NEW GENES FOR NEW BIOLOGY IN ANIMAL EVOLUTION

Asia Elizabeth Hoile
Student ID: 1092320
Keble College
Department of Biology

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

University of Oxford
Michaelmas 2025

Table of Contents

Declaration and Statement of Authorship

The work presented here was conducted by me (**Asia E. Hoile**) under the supervision of Professor Peter W. H. Holland and Dr Peter O. Mulhair in the Department of Biology at the University of Oxford. Any contributions of other individuals are explicitly stated and material referenced from other sources has been clearly indicated and its origins cited.

The work presented in Chapter 3 has been published as a first author original research article in the Journal *BMC Genomics* (2025). The work presented in Chapter 4 is under review at *Insect Molecular Biology* as a research article with myself as first author. All work on these manuscripts was completed during my time as a DPhil student at the University of Oxford. Detailed statements of authorship contributions are included at the beginning of the relevant results chapters; where published manuscripts have multiple authors the specific contributions of each author are stated.

I hereby confirm that this thesis has not been submitted for any other qualification or degree at this university or any other institution.

Asia E. Hoile

Keble College

9th October 2025

Acknowledgements

There are a number of people who I would like to thank, not just for supporting me throughout my DPhil, but also for inspiring me along this path in the first place.

I would firstly like to thank my primary supervisor, Professor Peter Holland for his guidance, support and helpful discussions throughout the past four years. Thank you to Professor Sebastian Shimeld for his support and to Dr Peter Mulhair for having endless patience whilst teaching me how to conduct bioinformatic analyses, always being on hand to help with coding issues and for all your encouragement when I didn't feel confident in my ability.

Thank you to all members of the Holland, Shimeld and Aboobaker labs for input on my research, but most importantly for always being on hand for a chat over lunch. Thank you to the BBSRC and DTC for their support in my research, especially to my DTC supervisors, Professor Esther Becker and Professor Gail Preston, in addition to Dr Berta Verd for their advice and support throughout the course of my DPhil.

Thank you to everyone who inspired me to pursue this path. I would particularly like to acknowledge my A level teachers at Ysgol Gyfun Ystalyfera for inspiring me to work hard and to strive to keep developing my work; Mr Richard Morgan, Mr David Lewis, Mrs Julie Jones, Mrs Angharad Lloyd, Mrs Emma Sandberg, Mr Danny Williams, Mrs Bethan Murphy, Dr Anita Rees, Mr Ashley Davies, Mrs Bethan Powell and Mrs Delyth Spurway – *Dysgu gorau, Dysgu byw*.

Thank you also to Dr Neil Gostling at the University of Southampton for inspiring my love and passion for evolutionary biology, demonstrating the true meaning of a good teacher and for believing I could get this far.

Although it may be argued to be more of a detriment than a support to my studies, I would like to acknowledge Oxford University Rifle Club as the reason I have spent far more time on the range than I have in the lab, but for somehow helping me maintain my sanity throughout my studies – Vs and 3s! I would particularly like to thank Adam Chidlow and Michael Horrell for their friendship and for the occasional word of advice on DPhil work too!

Finally, I would like to acknowledge my friends and family for all of their support in everything I do. Thank you to Kitty Barnard and Ethan Ross for ensuring that studying biology was nothing short of a comedy show, yet somehow always managing to come out on top! Thank you to Rebecca Cook for always offering a listening ear, a giggle and supporting me through all the ups and downs we've shared. Thank you to Hugo Malim for always being there for me and knowing how to make me laugh.

Thank you to my uncle, Leighton Hendra for all the nature walks in my younger days which likely inspired my love of biology in the first place. Thank you to Herbert the cat for his unwavering loyalty and many hours spent supervising my writing. His steady presence, quiet encouragement, and diligent oversight have undoubtedly enriched this thesis far beyond what I could have achieved alone.

And lastly, but most certainly not least, Mum and Dad. Words cannot express how truly blessed I am to have you as my parents. Thank you for being my biggest cheerleaders, my rock and for always believing in me even when I did not believe in myself. I dedicate this thesis to you both, because I am certain that I could not done this without you.

In loving memory of my grandmothers, Iris Hoile and Nora Hendra. Though you are no longer here, your love and wisdom live on in all that I do. I wish you were here to share this moment.

List of Figures

Chapter 1: General Introduction

Chapter 2: Identification of new genes arising at Metazoan nodes of interest

Chapter 3: Gene novelty and gene family expansion in the early evolution of Lepidoptera

Chapter 4: Gene expression in the reduced first thoracic legs of a Nymphalid butterfly

Chapter 5: Functional enrichment and genetic novelty in Bilaterian-specific tissues

Chapter 6: General Discussion and conclusions

List of Tables

List of Publications

Work for the following chapters has resulted in publications:

Chapter 3:

Hoile, A.E., Holland, P.W.H. and Mulhair, P.O. 2025. Gene novelty and gene family expansion in the early evolution of Lepidoptera. *BMC Genomics* 26(1), p. 161. Available at: https://doi.org/10.1186/s12864-025-11338-x.

Chapter 4:

Work submitted for this chapter is currently under review at *Insect Molecular Biology*.

Darwin Tree of Life Genome Notes

Over the course of my DPhil, I contributed to the following 'Genome Note' publication for species sequenced as part of the Darwin Tree of Life Project (https://www.darwintreeoflife.org/).

Boyes, D. and Hoile, A. 2023. The genome sequence of the Brick, Agrochola circellaris (Hufnagel, 1766). *Wellcome Open Research* 8, p. 44. doi: 10.12688/WELLCOMEOPENRES.18894.1.

Thesis Abstract

Gene novelty is one of the mechanisms by which phenotypic novelty may arise and has contributed to the animal diversity observed today. This thesis aims to advance our understanding of evolutionary innovation arising from gene novelty through the integration of multiple levels of analysis, incorporating genomic novelty, expression and morphological function.

Chapter 2 explores and discusses the rationale for the development of a pipeline to identify new genes, and how this requires modification when working with nodes spanning different evolutionary timescales. Chapter 3 demonstrates that gene duplication and divergence are the primary drivers of novel gene emergence and describes two examples of extensive duplication within Lepidoptera, one of which was identified as a putative HGT named 'Propellin'. Chapter 4 integrates these molecular insights with phenotypic evolution, exploring how changes in gene expression may contribute to the morphological reduction of T1 legs in Nymphalid butterflies, and how this may have resulted in the evolution of a chemosensory function in these appendages. Chapter 5 extends this perspective across Bilateria, specifically investigating novel genes which demonstrate tissue-specific enrichment in gut, nerve cord / nervous system and muscle tissue.

Together, this research highlights the coordination between gene sequence innovation, regulatory dynamics and morphological diversification, offering a comprehensive view of how genomic and developmental processes may contribute to evolutionary novelty. Through the combination of these multi-level analyses, this research offers both conceptual and methodological advances, establishing a model for studying the molecular mechanisms underlying trait evolution across diverse animal lineages, in the context of novel gene evolution.

Chapter 1: General Introduction

Genetic diversity is one of the factors that allow the variety of species to exist as observed today, and new genes have been major contributors to the origin of adaptive evolutionary novelties (Kaessmann 2010). Despite this being such a fundamental part of the diversity of life, comparatively little is understood about how evolutionary new genes arise, where they come from, whether they confer an adaptive benefit, and how this all relates to phenotypic change. In this research, I explore the methods by which new genes arise, their functions and how this relates to animal evolution. There are several major transitions in the tree of life at which gene evolution can be studied, however this thesis investigates two: Lepidoptera and Bilateria.

1.1 The explosion of animal diversity

The origin and evolution of life on Earth have been influenced by a combination of factors, both environmental and biological (Kaessmann 2010). As the Earth has changed throughout geological time, life has reflected this and resulted in increased species variety and complexity as a response to a changing environment and either more favourable or more challenging conditions for survival (Payne et al. 2020).

1.1.1 Biological evolution and geological change go hand in hand

Biological evolution and geological change go hand in hand in understanding the diversification of life on Earth (Herrera-Alsina et al. 2021). Studying fossils and their corresponding layers of rock formation allows for the understanding of Earth at different periods in historical time. It is also clear that geological events have contributed to the evolution and extinction of species (Herrera-Alsina et al. 2021; Foster et al. 2023). Examples of this include moving tectonic plates, changes in atmospheric composition and mass extinction events (Grant et al. 2017).

The Earth's crust and upper mantle comprise seven major tectonic plates and smaller minor plates which are able to move independently (Stern and Gerya 2023). It is this movement which resulted in the rearrangement of continents from the supercontinent Pangaea and geological features such as volcanoes, mountains and earthquakes. This resulted in the emergence of new environments and the isolation of species, thereby facilitating the evolution of different species on different continents (Willen 2003; Stern 2016).

This movement also gave rise to the Earth's early atmosphere. This was largely dominated by volcanic gases such as carbon dioxide, water vapour, ammonia and methane, with little to no oxygen (Zahnle et al. 2010). The oldest evidence of life on Earth are stromatolites which are found in rocks approximately 3.5 billion years old (Noffke et al. 2013), however it is possible that life began as early as 4.1 billion years ago (Bell et al. 2015). Stromatolites are round sedimentary rock formations that were made by cyanobacteria which photosynthesised, using sunlight and carbon dioxide from the atmosphere to chemically produce

carbohydrates as a food source. As a waste product of this reaction, oxygen was formed and resulted in an increased proportion of oxygen in the Earth's atmosphere (Blaustein 2016).

This increased availability of oxygen in the Earth's atmosphere stimulated the evolution of new species. An example of this is some bacteria acquiring the ability to use oxygen to digest carbohydrates to release energy and produce carbon dioxide as a waste product (respiration) (Peter Jurtshuk 1996). Around 2.3 billion years ago, the atmospheric oxygen level reached ~2% and resulted in the formation of a layer of ozone in the stratosphere. This allowed for the protection of life from solar radiation and hence facilitated the expansion of life to what we know today (Takahashi and Ohnishi 2004).

A sudden disappearance of fossil records often represents large-scale extinction of biological species and occur as a result of abrupt changes in the Earth. There have been at least five major extinction events in the last 500 million years of the Earth, the most well-known occurred 65 million years ago near the end of the Cretaceous and resulted in the extinction of the non-avian dinosaurs (Brusatte et al. 2015). Extinction events can be beneficial for the long-term evolution of life on earth, resulting in increased biodiversity and the evolution of new adaptations as it can provide opportunities for new species to emerge and diversify by clearing out existing species and freeing up ecological niches (Jablonski 2001). This is observed following the extinction of non-avian dinosaurs, resulting in the diversification of mammals from small rodent-like forms to the variety observed today (Upham et al. 2019).

1.1.2 Evolution and diversification of life on Earth

The Ediacaran period occurred between 635 and 542 million years ago and marked the first widespread appearance of complex multicellular organisms (Retallack 2013). This followed the end of the snowball Earth glaciation event which resulted in the Ediacaran Evolutionary Radiation (EER) and the emergence of diverse, soft-bodied organisms (Landing et al. 2018). The snowball Earth was a period when the entire Earth was likely covered in a sheet of ice, followed by rapid greenhouse warming resulting from a buildup of carbon dioxide over multiple cycles. It is also hypothesised that the end of the snowball Earth provided nutrient-rich conditions, resulting in a favourable environment for diversification of Ediacaran biota (Agić et al. 2024). The fossils identified from the Ediacaran period are widely considered evolutionary predecessors of the Cambrian explosion of marine animal phyla (Erwin et al. 2011).

The first widely accepted bilaterian fossil - an animal possessing bilateral symmetry - is *Ikaria warioota* which is roughly 555 million years ago and dates to the Ediacaran period. *Ikaria warioota* is a small (2-7mm length) worm-like organism which likely had burrowing habits. It represents a link between the Ediacaran and the subsequent explosion of bilaterian diversity observed during the Cambrian period (Evans et al. 2020).

Around 540 million years ago, during the early Cambrian period, the diversity of life expanded through a “three-phased explosion” of animal body plans, episodic biomineralization, pulsed change of generic diversity, body size variation and progressive increase of ecosystem complexity which is collectively known as the Cambrian Explosion

(Zhang and Shu 2021). Although this phenomenon does not have a single isolated cause, the Cambrian Explosion could not have taken place without molecular evolution (Zhang and Shu 2021). The Cambrian Explosion was first noticed as a result of a palaeontological phenomenon where there was a sudden emergence of the skeletal remains of diverse animals in the lowest known fossiliferous rocks of the Cambrian, which was first noted by William Buckland (Conway Morris 2000). During this period there was a rapid increase in the diversity of bilaterian animals in particular. While the Ediacaran period marks the emergence of early bilaterians, the Cambrian explosion (around 541 million years ago) saw a rapid diversification of many major animal groups, including the development of most bilaterian body plans (Zhang and Shu 2021).

