**Genome divergence in Lepidoptera provides insight into the relationship between rate of molecular evolution and diversification.**

**Abstract:**

Diversification rates are linked to rates of molecular evolution in a diverse array of taxa across the tree of life, but the underlying cause of the relationship is debated.

Characterising the extent and cause of this phenomenon is of great interest for understanding the link between microevolutionary processes and the macroevolutionary patterns of diversification. A number of theories have been proposed linking the rate of genome change to spatial patterns of biodiversity, suggesting that factors that increase the mutation rate should increase genome-wide substitutions rates, increasing the rate of formation of reproductive isolation through genome incompatibility and thus increasing the speciation rate. Testing these predictions is important for informing new evolutionary models for use in phylogenetics, molecular dating and comparative evolution. Lepidoptera (butterflies and moths) provide an interesting test case because, in many cases, speciation is thought to be driven by strong selection on a small number of loci, rather than gradual genome-wide divergence. We collate phylogenies for all of Lepidoptera to demonstrate that the association between host plant diversity and lepidopteran diversity is general across the whole order, not confined to well-studied groups, suggesting that diversification rates are driven by adaptation to shifting hosts.

**Introduction**

The pace of evolution varies throughout the tree of life. Dissecting the factors that drive variation in the tempo and mode of evolution is central to understanding the mechanics of evolution. It is also a critical factor in the construction of evolutionary models essential for phylogeny reconstruction (Arenas 2015), molecular dating (Bromham et al. 2018; dos Reis et al. 2016; Kumar and Hedges 2016), and process-based models of ecology (Graham et al. 2004; Guillory and Brown 2021; Heibl and Calenge 2013) and biogeography (Landis et al. 2020; Ronquist and Sanmartín 2011). One area that requires further investigation is the relationship between the rates of change at the genomic level and the rate of generation of biodiversity (Hua and Bromham 2017).

Several general theories of biodiversity generation link rates of genome change and speciation rates. For example, the Metabolic Theory of Ecology proposes that greater environmental energy will lead to increased metabolic rate, which will increase the rate of molecular evolution (either directly through an influence of metabolism on mutation rates, or indirectly through decreased generation times) which will lead to faster accumulation of genomic incompatibilities between diverging populations and therefore a faster rate of speciation (Allen et al. 2006). Similarly, the Evolutionary Speed Hypothesis suggests that greater rates of molecular evolution in warmer environments should lead to faster speciation rates, but through more substitutions driven by selection for ecological divergence (Gillman and Wright 2014). Examining the relationship between rates of genome change and generation of biodiversity provides a key mechanism for testing the generality and explanatory potential of these theories.

Clarifying the nature of the relationship between substitution rates and diversification rates will not only shed light on the drivers of tempo and mode of evolution, it is also essential for accurate analysis of the patterns of diversification. All macroevolutionary phylogenetic methods rest on the assumption that rates of molecular change that inform phylogeny construction are independent of processes of diversification. Violation of this assumption can cause serious errors in macroevolutionary analysis (Duchêne et al. 2015; Duchêne et al. 2017; Ritchie et al. 2020).

There is broadscale empirical evidence for an association between rate of molecular evolution and diversification rate (Bromham 2024). Phylogenetic studies have demonstrated an association between substitution rate and net diversification rate for a wide range of taxa, including reptiles, birds, fish and plants. This relationship has been detected in a wide range of datasets using a variety of analytical approaches, including whole-tree analyses that correlate root-to-tip path lengths with number of nodes along the path (Ezard et al. 2013; Pagel et al. 2006; Webster et al. 2003) and sister pairs analysis that correlate the average substitution rate with species richness in reptiles (Eo and DeWoody 2010 ), birds (Iglesias-Carrasco et al. 2019; Lanfear et al. 2010), angiosperms (Bromham et al. 2015; Lancaster 2010). The relationship between rate of molecular evolution and species richness has been detected over a wide range of taxonomic scales, for example comparing genera within the plant family Proteaceae (Duchene and Bromham 2013), between families of angiosperms (Bromham et al. 2015) and across animal phyla, orders and classes (Fontanillas et al. 2007).

However, the mechanism underlying these relationships is a matter of debate. It has been suggested that speciation could cause a burst of substitutions, increasing phylogenetic path lengths that include many speciation events (Venditti and Pagel 2010; Venditti and Pagel 2014). Alternatively, it has been argued that since species richness is associated with synonymous substitution rate in some taxa, variation in mutation rate must influence the rate of genome divergence, resulting in faster evolution of reproductive isolation between diverging populations (Lanfear et al. 2010). While comparative studies support a general relationship between genomic divergence and rate of speciation, they don't directly reveal the causal mechanism (Figure 1). Furthermore, the relationship does not appear to be universal, as it is weak or absent in some taxa. Is the failure to detect an association between rate of molecular evolution in some datasets due to low power, or because the relationship is not universal? It has been suggested that absence of a relationship between species richness and rate of molecular evolution might be due to variation in the underlying genetic basis of speciation in different lineages (Goldie et al. 2011). It would therefore be useful to examine the relationship between molecular rates and diversification in groups showing different modes of speciation.

Allopatric speciation occurs when a population is subdivided such that members of the subpopulations do not interbreed, so that genetic variants are no longer shared across the different parts of the population. In this case, a relationship between substitution rate and speciation rate is predicted under the Dobzhansky-Muller Incompatibility (DMI) model, due to gradual accumulation of substitutions which are not strongly deleterious in the population in which they arise, but may have negative effects when combined with substitutions in other populations, reducing the viability or fertility of hybrids between populations. If genomic incompatibility is generated through multiple substitutions of roughly equal effect, and every substitution has an equal chance of generating incompatibilities, this could generate a predictable relationship between substitution rate and speciation rate (Orr and Turelli 2001). There is empirical support for this model from a range of study systems (Dufresnes et al. 2021; Matute et al. 2010; Moyle and Nakazato 2010). However, the form of the relationship will change depending on the nature of speciation (ref).

While the basic formulation of the DMI predicts a “snowball effect” as incompatibilities accumulate exponentially, the relationship may be more linear if incompatibility is due primarily to compensatory substitutions that have occurred in pairs within each lineage (Presgraves 2010). Speciation may be completed by selection for reproductive isolation traits that reduce the chances of mating with members of the other population, reducing the incidence of less fit hybrids.The relationship between genetic divergence and speciation will also depend on whether incompatible substitutions accumulate in allopatry (and therefore determined by mutation rate and population dynamics) or in sympatry (therefore more likely to be driven by few substitutions each of large effect) (Scopece et al. 2007). Ecological speciation driven by adaptation to different niches should also change the relationship between substitution rate and speciation rate, driving population separation through relatively few adaptive changes rather than waiting for incompatibilities to accumulate (ref).

Therefore, we may be able to get some explanatory traction on the relationship between genomic divergence and speciation rate if we focus on case studies characterised by different modes of speciation. We might expect to observe a strong relationship between substitution rates and diversification rates for many taxa in which the predominant speciation mode is the gradual accumulation of incompatible substitutions in allopatry generating postzygotic incompability. But we might expect little or no relationship between substitution and diversification in groups where the predominant mode of speciation is selection on few key loci where a small number of substitutions have each of large and specific effects on prezygotic isolation .

Butterflies and moths are an ideal test case for this hypothesis, for several reasons. Firstly, diversification patterns in Lepidoptera are often attributed to strong selection for pre-zygotic isolation, for example due to host plant shifts, mimicry or symbioses, rather than being primarily driven by allopatry and gradual accumulation of substitutions that contribute to post-zygotic incompatibility (refs). Three major hypotheses have been articulated to describe how host plant shifts influence lepidopteran speciation (Jousselin and Elias 2019): (1) speciation occurs in bursts following adaptation to overcome the chemical defences of major plant groups (escape and radiate, Ehrlich and Raven 1964); (2) speciation occurs through the gradual expansion of diets to closely related species, leading to range expansion followed by fragmentation into specialist subpopulations (Oscillation, Ref); (3) speciation occurs upon colonisation of new host species, and that related taxa compete for a relatively narrow set of host plants that can be exploited (musical chairs, Hardy and Otto 2014).There is evidence for associations between lepidopteran diversity and the number of exploited host plant taxa (host breadth) (Janz et al. 2006; Wang et al. 2017), and between diversification and the proportion of generalists (Hardy 2017; WEINGARTNER et al. 2006).

Secondly, the relationship between speciation and genetic change has been well-studied in Lepidoptera. A recent phylogenetic study showed that lineages with host plant shifts gave rise to bursts of diversification as well as to a larger number of genes experiencing positive selection (Allio et al. 2021). Other authors have noted that the Lycaenid butterfly group, whose rapid diversification has been partly attributed to its recently evolved ant symbioses (Schär et al. 2018), also has notably higher substitution rates (Pellissier et al. 2017; Pellissier et al. 2012).

