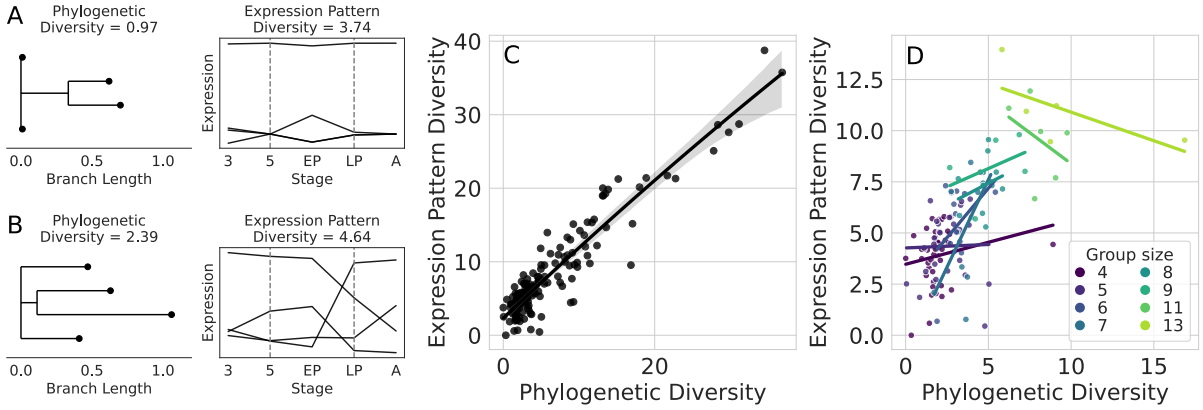

Figure 4. More phylogenetically diverse homologous groups exhibit more diverse patterns of expression. A) An example of a less phylogenetically diverse homologous group (a geranylgeranyl diphosphate synthase-like group) showing less diverse patterns of expression. B) An example of a more phylogenetically diverse group (an arrestin-like group) showing more diverse patterns of expression. A and B are meant to contextualize the broader analysis, not to lend interpretation about the specific homologous groups used for demonstration purposes. C) The relationship between phylogenetic and expression pattern diversity across all homologous groups. The solid black line depicts the fit quadratic model, and the light gray area indicates the 95% confidence interval for said model. D) The relationship between phylogenetic and expression pattern diversity by homologous group size. Each line represents the linear model fit to each homologous group size (only group sizes with five or more replicates were considered in this analysis).

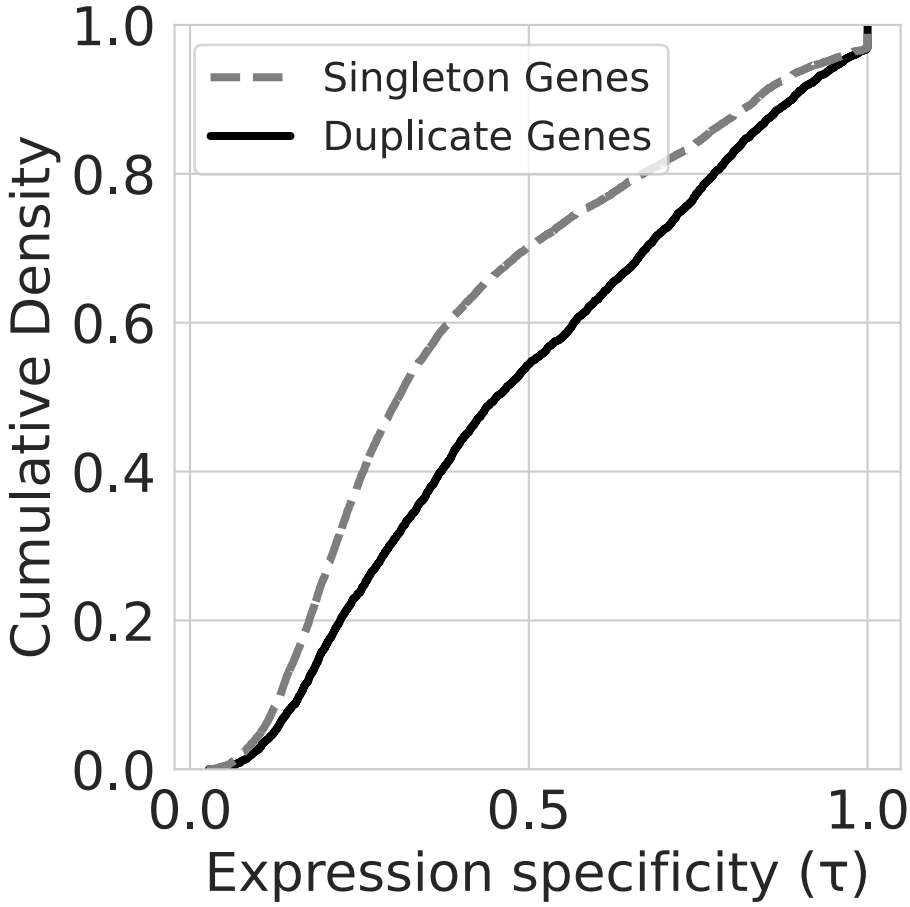

Figure 5. The expression patterns of duplicate genes show increased stage-specificity relative to singleton genes. The empirical cumulative density functions of expression specificity values for duplicate (solid line) and singleton (dashed line) genes. Higher expression specificity values indicate increased stage-specificity.