Although it is considered as a more controversial hypothesis, it has also been proposed that the diversification which occurred during the Cambrian may have taken place over a greater period of evolutionary time and was therefore more gradual. It could therefore be suggested that this was a “slow burn” evolution event rather than a “short fuse” as illustrated in Figure 1.1 (Conway Morris 2000). Despite this, most evidence points towards a rapid diversification i.e. a Cambrian ‘explosion’ (Zhang and Shu 2021).

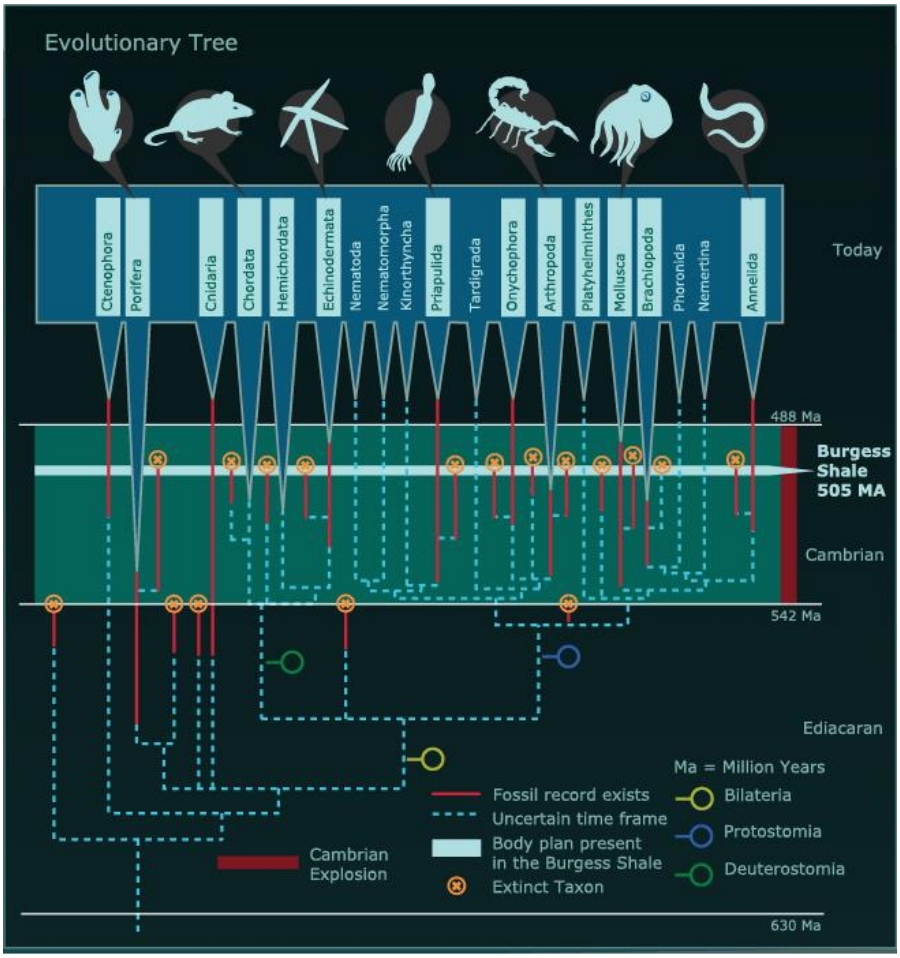


Figure 1.1 – One (controversial) proposal for the dating of Cambrian and Ediacaran fossils demonstrating a number of old "ghost" lineages. This evolutionary tree implies a "slow burn" as opposed to a "short fuse" when considering the Cambrian diversification event, which would have resulted in gradual diversification of organisms rather than the rapid divergence as described in the Cambrian Explosion hypothesis (Royal Ontario Museum 2025).

1.2 Mechanisms of gene novelty

Diversification and adaptation are dependent on genetic change; however identifying specific genes which result in phenotypic change can be difficult. Changes in existing genes cannot explain all adaptive evolution and therefore it is likely that this can be at least partly explained by gene number variation, gene duplications, and gene novelty in driving evolution and adaptation. Whatever the mode of origin, novel genes likely reflect novel biology as they will encode proteins with potentially distinct activity or function absent in

the outgroup taxa. Although further mechanisms for phenotypic novelty will be discussed, this thesis aims to focus on gene novelty in this context.

1.2.1 Gene duplication

Gene duplication occurs when a gene-containing region of DNA is altered such that a descendent individual inherits more than one copy (Ohno 1970; Birchler and Yang 2022). This could occur through recombination between misaligned homologous chromosomes causing tandem duplication, retro-transposition from the insertion of an RNA transcript, whole genome duplication resulting in ohnologues, replication slippage or accidental capture by a selfish genetic element that undergoes copying (Hastings Philip and Rosenberg Susan 1998; Suzuki David T et al. 2000; Zhang 2003; Eickbush and Jamburuthugoda 2008; Choudhuri 2014; Qiao et al. 2018; Velicky et al. 2018). Occasionally one or more of the duplicate copies may undergo radical sequence change, although it may retain a functional domain (Holland et al. 2017). In this case, the gene duplicate would have recognisable distant homologues but, by the definition of novelty explored through this thesis, it has become a novel gene.

1.2.2 Gene duplication outcomes

Gene duplication usually results in one of three outcomes (Figure 1.2). Firstly, gene duplication may result in sub-functionalisation (Birchler and Yang 2022). This is a neutral mutation process and means that no new adaptations are formed and instead the function of the original gene is distributed between the two duplicated paralogs. As a result, neither gene can be lost but equally neither can possess novel functionality, however this does not

mean that the function cannot be elaborated through this process (Ohno 1970; Conrad and Antonarakis 2007; Sémon and Wolfe 2008). Secondly, gene duplication may result in neofunctionalization. This is an adaptive mutation process where one of the gene copies undergo mutation to develop a new function that did not exist within the ancestral gene, while the second copy retains its original function present within the ancestral gene (Ohno 1970; Rastogi and Liberles 2005; Conrad and Antonarakis 2007; Kleinjan et al. 2008). Thirdly, gene duplication may result in degeneration or gene loss. Neutral, beneficial and deleterious mutations can be lost or spread through the population through genetic drift, while selection may fix beneficial mutations. When the mutation is detrimental, it is unlikely to be preserved (Lee and Lupski 2006).

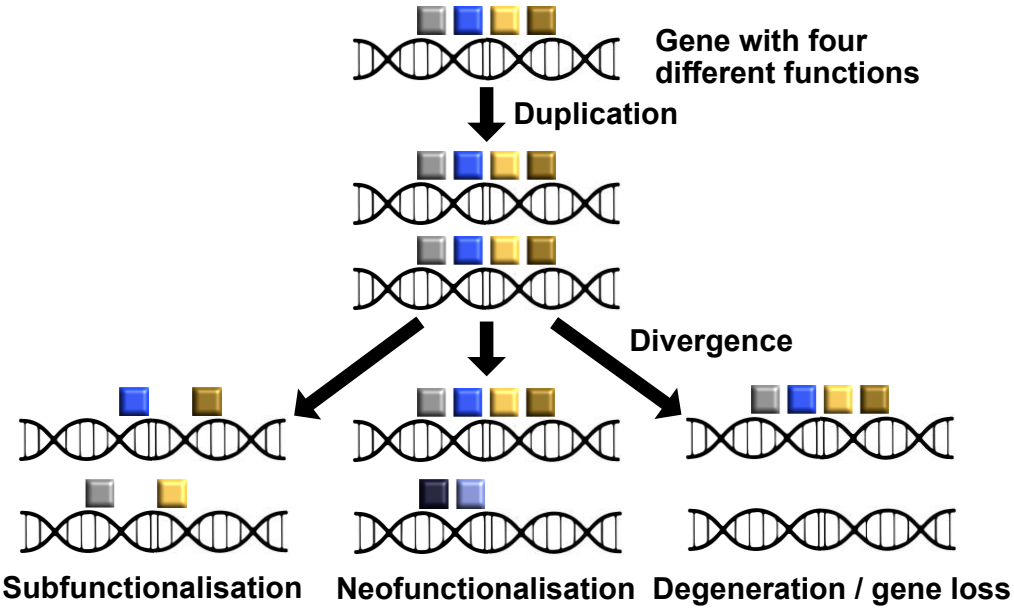


Figure 1.2 – Evolutionary fates of duplicate genes can result in one of three outcomes: ***sub-functionalisation*** *- a neutral mutation process of constructive neutral evolution and means that no new adaptations are formed and instead the function of the original gene is distributed between the two duplication paralogs. As a result the descendant duplicate genes split the original ancestral functions between them meaning they are only able to perform some of the original functions, while the ancestral gene was capable of performing all of the functions. Therefore neither gene can be lost but equally neither can poses novel functionality.* ***Neofunctionalization*** *- where one of the gene copies undergo mutation to develop a new function that did not exist within the ancestral gene, while the second copy retains its original function present within the ancestral gene.* ***Degeneration / gene loss*** *- this is where one copy is lost and is likely due to detrimental or lack of useful function.*

1.2.3 Transposable elements and gene shuffling

Transposable elements (TEs) are DNA sequences which can replicate themselves within a genome, often translocating to new genomic locations (McClintock 1950). For example, DNA transposons including terminal inverted repeat TEs and rolling-circle TEs (*Helitrons*) can capture coding DNA from neighbouring genes and copy them to elsewhere in the genome. This process can duplicate genes in the genome, which could then diverge in sequence, or they can contribute to generate novel genes by copying exons which become fused to other genes (Bourque et al. 2018; Catoni et al. 2019; Ma et al. 2023). Gene shuffling is the process by which existing genes and genetic elements within the genome are recombined to create novel genetic combinations which can result in the generation of novel genes or novel functions (Meyer et al. 2013).

1.2.4 Horizontal gene transfer

Horizontal gene transfer (HGT) is probably an underappreciated mode by which novel genes may arise in animals. It could be argued that HGT-derived genes are not 'novel genes' as they also exist in the donor genome, or a progenitor gene does. In the context of functional adaptation, however, they certainly give rise to novelty because an HGT event suddenly introduces new protein-coding potential into an evolutionary lineage (Eyres et al. 2015; Wybouw et al. 2016; Nowell et al. 2024). Furthermore, many HGT-derived genes seem to undergo radical sequence change, presumably during adaptation to the new cellular and physiological conditions (Barraclough 2015). For a horizontally transferred gene to be incorporated into a eukaryotic genome, it must be integrated into the germ line to be

transferred to offspring (Husnik and McCutcheon 2018). This could be achieved by exposure to ectopic DNA within the reproductive tissues, or during early developmental stages which include precursor cell lineages for the germ cells (Huang 2013). It is notable that some intracellular bacteria have a close association with the germ line, for example *Wolbachia* in arthropods (Kondo et al. 2002).

1.2.5 Genes arising 'de novo'

Genes may also arise '*de novo*', meaning they emerge from ancestrally non-protein coding DNA sequence which may or may not be transcribed (Andersson et al. 2015). There is evidence that some new genes have evolved from non-coding sequences in insects (Wu and Zhang 2013). Examples include the *Drosophila* genes *Goddard* and *Saturn* required for male fertility (Gubala et al. 2017).

1.2.6 Whole Genome duplication (Polyploidy)

Whole genome duplication (WGD) is the process by which a whole genome of a cell of an organism is doubled. This means the doubling of coding regions, non-coding regions and regulatory sequences, and results in the presence of an additional set of chromosomes (Moriyama and Koshiba-Takeuchi 2018). This duplication results from errors in cell division and is often considered to be an evolutionary 'dead end' (Van De Peer et al. 2017). Although the number of duplications which occur and remain fixed are small, evidence suggests that at least some evolutionarily successful lineages are derived from lineages which experienced whole genome duplications. Major ancient WGD events occurred during the origins of

vertebrates (Dehal and Boore 2005; Meyer and Van De Peer 2005) which were rapidly fixed at population level and resulted in the evolution of novel genes over 500-600 million years. Most WGD events have occurred in angiosperm lineages which were rapidly fixed, followed by evolution of new genes over 150-200 million years (Soltis et al. 2009). In fungi there is only evidence of one WGD (Scannell et al. 2007). In some instances therefore, WGD may be advantageous to evolutionary success.