* However, the potential association between diversification rates, adaptation and neutral substitution rates has yet to be characterised across the Lepidopteran genome and family tree.
* If the generation of diversity in Lepidoptera is typically driven by strong selection for mate choice, mimicry or host plant, then reproductive isolation is expected to be generated by selection on relatively few loci that generate pre-zygotic barriers to gene flow.
* If this is generally true, then we would not expect a strong relationship between genome-wide substitution rates and species richness in Lepidoptera.

Thirdly, Lepidoptera are a data-rich case study for investigating the causes and consequences of diversification. The major taxonomic groupings are well-characterised, with wealth of genomic-scale data informing both an understanding of the genetics of speciation and the inference of phylogeny (Allio et al. 2020; Espeland et al. 2018; Hamilton et al. 2019; Mitter et al. 2017; Wiemers et al. 2020). A long history of professional and amateur study has resulted in a rich understanding of the biology and geography of Lepidoptera (Mackintosh et al. 2019). Genetic diversity and species diversity vary greatly between lineages (Beck and Fiedler 2008; Ehrlich and Ehrlich 1978; Mackintosh et al. 2019). The group contains many well-studied cases of microevolutionary processes, for example industrial melanism of the peppered moth (Cook and Saccheri 2013), radiation tolerance of Fukushima pale grass blue butterflies (Nohara et al. 2017), and speciation with host plant preference in *Heliconius* (Jiggins 2017), as well as macroevolutionary studies of patterns and processes of diversification (refs). Genomic data is available for many Lepidopteran linages so we can evaluate the relationship between substitution rates and diversification rates over a large number of loci with greater power than earlier studies that relied on smaller numbers of housekeeping genes.

Lepidoptera provide an opportunity to test whether the association between diversification rate and rate of molecular evolution is found in invertebrate lineages. They also provide a useful test of the drivers of the relationship. If diversification rate in Lepidoptera is driven predominantly by host shifts, then it may depend predominantly on relatively few substitutions in key genes. Unless adaptation is mutation limited, this pattern of sympatric speciation may reduce the likelihood of a link between genome-wide rates of molecular evolution and diversification rate.

Here we employ a sister pairs approach to examine the association between host plant shifts, diversification and rates of molecular evolution in Lepidoptera. First we ask if net diversification rate is associated with host switches, then we ask if there is any association between either the supply of variation and diversification. In this way we can examine whether the nature of the association between substitution rates and diversification rates in cases where speciation is driven by selection to diversify resource use. We collate 24 published phylogenies and five different measures of host plant diversity to compile the largest dataset used to test for an association between host plant diversity and speciation rate in Lepidoptera. We then use two molecular datasets to examine the association between substitution rates, diversification rates, host plant diversity and specialisation in butterflies. One molecular dataset has fine-scale resolution, with nine nuclear protein-coding sequences (XX kb) representing 994 genera of butterfly (Chazot et al. 2019), the other has broad coverage, with 2098 nuclear genes (2.2 Mb) for 195 species representing 81 lepidopteran families (around half of all families in the order).

***Analysis rationale***

For many groups there has been an observation of a connection between diversification rate and rate of molecular evolution.



There are three broad explanations of this link.

A diagram of different types of text

Description automatically generated

One is that molecular change drives speciation, because accumulation of substitutions leads to hybrid incompatibility, so faster substitution leads to faster speciation rate (DMI model). We expect this to be predominantly associated with reproductive isolation in allopatry. If this is the case, then we expect an association between dN and clade size. The observation of a link between dS and clade size supports this path because mutations feed nearly neutral substitutions. There may or may not be a signal with dN/dS, depending on how whether allopatry leads to temporarily or permanently decreased population sizes (bearing in mind that dN/dS is not a great signal of anything much because so many things can affect it).

A diagram of a path

Description automatically generated

The second path is that the process of speciation drives faster rates of molecular evolution. For example, in Lepidoptera, speciation is often associated with a shift in host plant, or mimicry or sexual selection. These shifts are expected to affect relatively few loci in the genome: few genes, and few substitutions within those genes - so would not be expected to generate an association between genome-wide rates of change and diversity (unless hitchhiking is sufficient to generate genome wide substitution rate increases). Alternatively, host shifts and host specialisation could potentially result in reduction in average population size compared to widespread or generalist taxa so an increase in dN/dS.

A diagram of a path

Description automatically generated

And alternative (third) path is that both clade size and rate of molecular evolution are related to something else. Not sure what, in butterflies, but temperature/latitude might be a good thing to test (which would require some kind of clade-based measure of these things from distribution data).

What we need to know.

(1) Is rate of molecular evolution associated with clade size in lepidoptera: testing dS, dN (and dN/dS?).

* If dS is related to clade size, this supports path 1 (unless you want to make a case that reduction in population size erodes DNA repair through accumulation of mildly deleterious substitutions in which case you also expect dN and dN/dS to correlate with clade size). If adaptation is mutation limited then we could also see an association between dS and speciation rate.
* If dN/dS is related to clade size, then this could be compatible with path 1 or path 2 (if selection reduces population size).
* If dN is related to clade size, but not dN/dS, this is best explained by path 1.
* If none of the molecular rates are correlated with clade size then this is compatible with path 2.

(2) if none of the molecular rates are correlated with clade size, then we would like to know whether this is because the speciation mode in lepidoptera is not based on DMI accumulation of substitutions OR because of low power. So we want to ask if the data we have for Lepidoptera is compatible with speciation rates driven by adaptation relying on few genomic changes in which case we do not expect to see a relationship between substitution rates and rates of molecular evolution, or with a mechanisms that might affect rates of molecular evolution such as allopatry by host differentiation. We can consider several models of diversification in lepidoptera that have been proposed in the literature, and consider what predictions they make: Escape and Radiate, Oscillate, and Musical Chairs.

**I. Escape and radiate:** any lepidopteran lineage that can adapt to a new host type opens up new niches to radiate into. In particular, a genetic change that acts as the “key to the door” for a whole plant lineage could trigger a diversification event. In this case, a clade that has changed to access a new family should show a burst of speciation. For a sister pairs analysis, we are looking for a case where one clade has access to a wider range of families of hosts than its sister. There are two ways to measure this. One is by novel families that are not shared. If we have a close outgroup we can measure gain of families, if not we have to just measure different.

Without outgroups (sister comparison only)

Count host diversity: Here we assume number of families per clade represents the number of escape and radiate pulses.

A diagram of a number of host fans

Description automatically generated

Count host jumps: For this we could assume that any host family found in both sisters is ancestral, and only unique families not shared with other family represents a jump.

A diagram of a number of host jumping

Description automatically generated

If it is practical to add a close outgroup then we can refine the measure a little, because we can detect retention from ancestor, but its possibly not worth it unless we already have outgroups chosen and they are relatively close. For example here we can use the outgroup to show that the two families in clade A are ancestral retentions not host jumps, so B has more host jumps even though it has fewer host families.

A diagram of a number of host groups

Description automatically generated

We might also predict under this scenario that species richness of leps is associated with greater PD across hosts (independent of host number – that is, we want a measure of how distanly related the hosts of a lep clade are, independent of the number of host species or lep species. One way to do this is to compare the PD of all name hosts of the sister clades (that is, all the host plants of the species known to be associated with a sister clade, regardless of whether we include them in our sister pairs or not). We have to standardize the PD: we can do this by generating a null distribution of PD for different host numbers: make a tree of the list of all named hosts in our dataset and, for a sister clade of N species, draw 100xN species from the AllHosts tree, then compare PD of sisters to null PD for N species (it will be either positive if hosts are more distant than expected or negative if hosts are less distant than expected). Results are consistent with escape and radiate if

1. the clade with the greater value (on absolute scale) of host PD has more lep species (weak form) or
2. the clade with the greatest positive value (above expected host PD) has more lep species (stronger form).

We don’t expect a relationship between dN and clade size because the radiation is likely to be triggered by relatively few key genetic changes which are unlikely to be in the loci analysed or of large enough number to cause a significant difference in overall substitution rates (UNLESS there is a lot of hitchhiking).

A diagram of a network

Description automatically generated

**II. Oscillate**

Generalists expand host range, then give specialists as offshoots.

More generalists => more species (according to Hardy and Otto, though this seems a little odd as the generalists generate specialists which might cause the generalists to be lost by transtion to specialism?

A diagram of links and links

Description automatically generated

**III. Musical chairs**

Speciation occurs on colonization of a new host species

More hosts = more lep species

Because its one-to-one relationships between hosts and leps that generate (so more hosts = more niches) expect reduced or negative association with number of generalists.