Appendix for:
The link between gene duplication and divergent patterns of gene expression across a complex life cycle

James G. DuBose[1,*] and Jacobus de Roode[1]

[1] *Department of Biology, Emory University*
[*] *Corresponding author: James G. DuBose, james.g.dubose@gmail.com*

Contents

1 Extended Methods

1.1 Study system and experimental design

Holometabolous development involves the transition from a larval stage that is typically specialized for feeding and growth to a stationary or less mobile pupal stage. During the pupal stage, dramatic morphological restructuring occurs, resulting in a distinct adult stage that is typically specialized for dispersal and reproduction.

To quantify changes in gene expression the across the holometabolous development of *Danaus plexippus*, we sequenced mRNA extracted from third instars, fifth instars, early pupae (one day after pupation), late pupae (6-8 days after pupation), and adults (several hours after eclosion). A previous study has suggested that feeding on more toxic milkweed induces changes in gene expression during the second instar [1]. Therefore, we reared larvae on both *Asclepias incarnata* (less toxic) and *Asclepias curassavica* (more toxic) to ensure that our findings are robust to a major source of environmental variation. We collected five individuals at each stage and from each plant for mRNA quantification. All individuals sampled in this study were reared at the same time and in the same conditions.

1.2 Milkweed cultivation

A. incarnata and *A. curassavica* seeds were purchased from Joyful Butterfly (Blackstock, SC, USA). To break cold dormancy, seeds were placed in sand-filled bags and kept at 4°C for two months prior to sowing. Approximately two months before the start of the experiment, seeds were sown into Lambert LM-GPS germination soil and placed in a temperature-controlled greenhouse room that was held between 25°C and 29.4°C. *A. incarnata* germination rates tend to be relatively low, so seed trays were topped with vermiculite to aid in moisture retention. Seedlings were fertilized with approximately 20 PPM of Jack's LX 15-5-15 with 4% Ca and 2% Mg fertilizer three times a week until the majority of plants grew two sets of true leaves. All plants were then re-potted into Pro-mix BK25 soil, moved to a new temperature-controlled room that was held between 25.6°C and 29.4°C, and fertilized three times a week as described above. Approximately one week before the start of the experiment, plants were moved into the same greenhouse room that caterpillars were reared in (described below).

1.3 *D. plexippus* Rearing

Monarch butterflies were caught and labeled near St. Marks, Florida, U.S.A. (30°09'33"N 84°12'26"W) between October 21st and October 23rd, 2022. Clear tape was placed on the abdomen of each butterfly and examined under a stereomicroscope to ensure they were not infected by *Ophryocystis elektroscirrha*, a common parasite of monarch butterflies. Prior to mating season, wild-caught monarch butterflies were stored in glassine envelopes at 14°C to induce a state of diapause, and were fed approximately 10-20% honey water every 10 days. Between March 6th and March 15th, 2023, wild-caught monarchs were placed in mesh cages for mating. Each cage was set up in a climate-controlled growth chamber (25°C, 16-hour/8-hour day/night cycle) and contained three male and three female butterflies. All cages were provided with a petri dish containing a sponge soaked in approximately 10-20% honey water for butterfly feeding. Mating cages were checked every 14 hours, and copulated butterflies were transferred to their own separate cage. After a copulated pair had detached the next day, the male was removed from the cage and the female was given a potted *A. curassavica* plant for oviposition, as well as honey water as described above. After a given female was done laying eggs, the plant was taken out of the growth chamber and placed in a temperature-controlled greenhouse room that was held between 23.3°C and 27.8°C for.

F1 caterpillars were reared on *A. curassavica* in the same greenhouse room previously described. After pupation, the silk attached to the end of the pupal cremaster was used to hot glue the pupae to the lid of clear solo cups, which were then taken from the greenhouse to the laboratory (22°C) for eclosion. A piece of paper towel was placed in the bottom of cups to help absorb liquids produced during the eclosion process. After eclosion, butterflies were placed in glassine envelopes and stored as previously described.

Between April 23rd and May 1st, 2023, F1 butterflies from different lineages and that were not infected with *O. elektroscirrha* were mated as previously described in the F0 generation. F1 females were given either *A. curassavica* or *A. incarnata* for oviposition, and caterpillars were collectively placed on their treatment

plant species upon hatching. Care was taken to make sure caterpillars that had taken bites of the plant they were oviposited on to were placed on the same milkweed species. Likewise, only caterpillars that had not taken any bites of the plant they were oviposited on were placed on the other milkweed species. To reach the sample size needed for this experiment, we used F2 caterpillars from two different lineages that did not share F0 or F1 ancestors. Treatments of plant species and development stage were randomly distributed to caterpillars from both lineages to minimize confounding due to genetic background.

1.4 Sampling across life stages

To minimize changes in transcription due to sample handling, all caterpillars, pupae, and adults were snap frozen in liquid nitrogen before being stored in -80°C. Third instar caterpillars were pulled from their feeding plant and quickly placed into a sterile 2mL microcentrifuge tube that was then dipped in liquid nitrogen. Fifth instar caterpillars were frozen in the same way but were placed in sterile 5mL centrifuge tubes. Caterpillars that ate all of the leaves off of the plant they were originally placed on were placed on another plant of the same species.