Further examples include rotifers, barnacles, land snails and slugs. Bdelloid rotifers have likely undergone tetraploidisation followed by extensive gene loss (degenerate tetraploidy) which may facilitate survival in habitats which often involve desiccation and rehydration (Mark Welch et al. 2008). Barnacles have undergone a single WGD event which is hypothesised to have assisted with adaptation to the stressful intertidal environment, such as enrichment in the temperature stress pathway (Au et al. 2025). Similarly, slugs and land snails have also undergone WGD which is hypothesised to have facilitated the development of novel traits which have allowed their survival on land e.g. duplication of the HOX cluster (McHale et al. 2025). In contrast, although a single research article implies that WGD has taken place within the insects (Li et al. 2018), this has since been refuted (Nakatani and McLysaght 2019; Roelofs et al. 2020).

1.3 Further mechanisms for phenotypic novelty

Although this thesis will focus on gene novelty, it is important to acknowledge that this is not the sole mechanism for phenotypic novelty. Further mechanisms include regulatory mechanisms, developmental mechanisms, and ecological influences. All of the above can also result in species evolution and diversity, and mechanisms are often interconnected, with multiple mechanisms acting together to result in novelty within an organism (Bush et al. 2017). This section aims to outline these mechanisms and give examples of where they act in the animal kingdom.

1.3.1 Regulatory Mechanisms

In this context, regulation refers to changes in location, timing and how genes are activated to generate new traits without alteration in underlying coding sequence (Peter & Davidson, 2015). Examples of different regulatory mechanisms include cis/trans-regulation, promoters and enhancers, microRNAs, piRNAs, epigenetic changes, DNA methylation and histone modification. Regulatory changes may also be driven by co-option and repurposing existing genes (Rebeiz et al. 2011).

Cis and trans-regulatory elements maintain gene expression (Signor and Nuzhdin 2018). Cis-regulatory elements are DNA sequences in close proximity to a gene, such as promoters or enhancers, and influence the expression of a gene. Trans-regulatory elements such as transcription factors act on many genes (Hansen et al. 2024). Due to the direct, local impact on single genes, cis-regulatory changes are typically linked to species-specific traits and

morphological differences while trans-regulatory changes can globally influence multiple genes and regulatory networks despite typically being more conserved (Hansen et al. 2024).

Cis-regulatory changes are often linked to species-specific traits and morphological differences due to their direct and local impact on single genes, such as genetic variation on a gene's promoter or enhancer region, which has been linked to the emergence of novel traits. An example of this is the reduction of pelvic fins in freshwater sticklebacks resulting from deletion of a *Pitx1* enhancer (Chan et al. 2009). Trans-regulatory changes contribute to phenotypic novelty through globally influencing multiple genes and regulatory networks which can lead to new developmental patterns or traits within an organism. Mutations in trans-acting factors may be located anywhere across the genome and can affect downstream target genes, leading to new gene expression patterns over time which can result in phenotypic novelty (Signor and Nuzhdin 2018). An example of this is in *Drosophila* wing patterning where trans regulatory elements bind to cis-regulatory elements related to pigmentation genes, therefore resulting in the observation of new wing pigmentation patterns (Hanly et al. 2019).

Promoters and enhancers are cis-regulatory DNA elements which share common features such as transcription factor binding and histone modification, however differ slightly in their function and location in which they act (Kim and Shiekhattar 2015). Promoters are DNA sequences which are necessary for the initiation of transcription and are located at the start of a gene. Mutations or changes in promoters can impact timing, location and quantity of transcription which can lead to modification of existing traits or the emergence of new ones (Ruiz-Narváez 2013). Enhancers differ from this as they are located far away from the

relevant gene and therefore regulate and activate transcription from a distance (Agrawal et al. 2018). They act more like switches and bind to transcription factors to increase or decrease gene expression and are typically less constrained than enhancers, which can result in new gene expression patterns (Medina-Rivera et al. 2018).

MicroRNAs and piRNAs are examples of small noncoding RNAs (sncRNAs) which are involved in regulatory mechanisms giving rise to phenotypic novelty. MicroRNAs are small, single-stranded molecules which can block protein synthesis or cause the breakdown of messenger RNA (mRNA). This results in destabilising of the target mRNA and the resulting gene silencing of the target (O'Brien et al. 2018). MicroRNAs are therefore crucial for regulation of gene expression and allow for rapid cellular response to environmental change, developmental cues of disease states, allowing evolutionary adaptation and phenotypic plasticity (Huang et al. 2024). PiRNAs form complexes with the piwi proteins of the Argonaute family and as a result cause the silencing of transposable elements during the critical time of germ-line development, therefore maintaining genome integrity (Wu and Zamore 2021). Despite this, evolutionary gain and subsequent diversification of novel piRNA clusters can result in modification of gene expression patterns or the silencing of new genomic elements, therefore contributing to novel phenotypes (Grimson et al. 2008; Huang et al. 2024).

Phenotypic novelty may also be obtained without any alteration of the DNA sequence via epigenetic changes. Modifications such as DNA methylation or histone modification may lead to new expression patterns, while maintaining the original DNA sequence (Duncan et al. 2022). DNA methylation involves the addition of a methyl group to a cytosine. This chemical modification influences transcriptional activity by blocking the binding of transcription

factors and hence affecting transcriptional activity (Hanson and Liebl 2022). An example of this is the variation in seasonal colour morphs in Siberian hamsters, resulting from variation in extent of DNA methylation at the *DIO3* gene in the hypothalamus as a response to shifts in daylength (Husby 2020). This influences a lightening or darkening of coat colour in response to seasonal change (Duncan et al. 2022).

Histone modifications include acetylation and methylation, which alter chromatin structure. This directly impacts accessibility of genes for transcription, therefore regulating gene expression (Shirvaliloo 2022). This is observed in *D. melanogaster* where the H3K27R mutation prevents methylation of the histone H3 at Lysine 27, leading to an alteration in the expression of key genes such as *Ubx*, *Abd-B*, *Scr* and *en* (Lavarone et al. 2019; Sankar et al. 2022). This results in homeotic transformations, i.e. where differentiated cells develop into the wrong structure or body part and gives rise to new phenotypes (Yung et al. 2015). Epigenetic changes often occur as a result of environmental change and are partly responsible for phenotypic plasticity (Angers et al. 2020).

Co-option, often referred to as gene recruitment is the method by which pre-existing genetic networks or regulatory sequences are recruited into a new regulatory system. Here, they serve a novel function within an existing biological processes or pathway, which is often in a different tissue or developmental context (Shirai et al. 2012; Liang 2024). This process often results in the evolution of lineage specific traits (Xia et al. 2021). This can occur as a result of environmental change such as the resulting response to acute stress, leading to evolutionary change in populations of organisms (Eyck et al. 2019; Love and Wagner 2022). An example of

this is when *Drosophila melanogaster* is exposed to brief heat shocks at a young age, which can result in increased longevity and thermotolerance in adults (Le Bourg et al. 2001).

It is also possible for single genes to be repurposed for new functions, for example crystallins. The ancestral proteins, which include small heat-shock proteins and βγ-crystallins have gained new structural and functional properties such as high solubility, longevity and stability which have allowed crystallins to become the primary structural components of the eye lens (Slingsby et al. 2013). It is likely that this was a gradual process, occurring through stepwise modification to gene promoters which were already expressed at low levels in the lens (Cvekl et al. 2004).

1.3.2 Developmental mechanisms

Developmental mechanisms, in combination with genetic change, give rise to phenotypic diversity (Nijhout 2025). This includes heterochrony, heterotopy, heterometry, heterotypy, modularity and integration. The first four are mechanisms by which developmental processes can be altered during evolution to result in changes in an organism's form (Zhang et al. 2014). Modularity involves the organisation of an organism into semi-independent units and integration describes how each unit is constrained and influences others (Gilbert 2000; Hall and Hanken 2023).

Heterochrony can be described as the changes in timing of developmental events such as paedomorphosis or peramorphosis. Paedomorphosis is the retention of juvenile traits in an adult descendant, while peramorphosis is delayed maturation, resulting in the extension or

overdevelopment beyond the ancestral adult form (McNamara 2012). An example of paedomorphosis is the axolotl which is the juvenile form of the Mexican Salamander *Ambystoma mexicanum*. It retains juvenile features such as external gills into its sexually mature adult stage, allowing it to continue an aquatic lifestyle (De Groef et al. 2018). Other salamanders undergo thyroid-hormone regulated metamorphosis to transition to a terrestrial life, however a disruption in this process allows the axolotl to exploit aquatic habitats (Crowner et al. 2019). Peramorphosis can be observed in the Irish Elk and it's resulting large antlers as a result of extended development during their growth and maturation period (Moen et al. 1999; Mooi 2009).

Heterotopy occurs when there are spatial changes in where genes are expressed during development, leading to new developmental pathways and resulting in the alteration of a type or quality of a trait, often giving rise to new structures (A Dictionary of Biology. 2019a). Heterotopy can be observed in the muscles of tetrapod forelimbs which can originate from different segments of the early embryo (somites) (West-Eberhard 2003).

Heterometry can be defined as the change in quantity or level of gene product, and this directly influences the scale or "amount" of a feature (Yanai et al. 2011). Heterometry can be observed in the land snail *Capaea nemoralis* which can vary in the number of black bands observed on its shell (Johansen et al. 2023). Heterometry is responsible for the number of bands observed on an individual's shell in this species as it influences the quantity of expression of the supergene controlling shell traits (Gonzalez et al. 2019).

Heterotypy is responsible for change in the type of gene products through regulatory genes and coding mutations and can lead to the evolution of a novel structure or body part in an organism (A Dictionary of Biology. 2019). One of the best examples of heterotypy is in the Antennapedia mutant of *Drosophila melanogaster* where legs replace the antennae and in Bithorax mutants where wings grow in the location of the halteres (Webster and Zelditch 2005). This occurs due to a mutation in the *Antp* gene and BX-C gene cluster respectively, and results in the transformation of one segment into another (Bender et al. 1983; Emerald and Cohen 2004).

The interaction between modularity and integration can influence phenotypic novelty through the creation of developmental pathways which can facilitate rapid evolution or generate diverse and novel structures (Zhang et al. 2014). This can be observed in the evolution of flatfish cranial asymmetry which has resulted from coordinated changes in the skull and the production of a novel and highly integrated trait which has allowed adaptation to a new environment (Black and Berendzen 2020). Developmental studies have revealed that this change takes place through a combination of thyroid hormone expression and changes in swimming behaviour (Schreiber 2006). This innovation allowed for the domination of the benthic aquatic habitat and the modularisation of his trait allowed for the evolution of skull shape without impacting the remainder of the body plan (Evans et al. 2021).

1.3.3 Ecological and Evolutionary influences

Ecological and evolutionary influences, in combination with genetic change may also be considered as mechanisms for phenotypic novelty and include hybridisation, phenotypic plasticity, niche construction and environmental feedback, and sexual selection (Sommer 2020). Ecological factors such as resource availability, climate and biotic interactions shape these phenotypic novelties, while evolutionary processes such as natural selection refine these traits over generations (Riley et al. 2023).

Hybridisation occurs when two species or populations of the same species crossbreed which results in the mixing of genomes and can produce new trait combinations which are not observed in either parent species or population (Stelkens et al. 2009). If these traits are beneficial to the organism, they may become fixed over time (Atsumi et al. 2021). One of the most well-studied examples of hybridisation is observed in the Galapagos Finches, where the hybrid cross of two species has resulted in transgressive segregation on beak shape and size. This novelty allowed these finches to feed on previously inaccessible sources of food and contributed to their ecological success and formation of a new species (Grant and Grant 2008).

Phenotypic plasticity refers to the ability of an organism to vary its phenotype in response to environmental conditions, without permanently fixing that trait and without modifying its genotype (Sommer et al. 2017). The nematode *Pristionchus pacificus* exhibits feeding plasticity in response to environmental factors such as starvation and crowding (Bento et al. 2010; Sommer et al. 2017). These nematodes are capable of expressing two different mouth

phenotypes depending on whether they are under bacterial or predatory conditions to maximise survival (Bento et al. 2010). This contributes to evolution through allowing the selection of the most optimal phenotype, in response to environmental cues.