A diagram of a chair

Description automatically generated

Putting it all together you get something like this for a path diagram (grey lines only apply if hitchhiking can drive genome-wide substitution rates or if substitution rates are estimated from loci under selection).

A diagram of a network

Description automatically generated

**Methods**

***Sister pairs***

We use sister pairs as a general and robust approach to inferring the association between rates of molecular evolution and diversification rates (Barraclough and Savolainen 2001; Bromham et al. 2015; Davies et al. 2004; Lanfear et al. 2010). Sister pairs are clades defined by a most recent common ancestor that is not shared with any other such pair in the analysis, so they provide a phylogenetically independent set of contrasts for evaluating associations between lineages. Since both clades in a sister pair share a common ancestor, they began from the same starting point with identical values for all biological parameters, then since then have had the same amount of time to accumulate diversity and substitutions. Sister pairs analyses have the advantage of not requiring estimates of divergence times or knowledge of the deeper-level phylogeny, so they can be applied to widely available phylogenetic information and combined across different published phylogenies.

We generated three sets of sister clade comparisons (Table A). For the analysis of the relationship between molecular and diversification rates, we select pairs of sister taxa from two molecular data sets: one sampled at the genus level and restricted to butterflies (Papilionoidea; Chazot et al. 2019) and one sampled across Lepidoptera at the family level (Kawahara et al. 2019). These contain about half of all butterfly genera and lepidopteran families, respectively. To test for an association between lepidopteran species richness and host plant diversity, we require a data set that is sampled across all Lepidoptera at the finest resolution possible, so we construct a third data set by selecting sister clades from a range of published Lepidopteran phylogenies of single families or superfamilies. This contains a combination of pairs taken at the tribe, subfamily and family levels (Table A). To construct a data set for the analysis of host plant association across all Lepidoptera, we compile a set of 22 published phylogenies comprising families or superfamilies (Table A, III). Where no digital version of the tree was available, phylogenies were extracted from published diagrams using TreeSnatcher Plus (Laubach et al. 2012). For the method of sister pairs, it is not necessary to resolve the deep-level relationships between the subsets of taxa included in any of the phylogenies.

To select sister pairs from a phylogeny, we collapse all tips that belong to the same taxon (family, genus, tribe or subfamily) to generate a tree where each tip represents one taxon. We do not collapse taxa that do not comprise a monophyletic clade in the tree. We then select all pairs of tips (each representing a family, subfamily, tribe or genus) that are each other’s closest relatives, so that each pair shares a most recent common ancestor not shared by any other tip (ignoring those tips that don't represent monophyletic taxa). For the family-level tree, we check the chosen pairs using a review of previous literature given in Mitter et al. (2017). Since studies at lower taxonomic levels include more intra-family evidence, where a lower-level study has supported a sister family relationship that is not mutually monophyletic in the Kawahara et al. (2019) tree, we added that pair to the analysis. Where a pair of sister families is not mutually monophyletic and there is uncertain support for it as described in Mitter et al. (2017), we do not use that comparison. By this method, we retain the family-level comparisons of Sphingidae-Saturniidae and Pyralidae-Crambidae on the strength of literature evidence, even though they are not monophyletic in Kawahara et al (2019).

Studies relating molecular branch lengths to diversification dynamics must consider the node-density effect which could create spurious correlations between the number of inferred substitutions in a lineage and the number of speciation events, due to the tendency to under-estimate the number of substitutions on long unbroken branches (Hugall and Lee 2007; Venditti et al. 2006). To ameliorate the node-density effect in the family level dataset, we sample the clade in each pair with the most tips down to match the size of its sister clade. For each pair of sister families, we identify all tips in the alignment that are included in those clades in the published tree. If one sister clade has more tips than its sister, we randomly sample tips so that both sister clades have the same amount of tips. To provide outgroups for branch length estimation, we form a triplet from each pair by adding the nearest single tip (measuring distance in substitutions per site) from the full taxon-level phylogeny. The list of comparison selected is given in Supplementary Tables SA, SB, and SC.

***Species richness***

The relative number of described species in two sister clades is a proxy for the difference in their average diversification rates (Barraclough et al. 1996; Bromham et al. 2015; Cardillo 1999). We extract valid species names from the Global Lepidoptera Names Index (LepIndex; Beccaloni et al. 2003). For each family, subfamily, tribe or genus represented in our sets of sister clades, we accept all unique binomials with the ‘Current Status’ field listed as ‘Valid Name’. If a taxon is not represented in LepIndex, the corresponding sister pair is dropped from the analysis. Species counts for each clade are given in Supplementary Tables SA, SB, and SC.

***Host plant diversity***

Host plant data was obtained from the HOSTS index, which lists known host plant species exploited by lepidopteran species (Robinson et al. 2010). Host species breadth is represented by the number of named angiosperm species exploited by at least one member of the taxon (Total Host Species). We also calculate the mean number of hosts per species (Mean Hosts/Species) for each sister clade, by dividing the number of host plant species (Total Host Species) by the number of species in the clade (see tables S1, S2, S3 for clade sizes). As a measure of the number of host shifts between major plant groups, we calculate host family breadth for each Lepidopteran higher taxon as the number of plant families that are used as hosts by at least one member of that taxon (Total Host Families). As an alternative measure of host breadth, we calculate the phylogenetic diversity of host plants colonised by the taxon, we also calculate the Faith’s phylogenetic diversity index (FPD; Faith 1992) of the associated host plant species (Host Phylodiversity). This is calculated as the sum of molecular branch lengths connecting all associated host plant species for each clade using the recent maximum likelihood angiosperm megaphylogeny of Janssens et al. (2020). The proportion of generalists for each Lepidopteran taxon is a measure of within-species diet breadth or phenotypic plasticity (Hardy 2017; Hardy and Otto 2014; Wang et al. 2017). It is calculated as the proportion of taxa in a sister clade that exploit more than one host plant species.(Prop.Generalists).

***Molecular data***

For the genus-level butterfly data set (I, Table A), we use the molecular phylogeny of (Chazot et al. 2019). This data set contains 994 tips, each representing a butterfly genus. The tree is restricted to superfamily Papilionoidea, but about half of the constituent genera are represented. Chazot et al. (2019) conducted numerous analyses intended to explore the effect of priors and data set construction on the inferred phylogeny.

* For this project, we choose the tree inferred by their core analysis based on 9 nuclear gene fragments with topology was inferred by maximum likelihood in RAxML (Stamatakis 2014) and divergence times inferred using Bayesian Inference in BEAST 1 (Suchard et al. 2018).
* Total alignment length: perhaps add this to the comparisons table since it varies between pairs?
* It would also be useful to include the number of tips per comparison, as opposed to the number of species counted per clade. From the description its not obvious whether most comparisons have only one tip each or many.

For the family-level data set (II, Table A), we make use of the phylogenomic data set of Kawahara et al. (2019) based on a transcriptomic supermatrix alignment of 2098 nuclear genes (2.2 Mb) for 195 tips representing about half of Lepidopteran families (81 families) curated to remove gap-only codon sites and loci with less than 70% taxon coverage. This results in near-complete, pre-aligned data for all taxa in the phylogeny. In the analysis of Kawahara et al. (2019), this data was recoded to remove synonymous signal, but for the purposes of codon model analysis we use the complete ACGT-coded data. This means that the branch lengths estimated in our study may not match those estimated in the maximum likelihood analyses of Kawahara et al. (2019), even for those based on nucleotide data. The study produced several phylogenies using different methods. We use the topology and branch lengths in units of time from the main text of Kawahara et al (2019, Fig. 2), which was inferred using maximum likelihood from a concatenated amino acid alignment.

***Substitution rates***

Each comparison consists of a pair of sister clades, each with the same number of sampled tips, plus the outgroup (see Tables S1, S2, S3). We use the alignments from Kawahara et al. (2019) and Chazot et al. (2019) to compile a concatenated alignment of coding sequences for each comparison, removing all incomplete codons and missing genes. This resulted in different alignment lengths for each comparison, but no alignment was shorter than 4 kb. Synonymous substitutions (dS) and nonsynonymous substitutions (dN) are estimated for all branches on the tree using the program ‘codeml’ in the PAML software suite (v4) (Yang 2007). A separate codon model was fitted to each sister clade using the branch model in ‘codeml’ (model=2), with the background model applying to the ancestral branch of the sister pair and to the outgroup. As a check on the codon model analysis, and to increase power by pooling data, we also estimate Total substitutions using the program ‘baseml’ with a general time-reversible (GTR) substitution model (Yang 2007). Substitutions are estimated on the branches of the published topology for each of the Family, Genus and Major Lineages data sets. For the family- and genus-level data sets (Table A, I and II), we estimated phylogenetic average substitution rates (in substitutions/site) by successively collapsing tips and adding the average of their substitutions to the parent branch. This phylogenetic average was used as the final substitution rate value for calculating contrasts.