One day after pupation, early pupae were placed in 5mL centrifuge tubes and frozen in liquid nitrogen as described above. Three days after pupation, pupae assigned to late pupa and adult stages were removed from their plant and taped to the lids of clear solo cups using silk attached to the cremaster. In some cases, not enough silk detached with the pupa, and tape was applied directly to the cremaster. Solo cups were then placed on the bottom rack of the same shelf that the caterpillars were reared on, and shade was provided by placing plastic trays above and to the southeast facing side of the shelf to prevent pupae from burning. A piece of paper towel was placed in the bottom of the cups containing adult samples to absorb fluids produced during the pupation process. Since there is variation in how long it takes for a pupa to eclose, late pupae were collected 6-8 days after pupation. Care was taken to ensure the distributions of how many days after pupation late pupae were sampled were equal between plants. Adults were frozen several hours after eclosion to allow their wings to fully expand. Here, adults were removed from their solo cup and quickly placed in glassine envelopes, which were then quickly frozen in liquid nitrogen.

After flash freezing in liquid nitrogen, samples were stored in a styrofoam cooler full of dry ice until all freezing for that day was completed. This process took approximately one hour or less on any given day, so no sample was on dry ice for more than an hour before being transferred to the -80°C freezer. All freezing took place in the same greenhouse room that the caterpillars were reared in, and no monarch left said room before being frozen throughout the duration of the experiment.

1.5 RNA extraction and sequencing

We use a Promega SV Total Isolation System kit to extract total RNA from the monarch homogenate. Extractions were performed in batches of 11 samples with 1 negative control (per extraction batch). After each extraction, we used a NanoDrop to quantify the purity and concentration of the RNA. Samples with an A260/A280 or an A260/A230 of less than 1.95 were discarded and re-extracted to meet purity standards. While the general workflow followed the manufacturer's suggested protocol, we made some alterations to obtain higher quality RNA extract. Briefly, we doubled the recommended RNA lysis buffer to decrease the tissue concentration in the initial lysis step. All centrifugation steps were increased to 20,000 rcf to better remove organic contaminants and performed at 17°C to avoid sample heating. We also added an additional centrifugation step after the initial tissue lysis to further clear organic contaminants and improve final extract purity. The specific protocol is as follows:

1. Add homogenate to 2mL microcentrifuge tube.
2. Immediately add 590 uL of RNA Lysis Buffer (RLA+BME) into microcentrifuge tube with homogenate
3. Use sterile micropestle to crush and lyse monarch homogenate (vigorously crush and spin pestle in tube for approximately 1 minute).
4. Centrifuge at 20,000 rcf for 10 minutes at 17°C.
5. Transfer approximately 400 uL to 500 uL of aqueous phase to a new microcentrifuge tube.

6. Centrifuge at 20,000 rcf for 20 minutes at 17°C.
7. Transfer 175uL of the cleared lysate (aqueous layer) to a new microcentrifuge tube.
8. Add 200 uL of 95
9. Transfer lysate+ethanol to Spin Basket Assembly and centrifuge for 5 minutes at 20,000 rcf and 17°C.
10. While the centrifuge is running, Prepare DNase incubation mix: 40uL of Yellow Core Buffer + 5uL MnCl2 + 5uL of Dnase I (per sample). Mix gently via pipetting.
11. Discard eluate.
12. Add 50 uL of DNase incubation mix to the membrane of the Spin Basket, incubate for 15 minutes at room temperature.
13. Add 200 uL of DNase Stop Solution (DSA+ethanol) and centrifuge at 20,000 rcf for 1 minute at 17°C.
14. Discard eluate.
15. Add 600 uL of RNA Wash Solution (RWA); centrifuge at 20,000 rcf for 1 minute at 17°C.
16. Discard eluate.
17. Add 250 uL of RNA Was Solution (RWA); centrifuge at 20,000 rcf for 2 minutes at 17°C.
18. Transfer Spin Basket to Elution Tube.
19. Add 100 uL of Nuclease-Free water to the Spin Basket membrane.
20. Centrifuge at 20,000 rcf for 1 minute to elute RNA.
21. Store at -80°C.

After all extractions were completed, purified RNA was packaged in dry ice and sent to Novogene for sequencing. Briefly, Novogene used an Agilent 5400 Fragment Analyzer System to performed additional quality control. This involved reconfirming sample purity, ensuring that all samples had adequate concentrations and volumes, and checking that all sampled had acceptable RNA integrity numbers (minimum = 7.9). After additional quality assessment, mRNA was separated via poly-A tail selection, and 150bp paired-end sequencing was performed using a NovaSeq 6000 sequencing system, ensuring at least 20 million reads were obtained for each sample.

1.6 Sequence processing and gene expression quantification

Quality control of raw sequences was initially performed by Novogene. This entailed the removal adapter sequences, the removal of reads with ambiguous base calls in greater than 10% of the read, and the removal of reads with a phred score of less than or equal to 5 in 50% of the read. After receiving the sequences from Novogene, we used FASTQC to generate additional quality reports for each sample [2]. This showed that for each sample, the median phred score did not drop below 30 at any position along the reads. Therefore, no additional quality control was performed.

To quantify transcript abundances for each gene, we used kallisto (v.0.46.2) to pseudo-align reads to the coding sequences of the D. plexippus reference genome (v.Dpv3; GCA_000235995.2) [3]. Downstream analyses were performed using transcript per million normalized read counts (automatically generated by kallisto) to minimize biases due to unequal gene lengths and varying library sizes [4, 5].

1.7 Quantifying gene expression divergence between stages

Given the high dimensionality of gene expression data, we first computed the Manhattan distance between each sample using the dist R function [6]. We then used the adonis2 function from the vegan R package (v.2.6-4) [7] to perform a permutational multivariate analysis of variance (PERMANOVA) with 999 permutations, where developmental stage and plant were initially considered as factors. We then performed a PERMANOVA on each set of adjacent stages, as well as between each larval stage and the adult stage. To visualize global expression divergence between stages, we performed principal component analysis using the *prcomp* R function [6]. To ensure our findings were robust to different metrics for evaluating overall transcriptional differences, we performed the same analysis using Pearson correlation distances, which were calculated using the *cor* R function [6].