Niche construction is the process where organisms alter environmental states, directly modifying the conditions that they and surrounding organisms experience and therefore influencing the source of natural selection that acts upon them, which is known as environmental feedback (Lala 2024). Beavers constructing a dam is an example of both of these processes. The construction of a dam alters water flow and therefore results in the formation of new aquatic habitats (Laland and O'Brien 2011). This in turn influences biodiversity and selection pressures for the beaver and other organisms residing in these habitats (Wright et al. 2002).

Finally, sexual selection occurs when new or unique traits evolve directly as a result of providing a reproductive advantage in mate selection or competition, such as ornamentation or specialised calls and do not occur due to environmental pressures or influence alone (Fox et al. 2019). This is observed in male fiddler crabs in the family Ocypodidae, which have evolved one enlarged "major" claw which has an ornamental function to attract females and is used as a weapon when fighting against rival males (Swanson et al. 2013). This feature has a high metabolic cost and a lower endurance capacity, however is still maintained as it is favoured by females of the species and therefore increases the likelihood of mating (Allen and Levinton 2007).

1.4 Bilateria

Bilateria are a clade of animals which represent a major transition in the tree of life, representing phenotypic novelty on a wide scale (Holland 1998). The Cambrian Explosion involved a significant increase in ecosystem complexity coupled with rapid diversification and the development of new behaviours such as burrowing and swimming (Holland 2015). Bilateria are believed to have originated during the Ediacaran period, roughly 555 million years ago (Evans et al. 2020). They are one of the five main lineages of animals and are characterised by bilateral symmetry during embryonic development which is often maintained into adulthood (Namigai et al. 2014). Genetic novelty is hypothesised to have been a major driver in the phenotypic novelty observed within this lineage (Holland 2015).

1.4.1 Physiology

Bilateria are characterised by having bilateral symmetry as an embryo, which often continues into adulthood. They have a front and rear end (anterior-posterior axis) and top and bottom (dorsal-ventral axis) (Brusca and Shuster 2016). Bilateria also have a complete digestive tract with a separate mouth and anus, and the presence of cephalisation resulting from a concentration of sense organs such as eyes, ears and a brain at the anterior end of the body (Minelli 2009). Additionally, the Bilateria possess muscle blocks which allow crawling and burrowing in sediment, in addition to other forms of movement. In combination with the above features, this allowed the bilateria to explore this changing world in three dimensions (Holland 2015) (Figure 1.3).

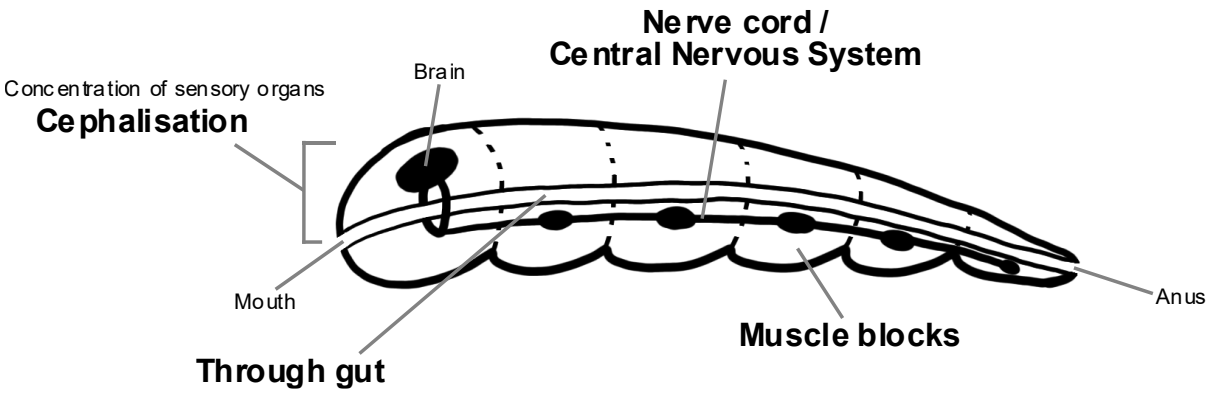


Figure 1.3 – Basic model of a bilaterian animal, demonstrating the presence of a complete digestive tract with a separate mouth and anus, cephalisation, muscle blocks and bilateral symmetry

One such theory is that developmental genetic changes may have contributed to the rapid emergence of bilaterian diversity of life observed during the Cambrian explosion. It has been observed that ANTP-class homeobox genes increased in number in the bilaterian stem lineage and earlier, resulting in the generation of a large array of ANTP class genes (Holland 2015). This includes three distinct gene clusters: NK, Hox and ParaHox. Evidence suggests that NK genes are mainly involved in patterning the bilaterian mesoderm, Hox genes were responsible for coding position along the central nervous system, and it is likely that ParaHox genes originally specified the mouth, midgut and anus of the newly evolved through-gut (Holland 2015). For this reason, it is implied that the diversification of the ANTP class genes mentioned above were involved in the patterning systems which resulted in animal bodies capable of high-energy directed locomotion, and therefore contributing to the Cambrian Explosion (Holland 2015). However it should be noted that Homeobox genes may play only one small part in the genetic change observed during this period. During the Cambrian explosion, phenotypic evolution was ~4 times faster and molecular evolution was ~5.5 times faster in comparison to all subsequent parts of the Phanerozoic eon (Lee et al. 2013a). The Phanerozoic eon is the period during which abundant animal and plant life has existed, covering 538.8 million years ago to the present (Cohen et al. 2013) .

1.4.2 Phylogeny

There are multiple hypotheses on the placement of outgroups at the base of the Bilateria (Li et al. 2021; Juravel et al. 2023) (Figure 1.4). The Phylum Xenacoelomorpha is an example of this and consists of two sister groups: Xenoturbella and Acoelomorpha. Xenacoelomorpha are marine worms which are characterised by bilateral symmetry, but an absence of other common bilaterian features such as an anus, nephridia, and a circulatory system (Hejnol and Pang 2016). There are two main debated locations for the Xenacoelomorpha: within the deuterostomes and sister to Ambulacraria, or as a sister to the Nephrozoa (protostomes and deuterostomes) (Kapli et al. 2021; Mulhair et al. 2022; Worsaae et al. 2023).

There are also debates surrounding the placement of non-bilaterian outgroups such as the aquatic invertebrates Ctenophora, Cnidaria, Placozoa and Porifera (Li et al. 2021). Ctenophora are commonly known as comb jellies and are the largest animals to swim with the aid of cilia and exhibit bi-radial symmetry (Malakhov and Gantsevich 2022). The Cnidaria are characterised by an un-centralised nervous system, which consists of a nerve net that is distributed throughout a gelatinous body and the presence of cnidocytes for capture of prey. They reproduce both sexually and asexually and this phylum includes jellyfish, hydroids, sea anemones and corals (Schierwater and DeSalle 2021). Placozoa are often described as “The simplest animals on earth”. They move along surfaces using a ciliary motion and consume food by engulfment (Pennisi 2021). Porifera are commonly known as sea sponges and lack true tissues or organs. They have a porous body structure and a network of internal channels, used for water circulation, filter feeding and respiration (Wörheide et al. 2012).

Although many variations of placement of these bilaterian outgroups exist (Figure 1.4), currently the most supported placement is Cnidaria as a sister to Placozoa, these as a sister clade to Bilateria, Ctenophora as sister to this and then Porifera. Another common iteration is to transpose Ctenophora and Porifera, making Ctenophora the outmost outgroup (Jékely et al. 2015). For the purpose of this thesis, in Chapter 5 I only focus on the placement of the Xenacoelomorpha in the context of Bilaterian phylogeny. As the aim of this chapter is exploration of gene novelty in the Bilateria, locations of these outgroups would not impact gene novelty on the Bilaterian node, however it is still important to acknowledge that there are many possible iterations of the Bilaterian tree and its outgroups.

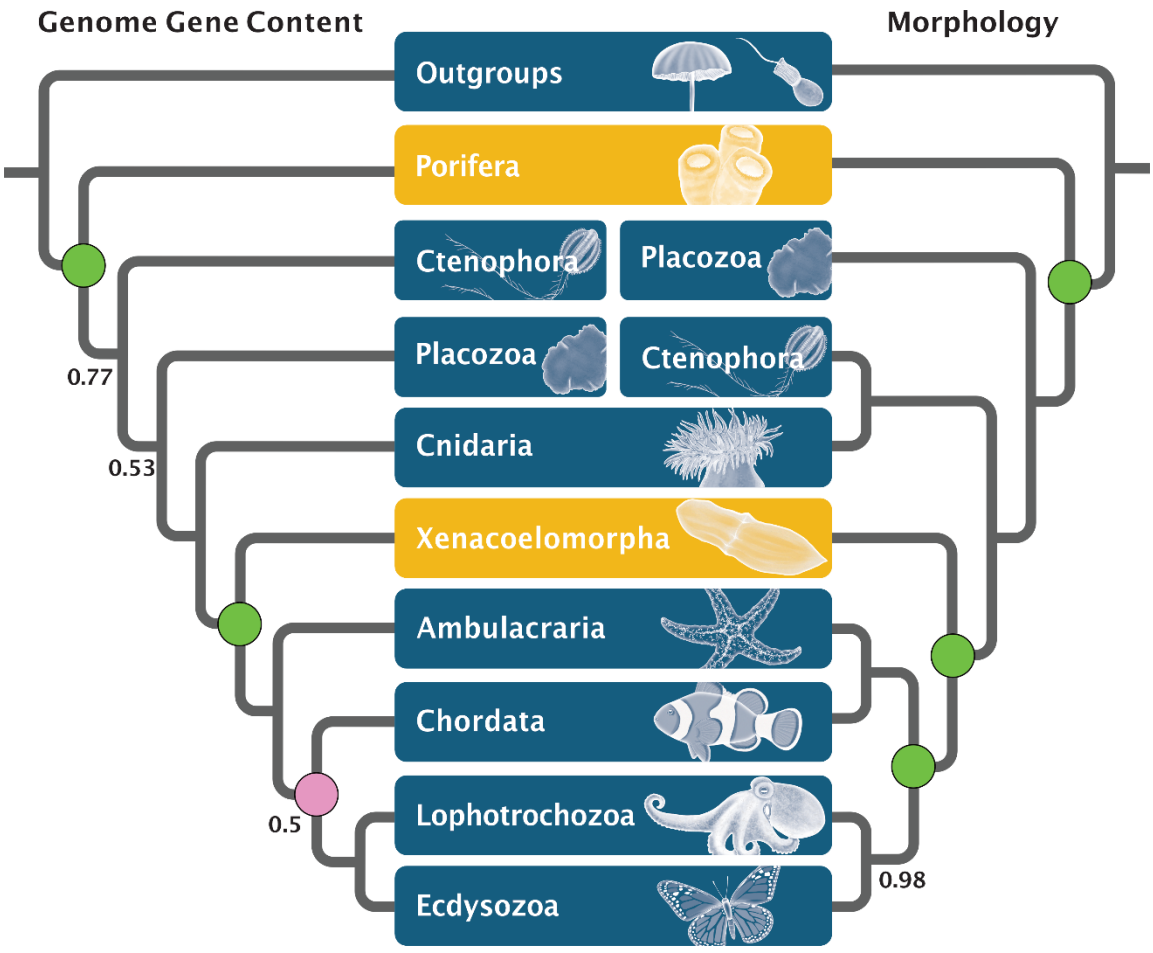


Figure 1.4 - Representation of the variety of the placements of bilaterian outgroups on a phylogenetic tree (Juravel et al. 2023).

1.5 The Lepidoptera

The Order Lepidoptera (moths and butterflies) are another example of a major transition in the tree of life. They form one of the largest insect radiations, comprising up to 10% of all described animal species and are believed to have diversified in sync with angiosperms which has facilitated their unique ecological roles as herbivores and pollinators (Kawahara et al. 2019; Wright et al. 2024). They also serve as important indicators of environmental change (Martay et al. 2016).

1.5.1 Physiology

Typically, Lepidoptera have bodies which are divided into a head, thorax and abdomen, have wings and three pairs of legs attached to the thorax (Figure 1.5A). They also undergo complete metamorphosis, meaning they go from egg to larva, pupa and finally adult (Powell 2009).