***Analysis***

Contrasts were calculated for all clade size, molecular rate and host plant variables by calculating the difference between the values for the two sister clades in each comparison after appropriate transformation. All variables were log-transformed except for proportion of generalists, which was arcsin-square root transformed due to being a proportion constrained between 0 and 1. Contrasts were calculated so that the value of each variable corresponding to the smaller clade was always subtracted from the value corresponding to the larger clade. The orientation of contrasts with a clade size difference of 0 was randomised. Univariate linear regressions were conducted with contrast in clade size as the dependent variable, using contrasts in each of the molecular rates and host plant variables as the dependent variable. Because the orientation of clades is arbitrary, all regressions were conducted through the origin.

For substitution rate contrasts, we additionally performed the filtering procedure of Welch and Waxman (2008) to check for an effect of pair age on the variance of the inferred substitution rate. This effect can violate the assumption of homoskedasticity and produce a confounding negative relationship between contrasts and pair age. To do this, we tested for a significant linear relationship between the square root of the age of the pair and the absolute size of the substitution rate contrast. If a relationship was found, we removed the shallowest pair and repeated the test until no such relationship remained. The remaining, filtered sister pair data sets formed the basis of all substitution rate analyses.

Multiple regression.

Path analysis?

**Results**

Our analysis provides strong support to the hypothesis that diversification rate in butterflies and moths is associated with the number and diversity of host plants. We show that this relationship holds for not only for the butterflies (Papilionoidea) but also across the whole of the Lepidoptera (Table B). Not only is lepidopteran species richness related to the number of host species (in butterflies and lepidopteran families) but also the phylogenetic diversity of host species (that is, how taxonomically divergent host plants are). Indicators of generalism – the mean number of hosts per species and the proportion of generalists – are significantly associated with species richness in lepidopteran families. There are no significant associations between any of the measures of substitution rate (dS, dN or total substitutions) and rates of molecular evolution (Table C). However, rates of non-synonymous substitutions are significantly positively associated with host family diversity for comparisons between families of Lepidoptera (Table D).

Multiple regression.

Path analysis?

**Discussion**

* Our analysis confirms that a general association between species richness in butterflies and moths and host plant diversity across the whole order Lepidoptera.
* Many studies of the drivers of variation in rate of diversification in Lepidoptera have largely focused on phytophagous habit and host plant specificity, particularly in butterflies (Mitter et al. 2017).
* Butterflies in particular are a model group for the study of insect-plant coevolution, many having very close relationships to their angiosperm hosts (Ehrlich and Raven 1964) which are frequently phylogenetically conserved (Menken et al. 2010; van der Linden et al. 2021).
* Mechanisms of speciation in Lepidoptera are debated, particularly with regard to the role of host plant shifts (Ehrlich and Raven 1964; Hardy 2017; Hardy and Otto 2014; Jousselin and Elias 2019).
* Phylogenetic diversification rate analysis does find a role for host plant differences in explaining varying diversification rates (Allio et al. 2021; Condamine et al. 2012; Fordyce 2010), but also indicates a historical effect of climate and alpine environment (Condamine et al. 2018).
* Faster diversification rates in the tropics (Cardillo 1999; Condamine et al. 2012) could be a function of increased host plant diversity, so that host specialisation would be the major driving force of butterfly species richness (Novotny et al. 2006). In this scenario, higher tropical speciation rates could be driven by mechanisms such as the repeated evolution of generalisation, followed by rapid population expansion caused by the availability of new resources not available to competitors, and finally by divergent evolution for greater specialisation in isolated populations (Scriber 2010).
* Any of these factors could also be mediated through life history and morphology, since temperature, altitude and host plant characteristics impact many insect characteristics such as body size, fecundity, and incidence of wingless or brachypterous forms (Hodkinson 2005).
* Adding to this complexity, both climate and host plant richness may also have different effects for generalists and specialists (Menéndez et al. 2007).
* Aside from host shifts, another factor potentially associated with higher speciation rates in insects is sexual conflict, which may cause speciation through divergent sexual selection (Arnqvist et al. 2000).
* Sexual conflict in Lepidoptera is common and driven by polyandry, sperm precedence of the most recent mate, and mating plug deposition by males (Ehrlich and Ehrlich 1978). However evidence for this is lacking at present in butterflies (Carvalho et al. 2020).

Furthermore, the theory behind Lepidopteran speciation is still debated, particularly with regard to the role of host plant shifts (Ehrlich and Raven 1964; Hardy 2017; Hardy and Otto 2014; Jousselin and Elias 2019). Three major hypotheses have been articulated to describe how host plant shifts influence Lepidopteran speciation. (Jousselin and Elias 2019). These propose (1) that speciation occurs in bursts following adaptation to overcome the chemical defences of major plant groups (escape and radiate, Ehrlich and Raven 1964); (2) that speciation occurs through the gradual expansion of diets to closely related species, leading to range expansion followed by fragmentation into specialist subpopulations (oscillation); or (3) that speciation occurs upon colonisation of new host species, and that related taxa compete for a relatively narrow set of host plants that can be exploited (musical chairs, Hardy and Otto 2014).Empirical relationships have been found between diversity and the number of exploited host plant taxa (host breadth) (Janz et al. 2006; Wang et al. 2017), and a relationship is also suggested between diversification and the proportion of generalists, which is often taken as supporting the oscillation hypothesis (Hardy 2017; WEINGARTNER et al. 2006).

Of particular interest in Lepidoptera is that the same patterns associated with macroevolutionary variation also appear to be associated with molecular adaptation. In a recent phylogenetic study, lineages with host plant shifts gave rise to bursts of diversification as well as to a larger number of genes experiencing positive selection (Allio et al. 2021). Other authors have noted that the Lycaenid butterfly group, whose rapid diversification has been partly attributed to its recently evolved ant symbioses (Schär et al. 2018), also has notably higher substitution rates (Pellissier et al. 2017; Pellissier et al. 2012). However, the potential association between diversification rates, adaptation and neutral substitution rates has yet to be characterised across the Lepidopteran genome and family tree.

These relationships could potentially confound or mediate a relationship between molecular evolution and diversification. For instance, if the primary driver in Lepidopteran diversification at the family level were adaptive radiation caused by access to major host plant groups, this may induce large differences between the diversity of related clades based on single adaptive events that could serve to mask the more subtle effect of substitutions in promoting speciation due to isolation. Microevolutionary mechanism could also affect our expectations for a relationship between molecular evolution and diversification. For example, if the oscillation hypothesis is true as originally described and speciation is driven by isolation in populations of generalist species, we should expect to see diversification rates related to both substitution rate and rates of generalism (Jousselin and Elias 2019). However, if instead speciation is driven by divergent selection on populations exploiting different host plants, diversification could be limited by ecological opportunity rather than the evolution of incompatibility due to isolation. A null or negative relationship between speciation and generalism may therefore lead us to expect a null relationship between molecular evolution and diversification. Although many comparative tests of these hypotheses have been carried out (Jousselin and Elias 2019), they are frequently only conducted on the scale of individual families, tribes, or genera, with the broadest being conducted across the butterfly superfamily (Papilionoidea; Hardy and Otto 2014). Given the importance of host plants in Lepidopteran diversification, it is of interest to determine whether these relationships extend to a broader taxonomic scale, and to speculate on how they may interact with the relationship between diversification and molecular evolution.

* If speciation is driven by host plant shifts, then we expect that when we compare sister clades of butterflies or moths, the one whose species are found on the greatest number of host plants will also have the greatest number of species. We find this is the case for butterfly genera (Papilionoidea, Dataset I) and lepidopteran families (Dataset II: see Table B). However, the number of hosts does not necessarily reflect the number of host shifts: if most species in a clade are generalists then it is possible that the ancestral state for the clade was a generalist that could exploit many species, so that the number of host species on the tips of the lepidopteran clades may not a reflection of the number of instances where a lineage changed hosts within that clade. So we show that the proportion of lepidopteran species in a clade that are known to be generalists (using more than one host plant) is related to species richness, in comparisons of tribes, subfamilies and families across Lepidoptera (Table B).

An alternative explanation for the lack of a clear realtionship between rates of molecular evolution in Lepidoptera is that the association may be obscured or overwhelmed by the influence of other species traits on rates of molecular evolution. Substitution rates can vary with life history characteristics, including generation time (Bromham et al. 1996; Thomas et al. 2010), body size (Barrera-Redondo et al. 2018; Berv and Field 2018; Fontanillas et al. 2007; May et al. 2020), longevity (Galtier et al. 2009; Hua et al. 2015) reproductive rates (Welch and Waxman 2008), as well as species traits such as parasitism (Bromham et al. 2013), sexual competition (Iglesias-Carrasco et al. 2019; Wong 2014), flight (Mitterboeck and Adamowicz 2013) and geographic distribution (Gillman et al. 2010; Lourenço et al. 2013). All of these factors can vary between lepidopteran clades, and since some of these factors also scale with diversification rates there may be complex interactions between factors shaping rates of molecular evolution.