1.8 Inferring homologous gene groups

To infer homology between genes, we first used PSI-BLAST (BLAST 2.5.0+) [8] with five iterations to align all D. plexippus protein sequences to each other. Genes were then inferred to be homologous if the query sequence showed at least 30% similarity across the length of the target sequence, as well as an E-value of at least 1x10-10. To examine how including more distant homologs could impact our analysis, we performed an additional analysis where homology was inferred based on at least 20% similarity across 70% of the target sequence and an E-value of less than 1x10-5. Homologous pairs were assembled into sets of two-node subgraphs, and subgraphs were then merged based on common node identity to assemble homologous groups.

To quantify the phylogenetic distance between members of inferred homologous gene groups, we first used MUSCLE (v.5.1) to create a multiple sequence alignment for each group [9]. We then used IQ-TREE2 (v.2.1.4) to identify the best fit sequence evolution model and infer maximum likelihood phylogenies for each multiple sequence alignment [10, 11].

1.9 Quantifying the relationship between gene phylogenetic divergence and expression pattern divergence within homologous groups

To quantify the relatinship between phylogenetic divergence and expression divergence within homologous groups, we used the *cophenetic.phylo* function from the *ape* R package (v. 5.7-1) [12] to calculate pairwise phylogenetic distances from each homologous group tree. To calculate pairwise expression pattern distances, we first mean centered and standardized the median transcripts/million value for each gene within each stage to better measure distance between temporal patterns as opposed to magnitude (which cannot be assessed with our data). We then calculated the pairwise Euclidian distance between each gene expression pattern within a given homologous group using the *dist* R function [6]. Finally, we used Mantel tests to calculate the correlation between phylogenetic and expression pattern distance matrices for each homologous group, which were implemented via the *mantel* function in the vegan R package (v.2.6-4) [7]. We then used a t-test to test if the distribution of correlation coefficients was positively shifted from 0, which was implemented using the *t.test* R function [6].

1.10 Quantifying the relationship between phylogenetic and expression pattern diversity

The diversity (D) of each tree was then calculated by summing all branch lengths: $D = \sum_{i=1}^{n} l_i$, where n represents the number of branches and l_i represents the length of the ith branch. To quantify expression pattern diversity, we first created hierarchical clustering graphs of the temporal expression patterns for each gene using the Ward method, as implemented by hclust R function [6]. Prior to clustering, the transcripts/million values for each gene were mean centered and standardized because hierarchical clustering will group expression patterns that show distinct temporal trends but have more similar relative abundances at each time point. For each hierarchical clustering graph, diversity was calculated as previously described for phylogenetic diversity. We then fit a linear model to examine the relationship between phylogenetic diversity and expression pattern diversity across all inferred homologous gene groups, which was implemented using the *lm* R function [6]. Because diversity was calculated additively (for each branch, diversity was added in

proportion to divergence), we also fit individual linear models to each homologous gene group size that had at least five replicates. In addition to removing the inherent positive correlation between group size and diversity, this approach also allowed us to contrast global and local patterns.

1.11 Expression specificity calculation and analysis

Stage-specificity for each gene was calculated using the tissue specificity index τ [13], which ranges from 0 (broad expression) to 1 (specific expression): $\tau = \frac{\sum_{i=1}^{N}(1-x_i)}{N-1}$, where N is the number of stages (for our purposes) and x_i is the expression level normalized to the maximum expression value across stages. Although τ was developed for assessing tissue specificity, it has been used to gain insight into temporal specificity as well [14]. We then performed a Kolmogorov–Smirnov test using the *ks.test* R function [6] to assess if the distribution of τ values was shifted in duplicated genes relative to singleton genes.

2 Methodological Summaries

2.1 RNA quality control report

Table S1: RNA extract quality control report.

Sample Name	Concentration (ng/ul)	Volume (ul)	Total amount (ug)	RIN
mtstp3cu2	120.15	91	10.93	9.7
mtstp3cu3	435.84	93	40.53	9.6
mtstp3cu4	99.07	88	8.72	9.6
mtstp3cu5	223.96	94	21.05	9.7
mtstp3cu8	111.95	91	10.19	9.6
mtstp3iu81	211.09	92	19.42	9.8
mtstp3iu82	194.64	93	18.1	9.8
mtstp3iu83	373.28	91	33.97	9.6
mtstp3iu84	196.04	93	18.23	9.6
mtstp3iu85	136.84	90	12.32	9.7
mtstp5cu17	363.74	101	36.74	9.4
mtstp5cu18	236.69	92	21.78	9.8
mtstp5cu19	544.95	93	50.68	9.5
mtstp5cu20	130	91	11.83	9.7
mtstp5cu21	400.58	94	37.65	9.8
mtstp5iu100	377.26	91	34.33	9.7
mtstp5iu101	98.09	89	8.73	9.6
mtstp5iu97	164.64	92	15.15	9.8
mtstp5iu98	258.2	91	23.5	9.8
mtstp5iu99	139.84	91	12.73	9.8
mtstpAcu65	132.79	92	12.22	9.5
mtstpAcu66	106.76	90	9.61	9.5
mtstpAcu67	37.05	94	3.48	8.9
mtstpAcu68	94.26	91	8.58	9.4
mtstpAcu69	176.26	91	16.04	9.8
mtstpAiu145	162.84	92	14.98	9.6
mtstpAiu146	283.45	92	26.08	9.6
mtstpAiu147	51.58	92	4.75	9.2
mtstpAiu148	89.37	92	8.22	9.3
mtstpAiu149	307.63	91	27.99	9.7
mtstpEcu33	203.48	92	18.72	7.4
mtstpEcu34	244.57	91	22.26	8.7