The word ‘Lepidoptera’ translates to ‘scale wing’ and refers to the small, modified scales which overlap to comprise the wings of all organisms in this order (Navarro et al. 2017). In addition to this, the majority of Lepidoptera lay terrestrial eggs, exhibit enhanced olfaction through the development sophisticated olfactory structures such as Androconia scales in males which attract females for mating, undergo metamorphosis and demonstrate phytophagy in larvae, allowing them to be well adapted to the variety of niches they occupy and allowing them to inhabit almost all terrestrial ecosystems (Li et al. 2024; Baral et al. 2025).

Larvae of the earliest lepidopteran lineages likely fed on nonvascular land plants as 'internal feeders' while adult forms had mandibulate chewing mouthparts which were likely used to feed on pollen (Young and Montgomery 2020). This is observed in basal extant lepidopteran species such as Micropterigidae (Erenler and Gillman 2010) (Figure 1.5B). The majority of Lepidoptera have a tube-like proboscis which allows for nectar feeding and may have resulted in the ability to colonise new niches through feeding on new plant food sources (Reinwald et al. 2022) (Figure 1.5A).

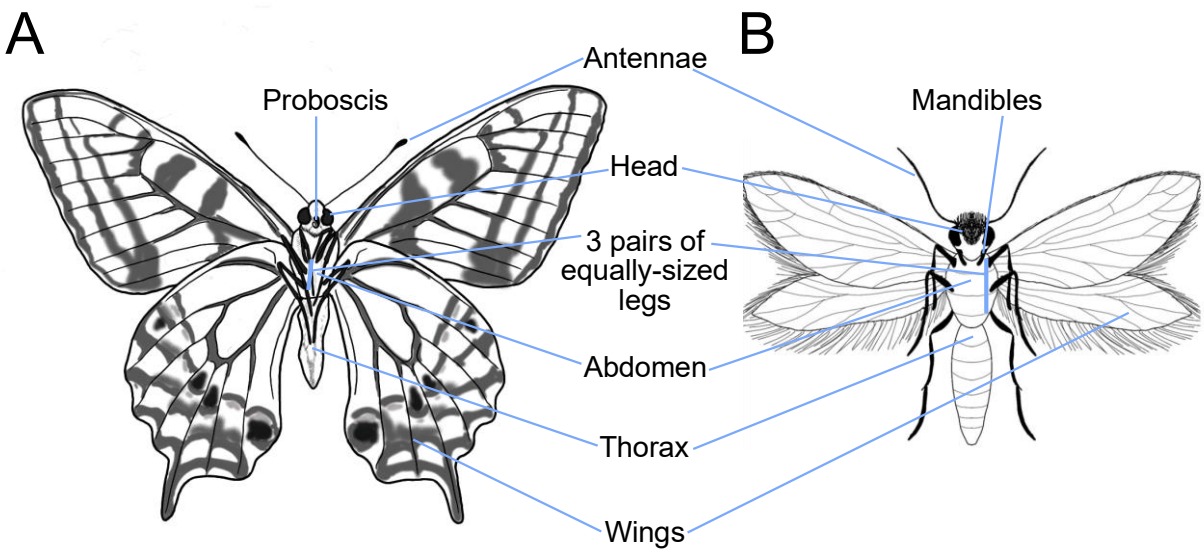


*Figure 1.5 – Labelled diagrams of lepidopteran anatomy. **A** Example of a "typical" lepidopteran e.g. Papilio machaon, demonstrating a proboscis for nectar feeding. **B** Micropterix aruncella, representing more basal moths with mandibles used for pollen feeding, and representing what the ancestral lepidopteran may have looked like.*

1.5.2 Phylogeny

Lepidopteran phylogeny is complex and undergoing constant revision (Kawahara et al. 2019). Although the monophyly of Lepidoptera is well established, research to resolve relationships within the Ditrysia (a major clade within the Lepidoptera) is still ongoing (Mitter et al. 2017). The earliest ancestors of the Lepidoptera crown group appeared approximately 300mya during the late Carboniferous period (Kawahara et al. 2019), however the oldest known lepidopteran fossil *Archaeolepis mane* is only believed to be 190 million years old, while the earliest evidence of scale wings dates to 201 million years old (Van Eldijk et al. 2018; Wang et al. 2022).

1.6 The lepidopteran family Nymphalidae

The lepidopteran family Nymphalidae are also known as the brush footed butterflies due to their unusual, reduced legs in the first thoracic segment (T1) (Hamm and Fordyce 2015). This family is the largest butterfly family, comprising over 6,000 species which occupy a variety of niches across the world (Wahlberg et al. 2009). The majority of nymphalid species have brightly coloured forewings and dull hindwings which are believed to mimic dead leaves to produce a cryptic effect for camouflage (Otak 2020).

1.6.1 Morphological differences

The most notable defining characteristic of the Nymphalidae is their reduced T1 legs (Figure 1.6), a trait which is shared with one other taxonomic family within the Lepidoptera, Riodinidae, which is believed to have evolved in parallel (Wolfe et al. 2011). What sets them apart is the reduced forelegs are observed in both sexes of nymphalid butterflies, but present only in male riodinids. This reduction in the prothoracic (T1) legs means that nymphalids walk on the legs of the second (T2) and third (T3) thoracic segments only. This loss of walking function in the T1 legs suggests that they may have acquired a new function which is unique to the Nymphalidae (Wolfe et al. 2011).

In addition to their unique leg morphology, the nymphalid butterflies have additional defining characteristics such as 12 veins on the forewing and tricarinate (three-ridged) antennae in the adult form (Berry et al. 2008). Nymphalid larvae typically have hairs or spines to deter predators (Mullen 2002).

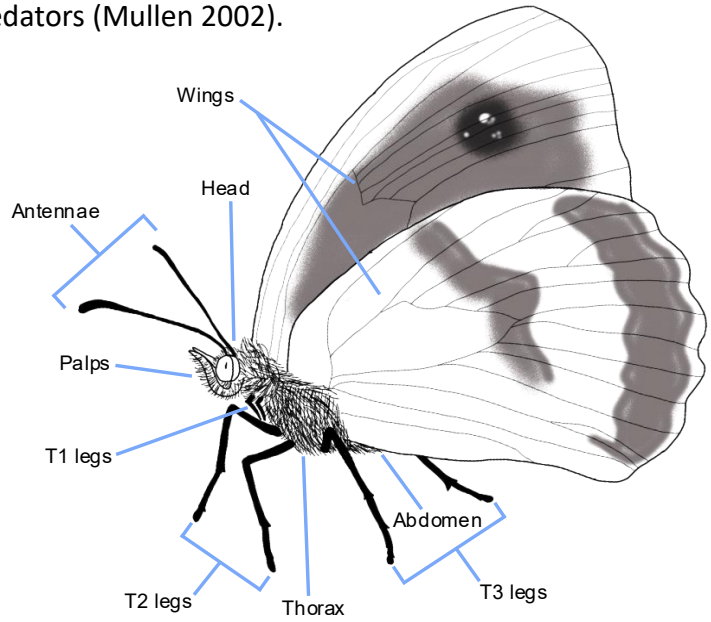


Figure 1.6 - Labelled diagram of Maniola jurtina as an example of a Nymphalid butterfly, highlighting the presence of reduced T1 legs which distinguish Nymphalids from other lepidopteran families.

1.6.2 Hypotheses for the use of brush feet

Although true function of the T1 legs in the Nymphalids remains unclear, it is speculated that T1 legs may have a role in sensory function. It is possible that labial palps and arthropod legs have evolved from a common ancestral appendage type as they are considered to be serially homologous structures (Panganiban et al. 1997; Hughes and Kaufman 2002). We know that lepidopteran labial palps are involved in chemoreception and tactile sensing which guides feeding behaviour (Myers 1969; Diiak et al. 2023). Furthermore, T1 legs are covered with small, sensory hairs called sensillae which are known to be implicated in host plant recognition and oviposition (Calvert and Hanson 1983; Baur et al. 1998). It has also been demonstrated that T1 legs react differently to T2 and T3 legs when exposed to stimuli such as sugar solution, a behaviour that is not observed in non-Nymphalid lepidopterans (Fox 1966). This evidence points toward a potential similarity of T1 legs to known sensory functions such as the palps, or a novel sensory function which may be unique to the Nymphalidae.

Female nymphalids have been known to ‘drum’ on host plant leaves as a taste-test. They rapidly brush their T1 legs against the desired leaf’s surface to release plant sap (Briscoe et al. 2013; Thiele et al. 2016). This is used as a means to sample the leaf’s chemical properties through the sensillae and assess its suitability for egg laying by determining whether it will be a good food source for offspring (Briscoe et al. 2013; Thiele et al. 2016). This again points to the suggestion of a specialised function within the Nymphalid butterflies.

1.6.3 Modes of feeding

Although most butterflies are nectar feeding, many adult nymphalid butterflies not only visit flowers, but also feed on carrion, dung, fungi, tree sap or juices from decaying fruits (Krenn et al. 2001). Morphometric studies have demonstrated that Nymphalidae have specialised a proboscis, including variations in proboscis length, wall composition, structure of tip wall and number of sensillae in comparison to other butterflies (Krenn et al. 2001). Among these adaptations, Nymphalid butterflies also have a specialised valve at the entrance to the sucking pump in the head which separates the counter current flow of nutrient fluid uptake and discharge of saliva (Eberhard and Krenn 2005). The purpose of this valve is to control liquid intake and saliva discharge and may assist with feeding modes of non-flower visiting Nymphalids as fluid discharge from the proboscis tip can assist the feeding process, such as extraction of amino acids from non-flower material (Penz and Krenn 2000; Knopp and Krenn 2003) and pollen (Krenn and Penz 1998; Estrada and Jiggins 2002; Young and Montgomery 2020).

It is possible that the function of T1 legs may have evolved in response to differing modes of feeding within the Nymphalidae, as observed in the insect family Corixidae (water boatmen) whose forelegs are modified into sensory scoop-like structures which assist with feeding on algae and detritus and assist with discrimination between edible and inedible material (Usinger 1956; Popham 1961; Nowińska et al. 2023).

1.7 Project aims and contributions

The concept of gene novelty is complex and multi-faceted. In this thesis I define 'novel genes' as protein-coding loci that are lineage-specific (i.e. taxonomically restricted genes), without close homologues in other taxa. This definition can be considered pragmatic rather than a mechanistic, as it is not always possible to determine the mechanism by which a novel gene arose. Although there are multiple mechanisms giving rise to phenotypic novelty across the animal tree of life, I focus on the emergence of gene novelty at different evolutionary timescales, the functional roles of these new genes, how genomic environment impacts the evolution of new genes and how these genes contribute to phenotypic variation in the aforementioned species.

In this thesis, I determine a consistent strategy to better understand the origin, function and importance of evolutionarily novel genes, and apply this to two big transitions in the tree of life: the Lepidoptera and the Bilateria. Until now, high-quality data from species across the animal tree of life has been limited. Data published by the Darwin Tree of Life project (DToL) (Blaxter 2022) provides high-quality genome annotation data across a breadth of species, allowing for new insights into the role of genetic novelty in the early evolution of the Lepidoptera, Ditrysia, Nymphalidae and Bilateria.

1.7.1 Chapter 2: Developing a methodology for identification of new genes arising at metazoan nodes of interest

At present a universally agreed approach on how to identify, classify and define functions of evolutionary novel genes does not exist. In this chapter, I establish the strategy used throughout my thesis to identify gene novelty on nodes of interest in the tree of life, determine the mechanism by which they arose, and downstream analyses to further understand their putative function. I also discuss limitations such as overcoming genomes annotated using different methods, defining and determining gene novelty and pitfalls concerning the pipelines used.

1.7.2 Chapter 3: Gene novelty and gene family expansion in the early evolution of Lepidoptera

Previous studies have either constructed deep-level phylogenies of Lepidoptera using a large density of species but relatively few loci or have studied specific gene families in depth. In this chapter I utilise the phylogenetically comprehensive dataset generated by the Darwin Tree of Life Project (DToL) to ask how many novel genes arise on the branch leading to Lepidoptera and ask what proportion arose by gene duplication, horizontal gene transfer or are entirely novel to the Lepidoptera. I also describe examples of new genes which were retained and duplicated further in all lepidopteran species, suggesting putative functions.