These relationships could potentially confound or mediate a relationship between molecular evolution and diversification. For instance, if the primary driver in Lepidopteran diversification at the family level were adaptive radiation caused by access to major host plant groups, this may induce large differences between the diversity of related clades based on single adaptive events that could serve to mask the more subtle effect of substitutions in promoting speciation due to isolation. Microevolutionary mechanism could also affect our expectations for a relationship between molecular evolution and diversification. For example, if the oscillation hypothesis is true as originally described and speciation is driven by isolation in populations of generalist species, we should expect to see diversification rates related to both substitution rate and rates of generalism (Jousselin and Elias 2019). However, if instead speciation is driven by divergent selection on populations exploiting different host plants, diversification could be limited by ecological opportunity rather than the evolution of incompatibility due to isolation. A null or negative relationship between speciation and generalism may therefore lead us to expect a null relationship between molecular evolution and diversification. Although many comparative tests of these hypotheses have been carried out (Jousselin and Elias 2019), they are frequently only conducted on the scale of individual families, tribes, or genera, with the broadest being conducted across the butterfly superfamily (Papilionoidea; Hardy and Otto 2014). Given the importance of host plants in Lepidopteran diversification, it is of interest to determine whether these relationships extend to a broader taxonomic scale, and to speculate on how they may interact with the relationship between diversification and molecular evolution.

References

Allen AP, Gillooly JF, Savage VM, Brown JH (2006) Kinetic effects of temperature on rates of genetic divergence and speciation. Proceedings of the National Academy of Sciences 103(24):9130-9135

Allio R, Nabholz B, Wanke S, Chomicki G, Pérez-Escobar OA, Cotton AM, Clamens A-L, Kergoat GJ, Sperling FAH, Condamine FL (2021) Genome-wide macroevolutionary signatures of key innovations in butterflies colonizing new host plants. Nature Communications 12(1):354

Allio R, Scornavacca C, Nabholz B, Clamens A-L, Sperling FAH, Condamine FL (2020) Whole Genome Shotgun Phylogenomics Resolves the Pattern and Timing of Swallowtail Butterfly Evolution. Syst Biol 69(1):38-60

Arenas M (2015) Trends in substitution models of molecular evolution. Frontiers in Genetics 6(319)

Arnqvist G, Edvardsson M, Friberg U, Nilsson T (2000) Sexual conflict promotes speciation in insects. Proceedings of the National Academy of Sciences 97(19):10460-10464

Barraclough TG, Harvey PH, Nee S (1996) Rate of *rbc*L gene sequence evolution and species diversification in flowering plants (Angiosperms). Proc. R. Soc. Lond. B 263(1370):589-591

Barraclough TG, Savolainen V (2001) Evolutionary rates and species diversity in flowering plants. Evolution 55:677-683

Barrera-Redondo J, Ramirez-Barahona S, Eguiarte LE (2018) Rates of molecular evolution in tree ferns are associated with body size, environmental temperature, and biological productivity. Evolution 72(5):1050-1062

Beccaloni GW, Scoble M, Kitching IJ, Simonsen T, Robinson G, Pitkin B, Hine A, Lyal C (2003) The Global Lepidoptera Names Index (LepIndex). In,

Beck J, Fiedler K (2008) Adult life spans of butterflies (Lepidoptera: Papilionoidea + Hesperioidea): broadscale contingencies with adult and larval traits in multi-species comparisons. Biological Journal of the Linnean Society 96(1):166-184

Berv JS, Field DJ (2018) Genomic Signature of an Avian Lilliput Effect across the K-Pg Extinction. Systematic Biology 67(1):1-13

Bromham L (2024) Combining Molecular, Macroevolutionary, and Macroecological Perspectives on the Generation of Diversity. Cold Spring Harbor Perspectives in Biology 19:a041453

Bromham L, Cowman PF, Lanfear R (2013) Parasitic plants have increased rates of molecular evolution across all three genomes. BMC Evolutionary Biology 13(1):126

Bromham L, Duchêne S, Hua X, Ritchie AM, Duchêne DA, Ho SYW (2018) Bayesian molecular dating: opening up the black box. Biological Reviews 93(2):1165-1191

Bromham L, Hua X, Lanfear R, Cowman P (2015) Exploring the relationships between mutation rates, life history, genome size, environment and species richness in flowering plants. American Naturalist 185:507

Bromham L, Hua X, Lanfear R, Cowman P (2015) Exploring the relationships between mutation rates, life history, genome size, environment and species richness in flowering plants. American Naturalist 185:507-524

Bromham L, Rambaut A, Harvey PH (1996) Determinants of rate variation in mammalian DNA sequence evolution. Journal of Molecular Evolution 43(6):610-621

Call E, Mayer C, Twort V, Dietz L, Wahlberg N, Espeland M (2021) Museomics: Phylogenomics of the Moth Family Epicopeiidae (Lepidoptera) Using Target Enrichment. Insect Systematics and Diversity 5(2)

Cardillo M (1999) Latitude and rates of diversification in birds and butterflies. Proceedings of the Royal Society of London. Series B: Biological Sciences 266(1425):1221-1225

Carvalho APS, St Laurent RA, Toussaint EFA, Storer C, Dexter KM, Aduse-Poku K, Kawahara AY (2020) Is Sexual Conflict a Driver of Speciation? A Case Study With a Tribe of Brush-footed Butterflies. Systematic Biology 70(3):413-420

Chazot N, Wahlberg N, Freitas AVL, Mitter C, Labandeira C, Sohn J-C, Sahoo RK, Seraphim N, de Jong R, Heikkilä M (2019) Priors and Posteriors in Bayesian Timing of Divergence Analyses: The Age of Butterflies Revisited. Systematic Biology 68(5):797-813

Condamine FL, Rolland J, Höhna S, Sperling FAH, Sanmartín I (2018) Testing the Role of the Red Queen and Court Jester as Drivers of the Macroevolution of Apollo Butterflies. Systematic Biology 67(6):940-964

Condamine FL, Sperling FAH, Wahlberg N, Rasplus J-Y, Kergoat GJ (2012) What causes latitudinal gradients in species diversity? Evolutionary processes and ecological constraints on swallowtail biodiversity. Ecology Letters 15(3):267-277

Cook LM, Saccheri IJ (2013) The peppered moth and industrial melanism: evolution of a natural selection case study. Heredity 110(3):207-212

Davies TJ, Savolainen V, Chase MW, Moat J, Barraclough TG (2004) Environmental energy and evolutionary rates in flowering plants. Proceedings of the Royal Society of London. Series B: Biological Sciences 271(1553):2195-2200

Doorenweerd C, Van Nieukerken EJ, Hoare RJB (2017) Phylogeny, classification and divergence times of pygmy leaf-mining moths (Lepidoptera: Nepticulidae): the earliest lepidopteran radiation on Angiosperms? Systematic Entomology 42(1):267-287

dos Reis M, Donoghue PCJ, Yang Z (2016) Bayesian molecular clock dating of species divergences in the genomics era. Nat Rev Genet 17(2):71-80

Duchene D, Bromham L (2013) Rates of molecular evolution and diversification in plants: chloroplast substitution rates correlate with species richness in the Proteaceae. BMC Evol. Biol. 13:65

Duchêne D, Duchêne S, Ho SYW (2015) Tree imbalance causes a bias in phylogenetic estimation of evolutionary timescales using heterochronous sequences. Molecular Ecology Resources 15(4):785-794

Duchêne DA, Hua X, Bromham L (2017) Phylogenetic estimates of diversification rate are affected by molecular rate variation. Journal of Evolutionary Biology 30(10):1884–1897

Dufresnes C, Brelsford A, Jeffries DL, Mazepa G, Suchan T, Canestrelli D, Nicieza A, Fumagalli L, Dubey S, Martínez-Solano I, Litvinchuk SN, Vences M, Perrin N, Crochet P-A (2021) Mass of genes rather than master genes underlie the genomic architecture of amphibian speciation. Proceedings of the National Academy of Sciences 118(36):e2103963118

Ehrlich AH, Ehrlich PR (1978) Reproductive Strategies in the Butterflies: I. Mating Frequency, Plugging, and Egg Number. Journal of the Kansas Entomological Society 51(4):666-697

Ehrlich PR, Raven PH (1964) BUTTERFLIES AND PLANTS: A STUDY IN COEVOLUTION. Evolution 18(4):586-608

Eo SH, DeWoody JA (2010 ) Evolutionary rates of mitochondrial genomes correspond to diversification rates and to contemporary species richness in birds and reptiles. Proceedings of the Royal Society B - Biological Sciences 277:3587-3592