Sample Name	Concentration (ng/ul)	Volume (ul)	Total amount (ug)	RIN
mtstpEcu35	491.84	89	43.77	7.9
mtstpEcu36	220.03	92	20.24	8.6
mtstpEcu38	220.92	93	20.55	8.6
mtstpEiu113	333.32	93	31	8.9
mtstpEiu114	446.98	91	40.68	7.9
mtstpEiu115	182.76	92	16.81	8.6
mtstpEiu116	257.7	89	22.94	8.4
mtstpEiu117	326.41	92	30.03	8
mtstpLcu49	95.18	92	8.76	9
mtstpLcu50	93.51	92	8.6	9.3
mtstpLcu52	220.52	93	20.51	8.8
mtstpLcu53	95.67	91	8.71	9.4
mtstpLcu56	399.94	89	35.59	8.2
mtstpLiu129	145.89	91	13.28	9.1
mtstpLiu130	91.59	91	8.33	8.7
mtstpLiu131	391.6	87	34.07	9.6
mtstpLiu133	85.42	89	7.6	8.4
mtstpLiu135	108.7	92	10	8.1

2.2 RNA sequencing statistics

After quantifying transcript counts per gene, we checked that our sequencing effort was adequate to downstream analyses. First, we examined the number and proportion of raw reads that passed quality control, as well as the number and proportion of quality-controlled reads that were pseudo-aligned to the *D. plexippus* genome (Table 2). We then generated a rarefaction plot see if our sequencing depth had sufficiently detected the expression of most transcripts that were expressed at a given stage (Figure 1).

Table S2: Sequence processing and mapping summary.

Sample	Raw PE reads	Passed QC	% Passed QC	Pseudo-aligned	% Pseudo-aligned
mtstp3cu2	21943709	2132928515	97.2	16582965.404739982	0.777474036
mtstp3cu3	22924309	2202796852	96.09	17321415.999555275	0.786337423
mtstp3cu4	24873976	2408547096	96.83	18718506.832656555	0.777170057
mtstp3cu5	27818742	2729296778	98.11	21155266.226326376	0.77511784
mtstp3cu8	21942370	2133237211	97.22	16603108.222044216	0.778305766
mtstp3iu81	25932842	2532860678	97.67	19868699.32840216	0.784437119
mtstp3iu82	22375940	2195750992	98.13	17106242.600410897	0.779061135
mtstp3iu83	23183162	2267545075	97.81	17444460.402408678	0.769310414
mtstp3iu84	23107037	2261023570	97.85	17945429.764851093	0.793686099
mtstp3iu85	21374435	2100251983	98.26	15988137.614267953	0.761248543
mtstp5cu17	24827892	2419229796	97.44	19638683.30078156	0.811774199
mtstp5cu18	23283921	2277633152	97.82	18962699.68340757	0.832561629
mtstp5cu19	20552516	1987839348	96.72	17087863.371912975	0.859619938
mtstp5cu20	22934654	2237734191	97.57	18115485.776358116	0.809545917
mtstp5cu21	23666289	2302019931	97.27	19702178.064739898	0.855864791
mtstp5iu100	31121258	3038368419	97.63	25782153.591624036	0.848552579
mtstp5iu101	21768932	2121164734	97.44	18187966.46325012	0.85745186
mtstp5iu97	23849683	2315565722	97.09	19239689.897892967	0.830885071
mtstp5iu98	21766794	2105284316	96.72	17675440.16384149	0.83957497
mtstp5iu99	23386986	2292860107	98.04	18686179.461194277	0.814972505
mtstpAcu65	23267178	2266921153	97.43	17092435.515985236	0.753993384
mtstpAcu66	26977628	2628700072	97.44	20072649.61273931	0.763596038
mtstpAcu67	24972220	2433542839	97.45	18546857.73716618	0.762134015
mtstpAcu68	22515202	2194331587	97.46	15951574.031512374	0.726944557
mtstpAcu69	21267172	2062915684	97	15662285.557752984	0.759230524
mtstpAiu145	29683116	2870950980	96.72	23068781.60188841	0.80352405
mtstpAiu146	22849815	2210262605	96.73	17606493.53499071	0.796579261
mtstpAiu147	22404070	2200079674	98.2	16188725.539826853	0.735824513
mtstpAiu148	25214198	2456367169	97.42	18462578.21672617	0.751621274
mtstpAiu149	22156963	2154321512	97.23	16710014.455382776	0.775650912
mtstpEcu33	22763953	2217664301	97.42	16862537.346874774	0.760373756
mtstpEcu34	25629454	2499384354	97.52	18633231.46842286	0.745512848
mtstpEcu35	22465556	2195109477	97.71	16842284.066866323	0.767263968
mtstpEcu36	22055561	2159239422	97.9	16253631.989667526	0.75274802
mtstpEcu38	23213904	2279605373	98.2	17463986.10284253	0.7660969
mtstpEiu113	23868417	2301870135	96.44	17119222.41677067	0.743709306
mtstpEiu114	22466254	2186865164	97.34	16425515.425250849	0.751098682
mtstpEiu115	23622008	2313067023	97.92	17237434.840021413	0.74521986
mtstpEiu116	33296616	3242424466	97.38	23547226.785158955	0.726222832
mtstpEiu117	22703490	2224942020	98	17101827.281335317	0.76864148
mtstpLcu49	27550907	2684560378	97.44	21053416.44129032	0.78424075
mtstpLcu50	27343921	2665485419	97.48	21414551.829342157	0.803401575
mtstpLcu52	26550778	2588435347	97.49	20277129.22413158	0.783373989
mtstpLcu53	21114739	2069033275	97.99	15824607.707317442	0.764830991
mtstpLcu56	36034764	3517353314	97.61	27879259.42907198	0.792620386
mtstpLiu129	20400384	1985773379	97.34	15447977.51483457	0.777932552
mtstpLiu130	34495919	3368181531	97.64	26660707.28941579	0.791546033
mtstpLiu131	22712422	2196972580	96.73	17262201.926361006	0.785726781
mtstpLiu133	22825856	2225064443	97.48	17685552.79521903	0.794833285
mtstpLiu135	24991444	2433166988	97.36	19590944.360775467	0.805162344