1.7.3 Chapter 4: Gene expression in the reduced first thoracic legs of a nymphalid butterfly

Previous studies have suggested that the reduced T1 legs of Nymphalid butterflies may have chemosensory functions. In this chapter I investigate whether nymphalid T1 legs transcriptomically resemble T2 and T3 legs (waking legs). I also ask whether nymphalid T1 legs have transcriptomic similarity or overlap to labial palps in nymphalids, thereby testing if they have co-opting biochemical or physiological characteristic of palps. Finally, I ask whether any transcriptomic expression differences between T1 and walking legs were related to whether this had arisen through genetic novelty originating during nymphalid evolution; for example, whether new genes had evolved specifically for roles in the T1 legs.

1.7.4 Chapter 5: Functional enrichment and gene novelty in bilaterian-specific tissues

Previous studies have investigated evolutionary novelty at the Bilaterian node of interest across a high number of species or have studied tissue-specific expression across smaller subsets of phyla within the Bilateria. In this chapter I utilise the phylogenetically comprehensive dataset generated by the Darwin Tree of Life Project (DToL); to ask how many novel genes arise on the Bilateria node using two different constrained tree topologies. I also ask which genes are enriched in Bilateria-specific tissues: namely gut, nervous system, and muscle, using a broader sampling of species across Bilateria and suggest a putative function and mode of origin

Chapter 2: Identification of new genes arising at metazoan nodes of interest

2.1 Abstract

Genetic diversity is partly responsible for the diversity of life observed today, with other factors such as environmental pressures contributing to this. Despite this, the understanding of the origin, function, and importance of evolutionary novel genes is currently limited, primarily due to the lack of available high-quality data from closely related species. This project used bioinformatic and laboratory methods to utilise the data published by the Darwin Tree of Life project (DToL) to further understand how the above factors relate to animal evolution and phenotype.

This project focussed specifically on species within the Lepidoptera and Bilateria. This was done using a large and phylogenetically comprehensive dataset of high-quality genomes to determine how new genes emerge at different evolutionary timescales, the functional roles of these new genes, how genomic environment impacts the evolution of new genes and how these genes contribute to phenotypic variation in aforementioned species. At present a universally agreed approach on how to identify, classify and define functions of evolutionary novel genes does not exist. This chapter therefore aims to define a consistent strategy to conduct analysis and discuss any arising limitations such as overcoming genomes annotated using different methods, defining and determining gene novelty, depths of nodes in evolutionary time and pitfalls concerning the pipelines used.

2.2 Introduction

Gene novelty and divergence are highly important for the evolution and adaptation of species observed today. In this context, the term 'novel' refers to protein-coding loci that are lineage-specific (i.e., taxonomically restricted genes), without close homologues in other taxa and may arise de-novo from non-coding DNA (Hoile et al. 2025). Divergence may also occur from duplication followed by divergence, horizontal gene transfer or transposable elements. It is possible that gene 'novelty' may give rise to proteins with distinct activity, function or expression patterns not observed in outgroup taxa. Determining the phylogenetic relationships between these genes is therefore integral to understanding the genetic diversity observed in extant species today. Unsurprisingly, a vast number of software tools have been developed to understand such a concept. This thesis utilises two of these tools: OrthoFinder (Emms and Kelly 2019) and GenEra (Barrera-Redondo et al. 2023).

OrthoFinder is considered the "most accurate orthology inference method available based on testing on community standard benchmarks" (Emms and Kelly 2019). OrthoFinder utilises an input of amino acid sequences of the protein-coding genes for species of interest. Using this it generates an unrooted tree for each orthogroup, which are rooted using a generated rooted species tree. An orthogroup is defined as the set of genes from multiple species descended from a single gene in the last common ancestor (LCA) of "that set of species and is the smallest set of genes such that, for all genes it contains, the orthologs of these genes are also in the same set" (Emms and Kelly 2019). The generation of orthogroups is useful for understanding gene novelty within groups of species e.g. the Lepidoptera and allows for greater comparability of genes. Defining orthogroups allows for more effective data analyses

and comparison across species and taxa, while considering gene families in their entirety provides greater biological context for interpreting results. For example, gene families may have shared function, while orthogroups are defined by shared evolutionary ancestry inferred from sequence similarity alone.

GenEra analyses the emergence of taxonomically restricted genes using a phylostratigraphy approach, annotating a given gene's likely maximum age based on shared homology in other species in context of a species tree. "This gene-family founder inference method aims to accurately infer gene ages using a small number of species of interest compared with the entire NCBI non-redundant (nr) database of all available sequences in the tree of life" (Barrera-Redondo et al. 2023). Genes identified using this method can also be referenced against the orthogroups identified through OrthoFinder analyses to further understand evolution across multiple species or taxa.

It is also important to acknowledge that it is possible that genes which duplicate and undergo extensive divergence may be placed in different orthogroups due to such a high degree of divergence in sequence. In this instance, the orthogroups would not contain the set of genes from multiple species descended from a single gene in the last common ancestor (LCA) of that set of species. This over-splitting can sometimes infer novelty at a node where this has not truly occurred and instead should exist in an orthogroup containing outgroup species belonging to more ancestral nodes.

In addition to the methods used to infer novelty within the taxa of interest, it was also important to develop a method to determine a mode of origin for each gene or orthogroups

identified. This chapter therefore gives an overview of the protocol developed and adapted through subsequent chapters to identify 'novel' genes, understand their mode of origin and infer further information about their function, expression and structure. Specimens were collected for RNA sequencing where sequence availability was limited for specific tissues, as described in the methods section. RNA data provides an additional layer of confidence that a gene we find through bioinformatics is real and also provides clues into its potential function by assessing its expression patterns (i.e. tissue or life stage-specific expression).

2.2.1 Aims:

- Define gene novelty in the context of this project.
- Utilise OrthoFinder and GenEra to develop a method to infer novelty.
- Develop a pipeline to overcome differences in annotation method between genomes at "shallow" nodes in evolutionary time e.g. order-level and at "deeper" nodes in evolutionary time.
- Define a consistent strategy to identify, classify and define functions of evolutionary novel genes.

2.3 Materials and methods

2.3.1 Specimen collection

Specimen collection was conducted at two sites within the Oxford area to obtain *Micropterix aruncella* (White-barred Gold), *Parage aegeria* (Speckled Wood) and *Maniola jurtina* (Meadow Brown) specimens for RNA extraction. Specimens were stored at -70 °C freezer and dissected on dry ice whilst frozen.

Micropterix aruncella (Micropterigidae) collection

15 *Micropterix* spp. samples were collected from Bagley Wood on 01-06-23 from three sites within the wood. Specimens were collected by sweeping through vegetation such as buttercups and brambles using small butterfly nets. Beating sheets were also used under hawthorn trees (UK grid reference SP511023).

Parage aegeria (Nymphalidae) collection

P. aegeria (Speckled wood butterfly) were collected using butterfly nets on 07-07-23 and 10-08-23 at Wytham woods (UK grid reference SP 468085). Five and seven specimens were collected on each date respectively.

Maniola jurtina (Nymphalidae) collection

33 female and 18 male *Maniola jurtina* (Meadow Brown) individuals were collected using butterfly nets between 05-07-24 and 22-08-24 at Wytham Woods, Oxford (UK grid reference SP 468085) and dissected as described in Chapter 4.

2.3.2 DNA Barcoding Protocol

It was not possible to identify female *Micropterix* species with 100% confidence from appearance alone and therefore DNA barcoding was required to clarify this. The barcoding was conducted as described in "DNA Barcoding Protocol for use with insects" (appendix). *Micropterix* specimens were dissected under a microscope on dry ice to ensure that the specimen remained frozen throughout. Dissection proceeded by removing wings and then placing the abdomen into one Eppendorf tube, the head into a second Eppendorf and combined thorax and legs into a third Eppendorf (Figure 2.1).

Following dissection, a DNA extraction was conducted on either the abdomen, head or combined thorax and legs of female *Micropterix* to determine the specific species of *Micropterix*. The tissue used to barcode each *Micropterix* specimen can be seen in Figure 2.1. The DNA extraction was conducted using ZymoResearch Quick-DNA insect DNA extraction kit. Originally, PCR amplification was conducted using LepF1-T1M13 and LepR1-T1RM13 primers, however this did not appear specific enough to the *Micropterix* sequence and instead produced a hit of *Rickettsia*. As a result, new primers were designed which were more specific to *Micropterix aruncella* and used the sequences:

Forward:

TATAATATTTAGAATATTAATTCGAATTGA

Reverse:

AAAGATGTATTTAAATTACGATCTGTTAAT

To ensure that PCR was successful, amplified DNA was run on a 1% agarose gel. Successful DNA extractions (with an approximate size of 800bp) were extracted from the gel using a UV light box and scalpel. DNA was purified using a QIAquick gel extraction kit and following the relevant protocol within the kit. DNA was eluted in 10µl of ultrapure water. Samples were sequenced by Source BioScience using the Sanger Sequencing pathway for PCR products.

2.3.3 RNA Extraction Protocol

Once all *Micropterix* DNA barcoded samples were determined to be *Micropterix aruncella*, samples underwent RNA extraction. *P. aegeria* and *M. jurtina* samples were identified from appearance and therefore did not require DNA barcoding. Further tissues were dissected for *P. aegeria* and *M. jurtina* as these were more likely to undergo RNA sequencing successfully in comparison to smaller *M. aruncella* tissues. *P. aegeria* samples were dissected into eight tissue types: head, palps, front legs, remaining legs, thorax, abdomen, gonads, and wings, while M. jurtina tissues were dissected into four tissue types: palps, T1, T2 and T3 legs. RNA was extracted from all species using RNeasy Plus Micro kit and the supplied protocol. Tissues from samples were pooled to obtain a greater concentration of RNA for each tissue type and eluted in 15µl of ultrapure water.

RNA sequencing was performed on *M. aruncella* and *P. aegeria* tissues using Illumina NovaSeq platform (Genewiz / Azenta), at a coverage of 10M paired end 150bp reads per sample. RNA sequencing was performed on *M. jurtina* replicates using the Illumina NovaSeq X Plus Series (PE150) Sequencing System (Novogene), at a coverage of 20M paired-end 150bp reads per sample.

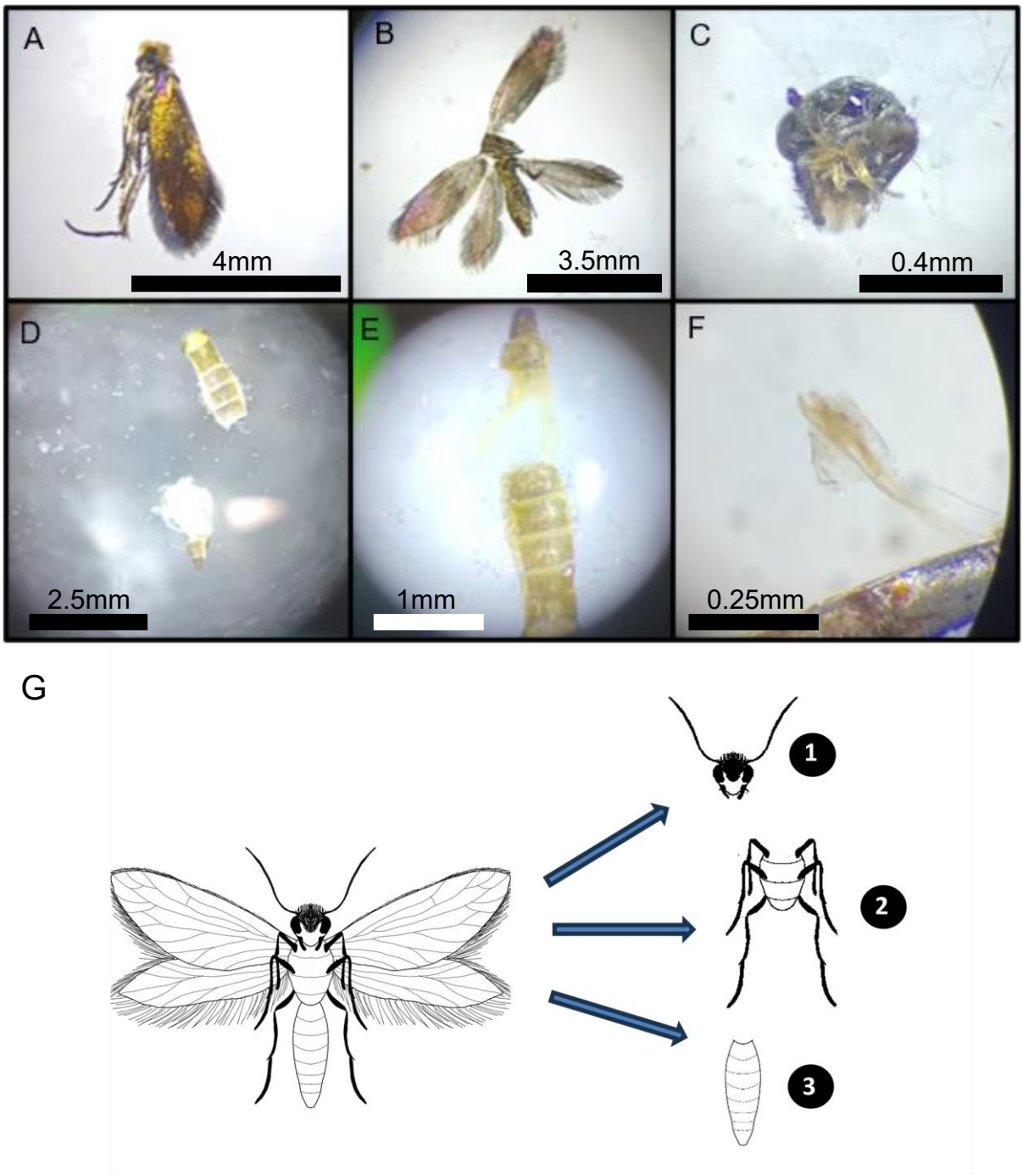


Figure 2.1 – Micropterix aruncella dissection protocol. ***A*** *Whole female M. aruncella specimen* ***B*** *Dissected female M. aruncella specimen with head removed and separated thorax and abdomen.* ***C*** *M. aruncella female head.* ***D*** *M. aruncella female abdomen (top) and Malpighian tubules.* ***E*** *M. aruncella female ovaries dissected from abdomen.* ***F*** *M. aruncella male genitals.* ***G*** *Illustration of M. aruncella dissection into three tissues: 1 – head, 2 – legs and thorax, 3 – abdomen.*