Espeland M, Breinholt J, Willmott KR, Warren AD, Vila R, Toussaint EFA, Maunsell SC, Aduse-Poku K, Talavera G, Eastwood R, Jarzyna MA, Guralnick R, Lohman DJ, Pierce NE, Kawahara AY (2018) A Comprehensive and Dated Phylogenomic Analysis of Butterflies. Current Biology 28(5):770-778.e775

Ezard THG, Thomas GH, Purvis A (2013) Inclusion of a near-complete fossil record reveals speciation-related molecular evolution. Methods in Ecology and Evolution 4(8):745-753

Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biological Conservation 61:1-10

Fontanillas E, Welch JJ, Thomas JA, Bromham L (2007) The influence of body size and net diversification rate on molecular evolution during the radiation of animal phyla. BMC Ecology and Evolution 7:95

Fordyce JA (2010) Host shifts and evolutionary radiations of butterflies. Proceedings of the Royal Society B: Biological Sciences 277(1701):3735-3743

Galtier N, Blier PU, Nabholz B (2009) Inverse relationship between longevity and evolutionary rate of mitochondrial proteins in mammals and birds. Mitochondrion 9(1):51-57

Gillman LN, Keeling DJ, Gardner RC, Wright SD (2010) Faster evolution of highly conserved DNA in tropical plants. Journal of Evolutionary Biology 23(6):1327-1330

Gillman LN, Wright SD (2014) Species richness and evolutionary speed: the influence of temperature, water and area. Journal of Biogeography 41(1):39-51

Goldie X, Lanfear R, Bromham L (2011) Diversification and the rate of molecular evolution: no evidence of a link in mammals. BMC Evolutionary Biology 11(1):286

Graham CH, Ron SR, Santos JC, Schneider CJ, Moritz C (2004) INTEGRATING PHYLOGENETICS AND ENVIRONMENTAL NICHE MODELS TO EXPLORE SPECIATION MECHANISMS IN DENDROBATID FROGS. Evolution 58(8):1781-1793

Guillory WX, Brown JL (2021) A New Method for Integrating Ecological Niche Modeling with Phylogenetics to Estimate Ancestral Distributions. Systematic Biology

Hamilton CA, St Laurent RA, Dexter K, Kitching IJ, Breinholt JW, Zwick A, Timmermans MJTN, Barber JR, Kawahara AY (2019) Phylogenomics resolves major relationships and reveals significant diversification rate shifts in the evolution of silk moths and relatives. BMC Evolutionary Biology 19(1):182

Hardy NB (2017) Do plant-eating insect lineages pass through phases of host-use generalism during speciation and host switching? Phylogenetic evidence. Evolution 71(8):2100-2109

Hardy NB, Otto SP (2014) Specialization and generalization in the diversification of phytophagous insects: tests of the musical chairs and oscillation hypotheses. Proceedings. Biological sciences 281(1795):20132960

Heibl C, Calenge C (2013) Phyloclim: integrating phylogenetics and climatic niche modelling. R Package version 0.9 4

Hodkinson ID (2005) Terrestrial insects along elevation gradients: species and community responses to altitude. Biological Reviews 80(3):489-513

Hua X, Bromham L (2017) Darwinism for the genomic age: connecting mutation to diversification. Frontiers in Genetics 8

Hua X, Cowman P, Warren D, Bromham L (2015) Longevity is linked to mitochondrial mutation rates in rockfish: a test using Poisson regression. Molecular biology and evolution 32(10):2633-2645

Hugall AF, Lee MS (2007) The likelihood node density effect and consequences for evolutionary studies of molecular rates. Evolution 61(10):2293-2307

Iglesias-Carrasco M, Jennions MD, Ho SYW, Duchene DA (2019) Sexual selection, body mass and molecular evolution interact to predict diversification in birds. Proceedings of the Royal Society B-Biological Sciences 286(1899)

Janssens SB, Couvreur TLP, Mertens A, Dauby G, Dagallier L-PMJ, Vanden Abeele S, Vandelook F, Mascarello M, Beeckman H, Sosef M, Droissart V, van der Bank M, Maurin O, Hawthorne W, Marshall C, Réjou-Méchain M, Beina D, Baya F, Merckx V, Verstraete B, Hardy O (2020) A large-scale species level dated angiosperm phylogeny for evolutionary and ecological analyses. Biodiversity Data Journal 8

Janz N, Nylin S, Wahlberg N (2006) Diversity begets diversity: host expansions and the diversification of plant-feeding insects. BMC Evolutionary Biology 6(1):4

Jiggins CD (2017) The ecology and evolution of Heliconius butterflies. Oxford University Press,

Jousselin E, Elias M (2019) Testing host-plant driven speciation in phytophagous insects: a phylogenetic perspective. arXiv preprint arXiv:1910.09510

Karsholt O, Mutanen M, Lee S, Kaila L (2013) A molecular analysis of the Gelechiidae (Lepidoptera, Gelechioidea) with an interpretative grouping of its taxa. Systematic Entomology 38(2):334-348

Kawahara AY, Mignault AA, Regier JC, Kitching IJ, Mitter C (2009) Phylogeny and Biogeography of Hawkmoths (Lepidoptera: Sphingidae): Evidence from Five Nuclear Genes. PLOS ONE 4(5):e5719

Kawahara AY, Plotkin D, Espeland M, Meusemann K, Toussaint EFA, Donath A, Gimnich F, Frandsen PB, Zwick A, dos Reis M, Barber JR, Peters RS, Liu S, Zhou X, Mayer C, Podsiadlowski L, Storer C, Yack JE, Misof B, Breinholt JW (2019) Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. Proceedings of the National Academy of Sciences 116(45):22657-22663

Kawahara AY, Plotkin D, Ohshima I, Lopez-Vaamonde C, Houlihan PR, Breinholt JW, Kawakita A, Xiao L, Regier JC, Davis DR, Kumata T, Sohn J-C, De Prins J, Mitter C (2017) A molecular phylogeny and revised higher-level classification for the leaf-mining moth family Gracillariidae and its implications for larval host-use evolution. Systematic Entomology 42(1):60-81

Kumar S, Hedges SB (2016) Advances in time estimation methods for molecular data. Molecular Biology and Evolution 33(4):863-869

Lancaster LT (2010) Molecular evolutionary rates predict both extinction and speciation in temperate angiosperm lineages. BMC Evol Biol 10(1):162

Landis M, Edwards EJ, Donoghue MJ (2020) Modeling Phylogenetic Biome Shifts on a Planet with a Past. Systematic Biology 70(1):86-107

Lanfear R, Ho SYW, Love D, Bromham L (2010) Mutation rate influences diversification rate in birds. Proceedings of the National Academy of Sciences 107(47):20423-20428

Lanfear R, Welch JJ, Bromham L (2010) Watching the clock: Studying variation in rates of molecular evolution between species. Trends in Ecology & Evolution 25(9):495-503

Laubach T, von Haeseler A, Lercher MJ (2012) TreeSnatcher plus: capturing phylogenetic trees from images. BMC Bioinformatics 13(1):110

Lee G-E, Han T, Park H, Qi M, Li H (2021) A phylogeny of the subfamily Thiotrichinae (Lepidoptera: Gelechiidae) with a revision of the generic classification based on molecular and morphological analyses. Systematic Entomology 46(2):357-379

Léger T, Mally R, Neinhuis C, Nuss M (2021) Refining the phylogeny of Crambidae with complete sampling of subfamilies (Lepidoptera, Pyraloidea). Zoologica Scripta 50(1):84-99

Lourenço JM, Glémin S, Chiari Y, Galtier N (2013) The determinants of the molecular substitution process in turtles. J Evol Biol 26(1):38-50

Mackintosh A, Laetsch DR, Hayward A, Charlesworth B, Waterfall M, Vila R, Lohse K (2019) The determinants of genetic diversity in butterflies. Nature Communications 10(1):3466

Matute DR, Butler IA, Turissini DA, Coyne JA (2010) A test of the snowball theory for the rate of evolution of hybrid incompatibilities. Science 329(5998):1518-1521

May JA, Feng Z, Orton MG, Adamowicz SJ (2020) The effects of ecological traits on the rate of molecular evolution in ray-finned fishes: a multivariable approach. J Mol Evol 88(8-9):689-702

Menéndez R, González-Megías A, Collingham Y, Fox R, Roy DB, Ohlemüller R, Thomas CD (2007) DIRECT AND INDIRECT EFFECTS OF CLIMATE AND HABITAT FACTORS ON BUTTERFLY DIVERSITY. Ecology 88(3):605-611

Menken SBJ, Boomsma JJ, Van Nieukerken EJ (2010) LARGE-SCALE EVOLUTIONARY PATTERNS OF HOST PLANT ASSOCIATIONS IN THE LEPIDOPTERA. Evolution 64(4):1098-1119