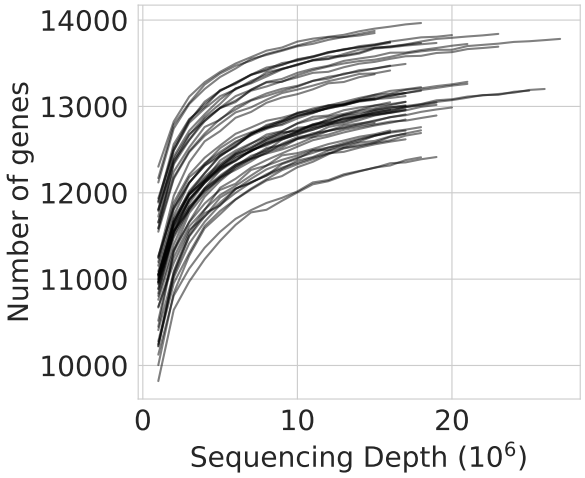

Figure S1: Rarefaction curves showing the number of genes detected on the y-axis and sequencing depth on the x-axis. Each line corresponds to an individual sample. Plateaus in the number of detected genes at higher sequencing depths suggest that our sequencing effort was sufficient.

2.3 Summary of homologous gene group inferences

Homologous gene group inference is described in section 1.8. The following plots show the summary histograms of homologous group size.

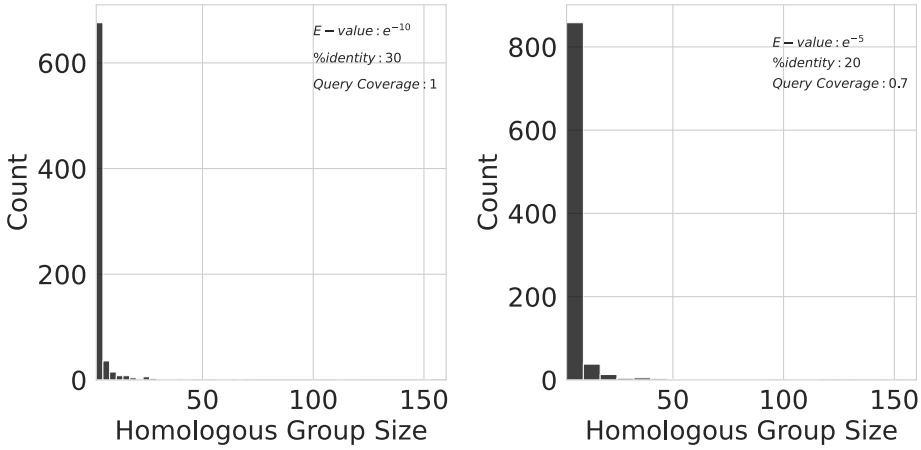

Figure S2: Histograms showing the distribution of homologous group sizes detected in the *D. plexippus* genome when using more (left) and less (right) stringent sequence similarity cutoffs.

Table S3: A table showing the number of duplicate and singleton genes identified in this study.

PSI-BLAST Parameters	n Homologs	n Singletons
e-value = e^-10, %identity = 30, query coverage = 1	3237	11995
e-value = e^-5, %identity = 20, query coverage = 0.7	9792	5440

3 Supporting Results

3.1 Analysis of transcriptional divergence using correlation-based distances

To ensure that our inference of transcriptional divergence between stages was robust, we performed an additional analysis using Pearson distances, as opposed to Manhattan distances (presented in the main text). We found that expression significantly varied by developmental stage ($F = 194.94$, $p < 0.001$) but not plant ($F = 0.87$, $p = 0.37$) (Figure S3). We then performed pairwise comparisons to test for differences between subsequent stages, as well as between larvae and adults. Following *D. plexippus* throughout metamorphosis: the transition form third instar to fifth instar involves some, but relatively few changes in gene expression ($F = 40.84$, $p < 0.001$). Then a substantial change in gene expression occurs during the transition from fifth instar to early pupa ($F = 182.74$, $p < 0.001$), followed by another substantial change from early pupa to late pupa ($F = 274.12$, $p < 0.001$). Finally, the transition from late pupa to adult involves another substantial change in gene expression ($F = 231.28$, $p < 0.001$). It is interesting to note that the difference between third instar and adults ($F = 458.90$, $p < 0.001$) is comparable to the difference between third instar larvae and early pupae ($F = 453.98$, $p < 0.001$). Likewise, the difference between fifth instar larvae and adults ($F = 217.05$, $p < 0.001$) is comparable to the difference between fifth instar larvae and early pupae. These findings are consistent with the analysis based on Manhattan distances, and highlight the same interesting point that pupae are approximately as transcriptioanlly diverged from larvae as adults are.