2.3.4 Computational analysis

OrthoFinder

Proteome data were obtained from Ensembl Rapid Release http://rapid.ensembl.org (accessed between October 2023 and August 2024 for all analyses); these proteome predictions are based on the Ensembl genebuild annotation pipeline. Primary transcripts were obtained from the predicted proteome data and OrthoFinder v2.3.14 was run to determine orthogroups within the dataset for both Lepidoptera sp analyses and OrthoFinder v3.0.1b1 was used for Bilateria analyses (Emms and Kelly 2019). To relate these to a species tree, amino acid sequences from single copy orthologues from the OrthoFinder output, and present in all species were aligned using MAFFT v7.505 (Katoh and Standley 2013), trimmed using trimAl v1.4.rev15 build (Capella-Gutiérrez et al. 2009), concatenated with PhyKIT (Steenwyk et al. 2021) and the concatenated alignment used to generate a species tree using IQ-TREE version 2.0-rc1 (Minh et al. 2020). Orthogroups gained on nodes of interest were extracted using Orthoparser (github.com/PeterMulhair/ortho_parser). To test further whether orthogroups specific to Lepidoptera were present in outgroups but missing from predicted proteomes, and to identify gene mode of origin, an additional pipeline was developed (Figure 2.2).

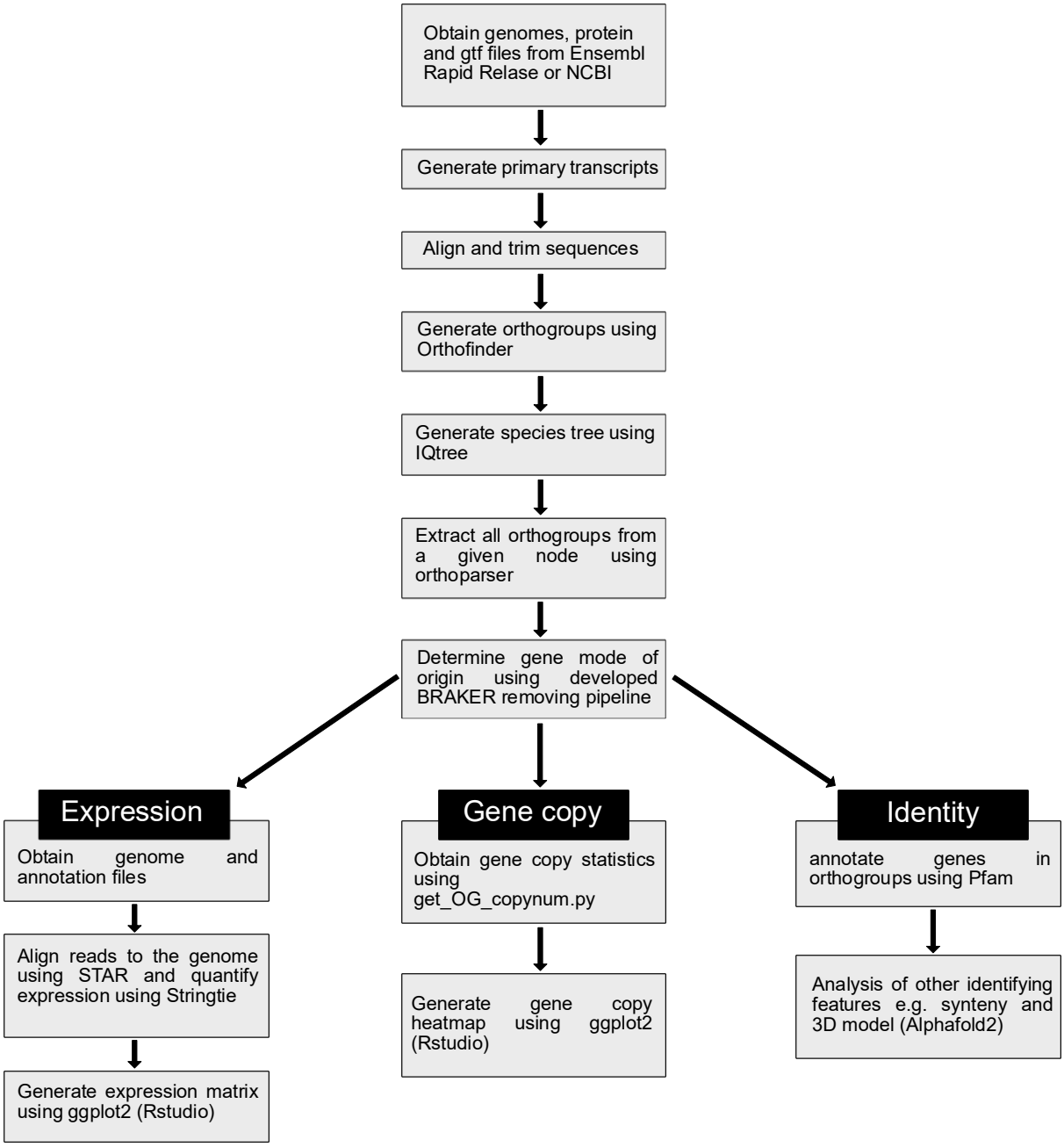


Figure 2.2 – OrthoFinder analysis pipeline demonstrating required inputs and downstream analysis to further understand potential function of orthogroups / genes of interest. This constitutes three main pathways. Firstly is using RNAseq tissue specific expression data to determine expression patterns. If replicates are available this can also be used for a differential gene expression analysis. Secondly, a python script can be used to generate gene copy number for orthogroups of interest to determine whether there are duplication patterns within specific groups e.g. Nymphalid butterflies. Lastly, BLAST and Pfam can be used to annotate genes to give further information regarding their function. Sequence data can also be used to determine synteny and generate 3D models to determine structural homology.

GenEra

GenEra is a method which can estimate gene-family founder events which uses genomic phylostratigraphy (Domazet-Lošo et al. 2007). In this thesis GenEra was utilised for nodes in deep time such as the Bilateria, in combination with OrthoFinder. HOGs are sets or groups of genes which have descended from a single common ancestor within a taxonomic range of interest (Altenhoff et al. 2013). Hierarchical orthogroups (HOGs) were obtained for the bilaterian node as HOGs can distinguish whether a gene family expansion may have taken place before or after the emergence of the Bilateria, whereas just using orthogroups may group paralogs together which would result in a lack of clarity as to whether a gene may be Bilateria-wide or lineage specific within the Bilateria (Zhou et al. 2020). Additionally, HOGs typically handle gene loss more effectively in comparison to regular orthogroups, aiming not to overspilt orthogroups when losses occur (Waterhouse et al. 2012). This therefore means that HOGs typically provide better clarity across multiple evolutionary depths, reducing the error and stabilizing inferences at deep time. This is less likely to occur on “shallower” nodes and therefore using orthogroups alone is sufficient for these analyses, as observed in the case of the Lepidoptera node.

For the same reasons as above, using the pipeline developed in Figure 2.4 would not be suitable for deep node analyses and therefore GenEra was used to assist in determining the mode of origin of HOGs of interest within the Bilateria analysis (Figure 2.3).

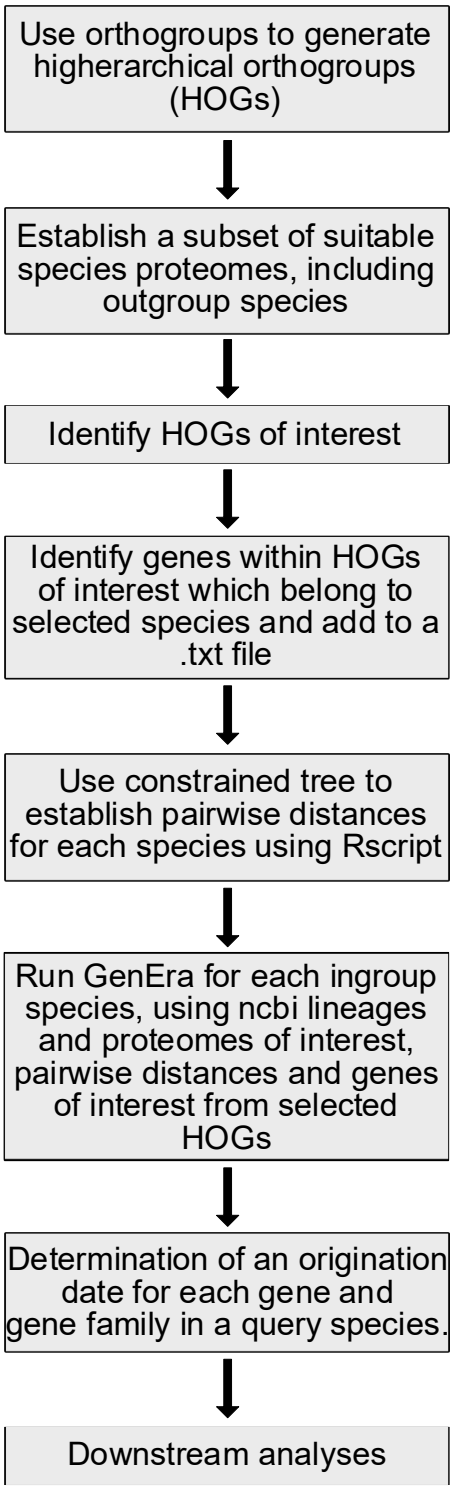


Figure 2.3 – GenEra analysis pipeline and required inputs to determine the origination date for gene families / hierarchical orthogroups (HOGs) of interest. This can be used to infer the mode of origin of a HOG of interest for deeper nodes such as the bilaterian node.

2.4 Results and discussion

2.4.1 Development of a pipeline to identify gene mode of origin

Originally, synteny analyses were conducted by hand on the Lepidoptera dataset to determine the most appropriate methods. First, the ID of the gene of interest from the relevant orthogroup was obtained from the OrthoFinder output. This gene ID was entered into Ensembl Rapid Release for the relevant species. From the results of this search, chromosome number, location and amino acid size of the gene was noted. Two "marker genes" either side of the gene of interest which had an identifiable domain (four in total) were selected and noted. Amino acid sequences from the orthogroup gene of interest, plus four marker genes were used in BLAST searches against the genomes of the Lepidoptera species: *Danaus plexippus*, *Papilio machaon*, *Tinea trinotella* and *Micropterix aruncella*, plus outgroup arthropod species *Limnephilus lunatus*, *Limnephilus marmoratus*, *Limnephilus rhombicus*, *Glyphotaelius pellucidus*, *Bibio marci*, *Drosophila melanogaster*, *Adalia bipunctata* and *Vespula vulgaris*. Information from BLAST results was used to generate synteny diagrams and determine if the gene of interest was one of four classes: 1. Arising from a gene duplication event / fast evolving gene 2. Not a novel Lepidopteran gene 3. Possible horizontal gene transfer 4. Novel Lepidopteran gene.