Milla L, van Nieukerken EJ, Vijverberg R, Doorenweerd C, Wilcox SA, Halsey M, Young DA, Jones TM, Kallies A, Hilton DJ (2018) A preliminary molecular phylogeny of shield-bearer moths (Lepidoptera: Adeloidea: Heliozelidae) highlights rich undescribed diversity. Molecular Phylogenetics and Evolution 120:129-143

Mitter C, Davis DR, Cummings MP (2017) Phylogeny and Evolution of Lepidoptera. Annual Review of Entomology 62(1):265-283

Mitterboeck TF, Adamowicz SJ (2013) Flight loss linked to faster molecular evolution in insects. Proceedings of the Royal Society B: Biological Sciences 280(1767):20131128

Moyle LC, Nakazato T (2010) Hybrid incompatibility “snowballs” between Solanum species. Science 329(5998):1521-1523

Murillo-Ramos L, Brehm G, Sihvonen P, Hausmann A, Holm S, Reza Ghanavi H, Õunap E, Truuverk A, Staude H, Friedrich E, Tammaru T, Wahlberg N (2019) A comprehensive molecular phylogeny of Geometridae (Lepidoptera) with a focus on enigmatic small subfamilies. PeerJ 7:e7386

Niehuis O, Yen SH, Naumann CM, Misof B (2006) Higher phylogeny of zygaenid moths (Insecta: Lepidoptera) inferred from nuclear and mitochondrial sequence data and the evolution of larval cuticular cavities for chemical defence. Mol Phylogenet Evol 39(3):812-829

Nohara C, Hiyama A, Taira W, Otaki JM (2017) Robustness and Radiation Resistance of the Pale Grass Blue Butterfly from Radioactively Contaminated Areas: A Possible Case of Adaptive Evolution. Journal of Heredity 109(2):188-198

Novotny V, Drozd P, Miller SE, Kulfan M, Janda M, Basset Y, Weiblen GD (2006) Why Are There So Many Species of Herbivorous Insects in Tropical Rainforests? Science 313(5790):1115-1118

Orr HA, Turelli M (2001) The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities. Evolution 55(6):1085-1094

Pagel M, Venditti C, Meade A (2006) Large punctuational contribution of speciation to evolutionary divergence at the molecular level. Science 314(5796):119

Pellissier L, Kostikova A, Litsios G, Salamin N, Alvarez N (2017) High Rate of Protein Coding Sequence Evolution and Species Diversification in the Lycaenids. Frontiers in Ecology and Evolution 5(90)

Pellissier L, Litsios G, Guisan A, Alvarez N (2012) Molecular substitution rate increases in myrmecophilous lycaenid butterflies (Lepidoptera). Zoologica Scripta 41(6):651-658

Presgraves DC (2010) Speciation genetics: search for the missing snowball. Current Biology 20(24):R1073-R1074

Regier JC, Brown JW, Mitter C, Baixeras J, Cho S, Cummings MP, Zwick A (2012) A Molecular Phylogeny for the Leaf-Roller Moths (Lepidoptera: Tortricidae) and Its Implications for Classification and Life History Evolution. PLOS ONE 7(4):e35574

Regier JC, Mitter C, Davis DR, Harrison TL, Sohn J-C, Cummings MP, Zwick A, Mitter KT (2015) A molecular phylogeny and revised classification for the oldest ditrysian moth lineages (Lepidoptera: Tineoidea), with implications for ancestral feeding habits of the mega-diverse Ditrysia. Systematic Entomology 40(2):409-432

Regier JC, Mitter C, Kristensen NP, Davis DR, Van Nieukerken EJ, Rota J, Simonsen TJ, Mitter KT, Kawahara AY, Yen S-H, Cummings MP, Zwick A (2015) A molecular phylogeny for the oldest (nonditrysian) lineages of extant Lepidoptera, with implications for classification, comparative morphology and life-history evolution. Systematic Entomology 40(4):671-704

Regier JC, Mitter C, Mitter K, Cummings MP, Bazinet AL, Hallwachs W, Janzen DH, Zwick A (2017) Further progress on the phylogeny of Noctuoidea (Insecta: Lepidoptera) using an expanded gene sample. Systematic Entomology 42(1):82-93

Regier JC, Mitter C, Solis MA, Hayden JE, Landry B, Nuss M, Simonsen TJ, Yen S-H, Zwick A, Cummings MP (2012) A molecular phylogeny for the pyraloid moths (Lepidoptera: Pyraloidea) and its implications for higher-level classification. Systematic Entomology 37(4):635-656

Ritchie AM, Hua X, Cardillo M, Yaxley KJ, Dinnage R, Bromham L (2020) Phylogenetic diversity metrics from molecular phylogenies: modelling expected degree of error under realistic rate variation. Diversity and Distributions n/a(n/a)

Robinson GS, Ackery PR, Kitching IJ, Beccaloni GW, Hernández LM (2010) HOSTS - A database of the world's Lepidopteran hostplants. In: Natural History Museum, London

Ronquist F, Sanmartín I (2011) Phylogenetic Methods in Biogeography. Annual Review of Ecology, Evolution, and Systematics 42(1):441-464

Rota J, Wahlberg N (2012) Exploration of data partitioning in an eight-gene data set: phylogeny of metalmark moths (Lepidoptera, Choreutidae). Zoologica Scripta 41(5):536-546

Schär S, Eastwood R, Arnaldi KG, Talavera G, Kaliszewska ZA, Boyle JH, Espeland M, Nash DR, Vila R, Pierce NE (2018) Ecological specialization is associated with genetic structure in the ant-associated butterfly family Lycaenidae. Proceedings of the Royal Society B: Biological Sciences 285(1886):20181158

Scopece G, Musacchio A, Widmer A, Cozzolino S (2007) Patterns of reproductive isolation in Mediterranean deceptive orchids. Evolution: International Journal of Organic Evolution 61(11):2623-2642

Scriber JM (2010) Integrating ancient patterns and current dynamics of insect–plant interactions: Taxonomic and geographic variation in herbivore specialization. Insect Science 17(6):471-507

Seraphim N, Kaminski LA, Devries PJ, Penz C, Callaghan C, Wahlberg N, Silva-Brandão KL, Freitas AVL (2018) Molecular phylogeny and higher systematics of the metalmark butterflies (Lepidoptera: Riodinidae). Systematic Entomology 43(2):407-425

Sohn J-C, Regier JC, Mitter C, Davis D, Landry J-F, Zwick A, Cummings MP (2013) A Molecular Phylogeny for Yponomeutoidea (Insecta, Lepidoptera, Ditrysia) and Its Implications for Classification, Biogeography and the Evolution of Host Plant Use. PLOS ONE 8(1):e55066

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics (Oxford, England) 30(9):1312-1313

Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A (2018) Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evolution 4(1)

Thomas JA, Welch JJ, Lanfear R, Bromham L (2010) A generation time effect on the rate of molecular evolution in invertebrates. Molecular Biology and Evolution 27(5):1173-1180

van der Linden CFH, WallisDeVries MF, Simon S (2021) Great chemistry between us: The link between plant chemical defenses and butterfly evolution. Ecology and Evolution 11(13):8595-8613

Venditti C, Meade A, Pagel M (2006) Detecting the node-density artifact in phylogeny reconstruction. Syst Biol 55(4):637-643

Venditti C, Pagel M (2010) Speciation as an active force in promoting genetic evolution. Trends in Ecology & Evolution 25(1):14-20

Venditti C, Pagel M (2014) Plenty of room for punctuational change. Trends Ecol Evol 29(2):71-72

Wahlberg N, Rota J, Braby MF, Pierce NE, Wheat CW (2014) Revised systematics and higher classification of pierid butterflies (Lepidoptera: Pieridae) based on molecular data. Zoologica Scripta 43(6):641-650

Wang H, Holloway JD, Janz N, Braga MP, Wahlberg N, Wang M, Nylin S (2017) Polyphagy and diversification in tussock moths: Support for the oscillation hypothesis from extreme generalists. Ecology and Evolution 7(19):7975-7986

Wang Q-Y, Li H-H (2020) Phylogeny of the superfamily Gelechioidea (Lepidoptera: Obtectomera), with an exploratory application on geometric morphometrics. Zoologica Scripta 49(3):307-328

Webster AJ, Payne RJH, Pagel M (2003) Molecular phylogenies link rates of evolution and speciation. Science 301:478

WEINGARTNER E, WAHLBERG N, NYLIN S (2006) Dynamics of host plant use and species diversity in Polygonia butterflies (Nymphalidae). Journal of Evolutionary Biology 19(2):483-491