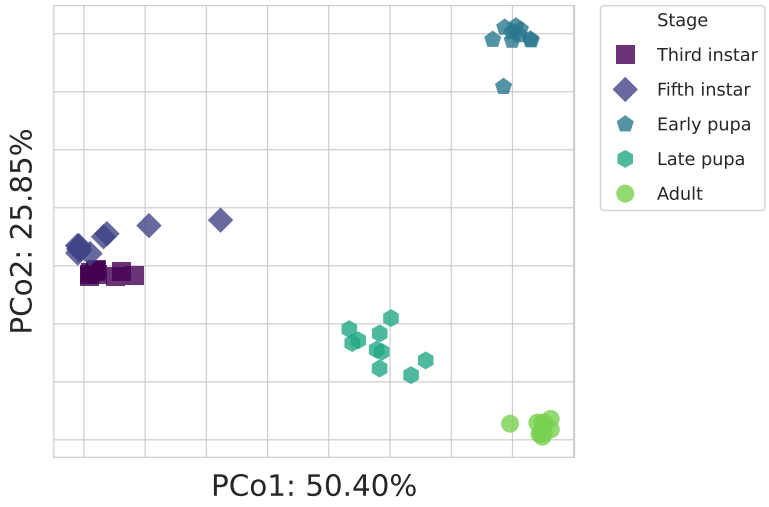

Figure S3: A principal coordinate analysis plot showing substantial transcriptional divergence between life stages. Each point represenst the global gene expression profile of an individual, and closer points indicate more similar gene expression profiles. Axis labels indicate principle coordinate rank and the proportion of explained variance.

To visualize of the lack of an effect of plant on transcriptional divergence, we also created a visualization of the Manhattan-distance based PCoA (Figure S4).

3.2 Reanalysis based on less stringent homology inference

To ensure that our findings were robust, we re-analyzed our data using less stringent sequence identity cutoffs to infer gene homology (see section 1.8). This included more divergent genes in our analysis, thus increasing the amount of phylogenetic diversity captured.

3.2.1 Correlations between phylogenetic and expression pattern diversity

The overall relationships between phylogenetic and expression pattern diversity were consistent with our primary analysis (Figure S5).

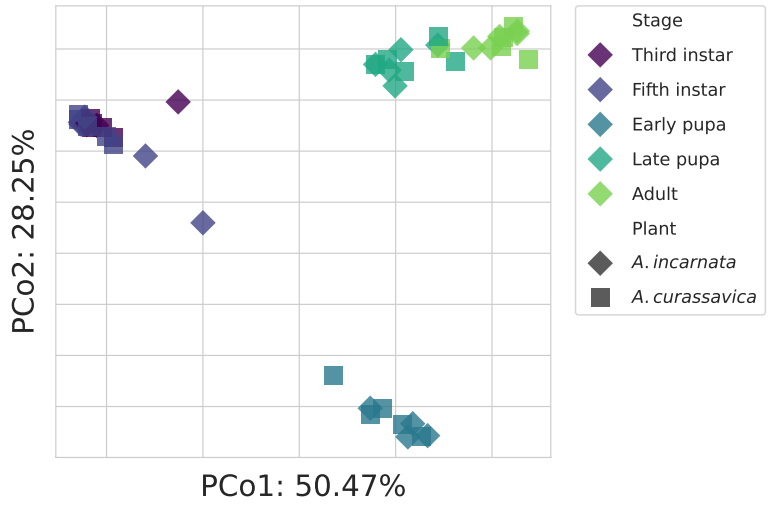

Figure S4: A principal coordinate analysis plot showing substantial transcriptional divergence between life stages, and a lack of differentiation based on which plant larvae were reared on. Each point represents the global gene expression profile of an individual, and closer points indicate more similar gene expression profiles. Axis labels indicate principle coordinate rank and the proportion of explained variance.

Table S4: A table showing the results for each correlation between phylogenetic and expression pattern diversity that were summarized in the main text.

Homologous group size	ρ	p-value
Total	0.8345044	$< 2.2e-16$
4	0.3434008	0.01234
5	0.18	0.1938
6	0.1178571	0.3382
7	0.6454545	0.0185
8	0.3212121	0.1838
9	0.2	0.3917
11	-0.5428571	0.8792
13	-0.5	0.825

Table S5: A table showing the results for each correlation between phylogenetic and expression pattern diversity that were produced by our reanalysis based on less stringent similarity cutoffs for homology inference.