Welch JJ, Waxman D (2008) Calculating independent contrasts for the comparative study of substitution rates. J Theor Biol 251(4):667-678

Wiemers M, Chazot N, Wheat CW, Schweiger O, Wahlberg N (2020) A complete time-calibrated multi-gene phylogeny of the European butterflies. ZooKeys 938:97-124

Wong A (2014) Covariance between testes size and substitution rates in primates. Molecular Biology and Evolution 31(6):1432-1436

Yang ZH (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24(8):1586-1591

Zahiri R, Holloway JD, Kitching IJ, Lafontaine JD, Mutanen M, Wahlberg N (2012) Molecular phylogenetics of Erebidae (Lepidoptera, Noctuoidea). Systematic Entomology 37(1):102-124

**Table A**

Datasets used in this study

I and II are used to test the relationship between molecular and diversification rates.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Taxonomic level** | **Taxon** | **Contrasts** | **Phylogeny** |
| I | Genera | Papilionoidea | 18 | Chazot et al (2019) |
| II | Families | Lepidoptera | 66 | Kawahara et al (2019) |
| II | Whole order | Non-Ditrysian (basal) clades |  | Regier et al. (2015) |
|  | Superfamily | Bombycoidea |  | Hamilton et al. (2019) |
|  |  | Gelechioidea |  | Wang and Li (2020) |
|  |  | Noctuoidea |  | Regier et al. (2017) |
|  |  | Pyraloidea |  | Regier et al. (2012) |
|  |  | Tineoidea |  | Regier et al. (2015) |
|  |  | Yponomeutoidea |  | Sohn et al. (2013) |
|  | Family | Choreutidae |  | Rota and Wahlberg (2012) |
|  |  | Crambidae |  | Léger et al. (2021) |
|  |  | Epicopeiidae |  | Call et al. (2021) |
|  |  | Erebidae |  | Zahiri et al. (2012) |
|  |  | Gelechiidae |  | Karsholt et al. (2013) |
|  |  | Geometridae |  | Murillo-Ramos et al. (2019) |
|  |  | Gracilariidae |  | Kawahara et al. (2017) |
|  |  | Heliozelidae |  | Milla et al. (2018) |
|  |  | Nepticulidae |  | Doorenweerd et al. (2017) |
|  |  | Pieridae |  | Wahlberg et al. (2014) |
|  |  | Riodinidae |  | Seraphim et al. (2018) |
|  |  | Sphingidae |  | Kawahara et al. (2009) |
|  |  | Thiotrichidae |  | Lee et al. (2021) |
|  |  | Tortricidae |  | Regier et al. (2012) |
|  |  | Zygaenidae |  | Niehuis et al. (2006) |

Table V: Variables analysed in this study

|  |  |  |
| --- | --- | --- |
| **Variable** | **Name** | **Description** |
| Synonymous substitution rate | dS | How calculated for each sister lineage |
| Nonsynonymous substitution rate | dN |  |
|  | dN/dS |  |
| Total substituions |  | Total substitutions using the program ‘baseml’ with a general time-reversible (GTR) substitution model (Yang 2007) |
| Clade size |  | For each family, subfamily, tribe or genus represented in our sets of sister clades, we count all unique binomials with the ‘Current Status’ field listed as ‘Valid Name’ in the Global Lepidoptera Names Index (LepIndex; Beccaloni et al. 2003).. |
| Total Host Species |  | Host species breadth is represented by the number of named angiosperm species exploited by at least one member of the taxon from the HOSTS index, which lists known host plant species exploited by lepidopteran species (Robinson et al. 2010). |
| Mean Hosts/Species |  | mean number of hosts per species (Mean Hosts/Species) for each sister clade, by dividing the number of host plant species (Total Host Species) by the number of species in the clade (?? meaning Clade size?}} |
| Host Jumps |  | Im not sure we have this at the moment but we could work it out. Either I can do this manually per comparison (would need sister pairs with identified host families) or if there is a simple way to automate it we could do that. |
| Host Families |  |  |
| Generalists | Prop.Generalists | The proportion of generalists for each Lepidopteran is calculated as the ratio of the number of listed lepidopteran species [is this clade size?] that exploit more than one host plant species, to the total number of listed species (Prop.Generalists). |
| Host Phylodiversity | Hosts\_pd\_std | Faith’s phylogenetic diversity index (FPD; Faith 1992) of the associated host plant species. This is calculated as the sum of molecular branch lengths connecting all associated host plant species for each clade using the recent maximum likelihood angiosperm megaphylogeny of Janssens et al. (2020)   * Is this corrected for number of host speices?   Does std mean standardized and if so how? |
|  |  |  |

**Table B: Correlation between lepidopteran diversity and their host diversity.**

Informative table legend defining all of the columns.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset | Taxon | Host diversity | Coefficient | Std. Error | *t* | P-value |
| I | Papilionoidea | Total Host Families | 0.515 | 0.290 | 1.775 | 0.084 |
|  |  | **Total Host Species** | 0.455 | 0.132 | 3.443 | **0.001** |
|  |  | Mean Hosts / Species | 0.246 | 0.196 | 1.256 | 0.217 |
|  |  | Prop.Generalists | 0.189 | 0.298 | 0.632 | 0.532 |
|  |  | **Host phylodiversity** | 0.692 | 0.181 | 3.818 | **0.002** |
| II | Lepidoptera | **Total Host Families** | 1.230 | 0.138 | 8.886 | **< 0.001** |
|  |  | **Total Host Species** | 0.802 | 0.102 | 7.864 | **< 0.001** |
|  |  | **Mean Hosts / Species** | 3.288 | 0.730 | 4.505 | **0.001** |
|  |  | **Prop. generalists** | 4.986 | 1.569 | 3.177 | **0.010** |
|  |  | Host phylodiversity | -0.875 | 1.491 | -0.587 | 0.567 |
| III | Lepidoptera | **Total Host Families** | 0.858 | 0.137 | 6.263 | **< 0.001** |
|  |  | Total Host Species | 0.006 | 0.090 | 0.072 | 0.943 |
|  |  | **Mean Hosts / Species** | 0.871 | 0.415 | 2.100 | **0.043** |
|  |  | Prop. generalists | 0.585 | 1.010 | 0.532 | 0.598 |
|  |  | **Host phylodiversity** | 0.444 | 0.087 | 5.120 | **< 0.001** |

**Table C: Correlation between lepidopteran diversity and rates of molecular evolution**

Informative table legend defining all of the columns.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset | Taxon | Substitutions | Coefficient | Std. Error | t | P-value |
| I | Papilionoidea | Total | -0.033 | 0.641 | -0.052 | 0.959 |
|  |  | dS | 0.401 | 0.644 | 0.621 | 0.537 |
|  |  | dN | -0.318 | 0.203 | -1.571 | 0.121 |
| II | Lepidoptera | Total | -0.585 | 5.062 | -0.116 | 0.909 |
|  |  | dS | -0.668 | 3.594 | -0.186 | 0.855 |
|  |  | dN | -0.146 | 0.233 | -0.624 | 0.541 |

**Table D: Linear models of log host family richness against log substitution rates**

Informative figure legend defining all of the columns.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset | Taxon | Substitutions | Estimate | Std. Error | t value | Pr(>|t|) |
| I | Papilionoidea | Total | 5.486 | 5.354 | 1.025 | 0.330 |
|  |  | dS | 5.516 | 3.996 | 1.380 | 0.198 |
|  |  | dN | 0.962 | 5.150 | 0.187 | 0.856 |
| II | Lepidoptera | Total | 0.735 | 0.581 | 1.267 | 0.213 |
|  |  | dS | 0.451 | 0.555 | 0.812 | 0.422 |
|  |  | dN | **0.626** | **0.255** | **2.458** | **0.019** |

**Table. Multiple regressions of clade size against host diversity and molecular rates**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Phylogeny | Substitutions | Coefficient | Std. Error | *t* | P-value |
| Lepidoptera – Families | dS | -1.128 | 2.163 | -0.521 | 0.615 |
|  | **Host Families** | **1.263** | **0.157** | **8.051** | **> 0.001** |
| Papilionoidea - Genera | dS | 0.277 | 0.897 | 0.309 | 0.759 |
|  | **Host Species** | **0.449** | **0.134** | **3.359** | **0.002** |

**Table. Number of sister pair comparisons available for each dataset used in the regression analysis.** Numbers of pairs differ when analysing the three different substitution types because different numbers of pairs are removed by the Welch filter. The number of pairs is reduced when analysing host plant data because not all taxa have entries in the HOSTS database.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable analysed | Lepidoptera-Families | Lepidoptera-MajorLineages | Papilionoidea-Genera |
| Total subst. | 16 | - | 66 |
| dS | 18 | - | 66 |
| dN | 17 | - | 66 |
| HOSTS data | 9 | 31 | 16 |