Homologous group size	ρ	p-value
Total	0.8268648	$< 2.2e-16$
4	0.3319884	0.0007816
5	0.3084583	0.02998
6	0.03387097	0.4282
7	0.3169231	0.0614
8	0.6363636	0.02722
9	0.3818182	0.1395
10	0.04242424	0.4593
12	-0.6	0.8833
14	-0.5	0.825
16	0.3	0.3417

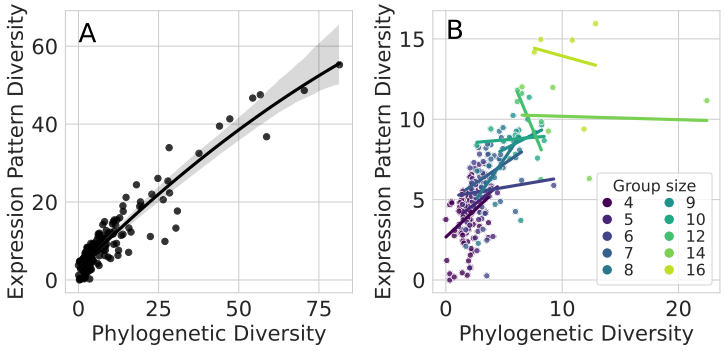

Figure S5: The relationship between phylogenetic and expression pattern diversity A) across all homologous gene groups, B) within each homologous gene group size. In A, the solid black line depicts the fit polynomial model, the light gray area indicates the 95% confidence interval for said model. In B, each line represents the linear model fit to each homologous group size (only group sizes will 5 or more replicates were considered in this analysis)

3.2.2 Expression pattern specificity

Our comparisons of stage-specificity between duplicate and singleton genes based on less stringent similarity cutoffs for homology inference were consistent with the analysis presented in the main text. Specifically, genes that are part of homologous groups tend to show increased stage-specificity relative to singleton genes ($D = 0.13$, $p < 2.2 * 10^{-16}$). Although this pattern is statistically supported, we note that the effect size is slightly smaller than the analysis presented in the main text analysis.

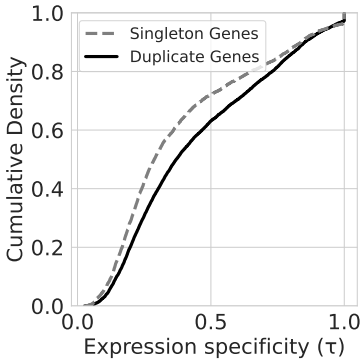

Figure S6: The empirical cumulative density functions of expression specificity values τ for duplicate (solid line) and singleton (dashed line) genes. Higher expression specificity values indicate more stage-specific expression patterns.

3.3 Broad functional overview

To gain a general sense of what high level functional differences occurred between stages, we used the KEGG [15] to infer gene functions and examined the relative transcriptional investment in the highest level KEGG BRITE groupings (Figure 6). Genes that were classified to multiple high-level KEGG categories were excluded from the analysis for more conservative estimates of functional investments.

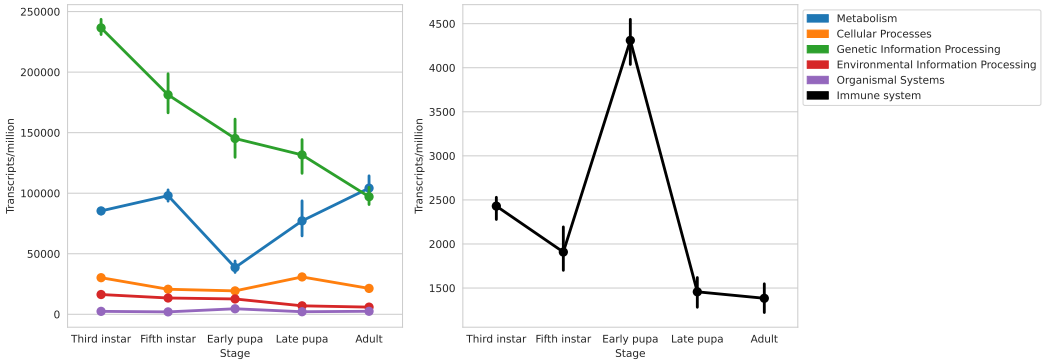

Figure S7: A line plot showing the relative transcriptional investment (transcripts per million) in each high level functional group across life stages. Note that these groupings reflect the overall transcriptional investment in each broad functional group listed, not the activity of individual pathways or genes. Therefore, each individual pathway or gene within each group is not expected to necessarily follow the exact trend exhibited by the whole group. Error bars represent 95% confidence intervals calculated across individual samples.

References

[1] W Tan, et al., Transcriptomics of monarch butterflies (*Danaus plexippus*) reveals that toxic host plants alter expression of detoxification genes and down-regulate a small number of immune genes. *Molecular Ecology* **28**, 4845–4863 (2019).

[2] S Andrews, FastQC: A Quality Control Tool for High Throughput Sequence Data. (2010).

[3] S Zhan, C Merlin, J Boore, S Reppert, The Monarch Butterfly Genome Yields Insights into Long-Distance Migration. *Cell* **147**, 1171–1185 (2011).

[4] ZB Abrams, TS Johnson, K Huang, PRO Payne, K Coombes, A protocol to evaluate RNA sequencing normalization methods. *BMC Bioinformatics* **20**, 679 (2019).

[5] GP Wagner, K Kin, VJ Lynch, Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences* **131**, 281–285 (2012).

[6] R Core Team, *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria), (2022).

[7] J Oksanen, et al., *vegan: Community Ecology Package*. (2022).

[8] S Altschul, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402 (1997).

[9] RC Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792–1797 (2004).

[10] BQ Minh, et al., IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).

[11] S Kalyaanamoorthy, BQ Minh, TKF Wong, A Von Haeseler, LS Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**, 587–589 (2017).

[12] E Paradis, K Schliep, ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

[13] I Yanai, et al., Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).

[14] M Cardoso-Moreira, et al., Gene expression across mammalian organ development. *Nature* **571**, 505–509 (2019).

[15] M Kanehisa, S Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28** (2000).