This was later developed into a decision tree-style Python script to automate the process and reduce analysis time (Figure 2.4). This pipeline began by conducting a BLASTp search of the genes in the orthogroups at the node of origin against all lepidopteran and chosen outgroup Ensembl Genebuild genomes. If a hit is identified, this suggests that this gene has arisen via gene duplication, or that this is a fast-evolving gene. This can be further clarified

by using all the genes in the search subject's orthogroups and outgroups to build a synteny map and gene tree. If no hits are identified within the selected Ensembl Genebuild genomes, there are two possible outcomes. Any genes which did not have a hit were then BLASTed against BRAKER (Augustus-Gaius) annotated genomes of selected outgroups. In the case of the Lepidoptera, four Trichoptera species were used. If hits were found to these genomes, then genes were excluded from further analysis and not deemed to be novel to the lepidopteran lineage. If no hits were identified in BRAKER-annotated genomes, genes were BLASTed against all genomes available in the tree of life. If a hit was found in a distantly related species e.g. fungi or bacteria, this is a potential horizontal gene transfer (HGT) event and further analyses should be conducted. If no hits were found this is deemed to be a novel gene, unique to Lepidoptera.

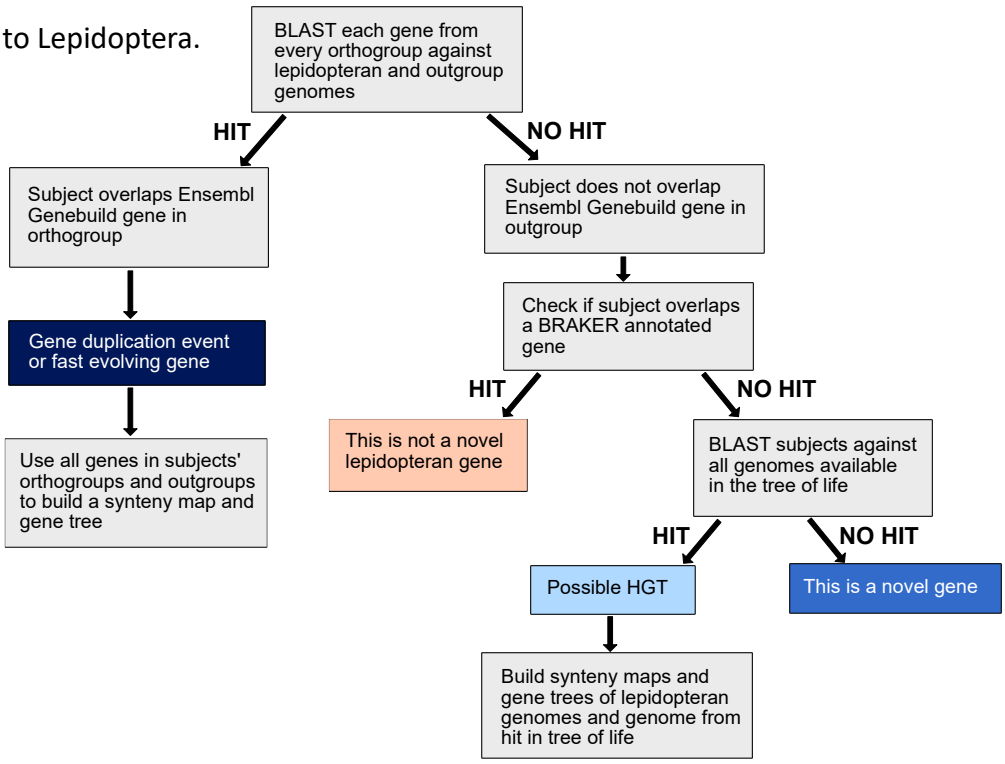


Figure 2.4 – Gene origin pipeline which leads to four potential outcomes. Genes of interest are used in BLASTp analysis against Lepidopteran and close outgroup Ensembl Genebuild genomes. If a hit is identified this gene has arisen from gene duplication or is a fast-evolving gene. If no hit is observed a BLASTp search is conducted against BRAKER (Augustus-Gaius) annotated genomes for closely related outgroups. If a hit is observed, the gene of interest is excluded from further analysis. If no hit is observed, the gene of interest is used in a BLASTp search against all genomes in the tree of life. If a hit is observed in a distantly related species e.g. bacteria or fungi, this is a potential horizontal gene transfer (HGT) event and further analysis is required to confirm this. Finally, if no hit is observed, this is deemed a novel lepidopteran gene.

2.4.2 Removal of potential artifact genes increases confidence in 'novel' gene discovery

Gene copy number was calculated for every orthogroup originating on the desired nodes of interest throughout this project. Figure 2.5 shows the Lepidoptera node plotted against a phylogenetic tree as an example. Gene copy number was separated into categories of 1,2,3,4,5 or more than 5 copies (Figure 2.5A and C). It was clear from Figure 2.5A that a large number of orthogroups had genes present in a small number of relatively distantly related Lepidoptera species. This raised concern that some of these results may be noise as a result of incorrectly grouped proteins due to spurious homology. This is because a genuine orthogroup would be expected to contain genes in a number of closely related species rather than scattered through a number of distantly related individuals. As a result, an occupancy graph of number of species present in each orthogroup was created (counting each species with the gene present only once, therefore discounting more than one gene copy present). From this graph, a 75% cutoff was applied, meaning that if fewer than 75% of all Lepidopteran species used in this analysis were present in a given orthogroup, it would be excluded from the final gene copy number plot (Figure 2.5B). To give an idea of the extent of this cutoff threshold, of the initial 184 orthogroups identified at this node, 87 were included following application of the 75% cutoff. Although there is a risk of excluding true novel orthogroups, this threshold value ensures a high degree of confidence that all genes that meet the threshold are true novel orthogroups, excluding noise and spurious homologies.

In the Bilateria dataset, a more relaxed cutoff was used due to the extensive gene loss which has occurred across the Bilateria (Guijarro-Clarke et al. 2020). As a result, genes from a given orthogroup would only need to be present in 25% of Bilaterian species and present in both

Amphioxus and Octopus to be considered a real HOG. These species were selected due to the availability of RNAseq data which was used in subsequent analyses in Chapter 5.

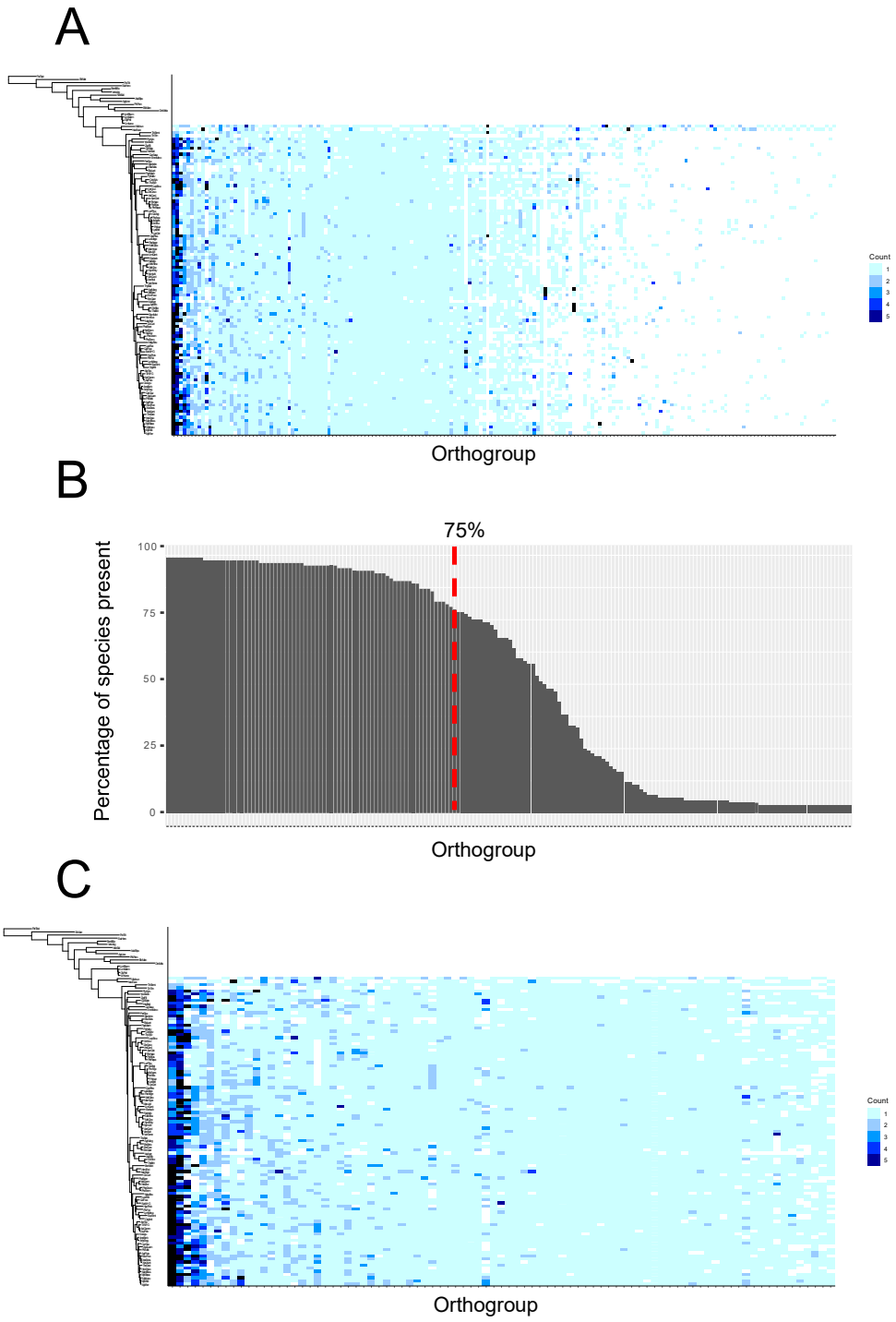


Figure 2.5 – Demonstration of filtering step employed to reduce the likelihood of identified orthogroups arising from spurious homology. ***A*** *gene copy number heatmap of all orthogroups at the* Lepidoptera *node and phylogenetic tree (184). Orthogroups on the right-hand side of the figure have genes present in minimal species and are likely arising from noise.* ***B*** *Occupancy plot of number of species which have a gene present from a specific orthogroup, showing 75% cutoff.* ***C*** *Gene copy number heatmap post filtering for the 75% cutoff level leaving 87 orthogroups remaining. Legend denotes number of genes present in a given species.*

2.4.3 Gene identification protocol assists in novel gene identification and potential functions

The following protocol has been developed and adapted throughout each of the subsequent results chapters in this thesis (Figure 2.6). All analyses (Lepidoptera and Bilateria) began with identification of species of interest within the taxon studied, usually to ensure a strong representation of phyla or families (depending on the breadth of species in the study). Where possible, genomes, protein and gtf files were downloaded from Ensembl rapid release, and if not possible, these were obtained from NCBI. RNAseq data was either downloaded from NCBI SRA or generated in the laboratory.

Initially, only the OrthoFinder pipeline was used to identify novel genes, however the GenEra pipeline was also used in the Bilateria analysis (Chapter 5). Both OrthoFinder and GenEra pipelines were run as described in the methods above. For analysis on the Lepidoptera and subsequent internal nodes, genes were then run through the gene origin pipeline as described in Figure 2.4.

Alongside this, RNAseq data was obtained from NCBI SRA. Where this was not possible, specimens were collected and upon successful identification, dissections were completed on dry ice, RNA was extracted, and sequencing was conducted to generate tissue-specific expression data. This allows for creation of expression matrices and differential gene expression analyses.

In addition to evaluating expression data, gene copy number was calculated for relevant orthogroups, in addition to annotation using Pfam (Finn, et al., 2014) and BLAST, synteny analyses and 3D modelling to understand features of genes of interest, with the aim of understanding more about the function and importance of these genes. Figures including phylogenetic trees and heatmaps generated in R used ggtree (Yu, et al., 2016), ggplot2 (Wickham, 2016), and Pheatmap (Kolda, 2019), Chromosome plots were created using RIdeogram (Hao, et al., 2020) and were edited using Affinity Designer 2 (Affinity, 2024) (used for data in Chapters 3,4 and 5).

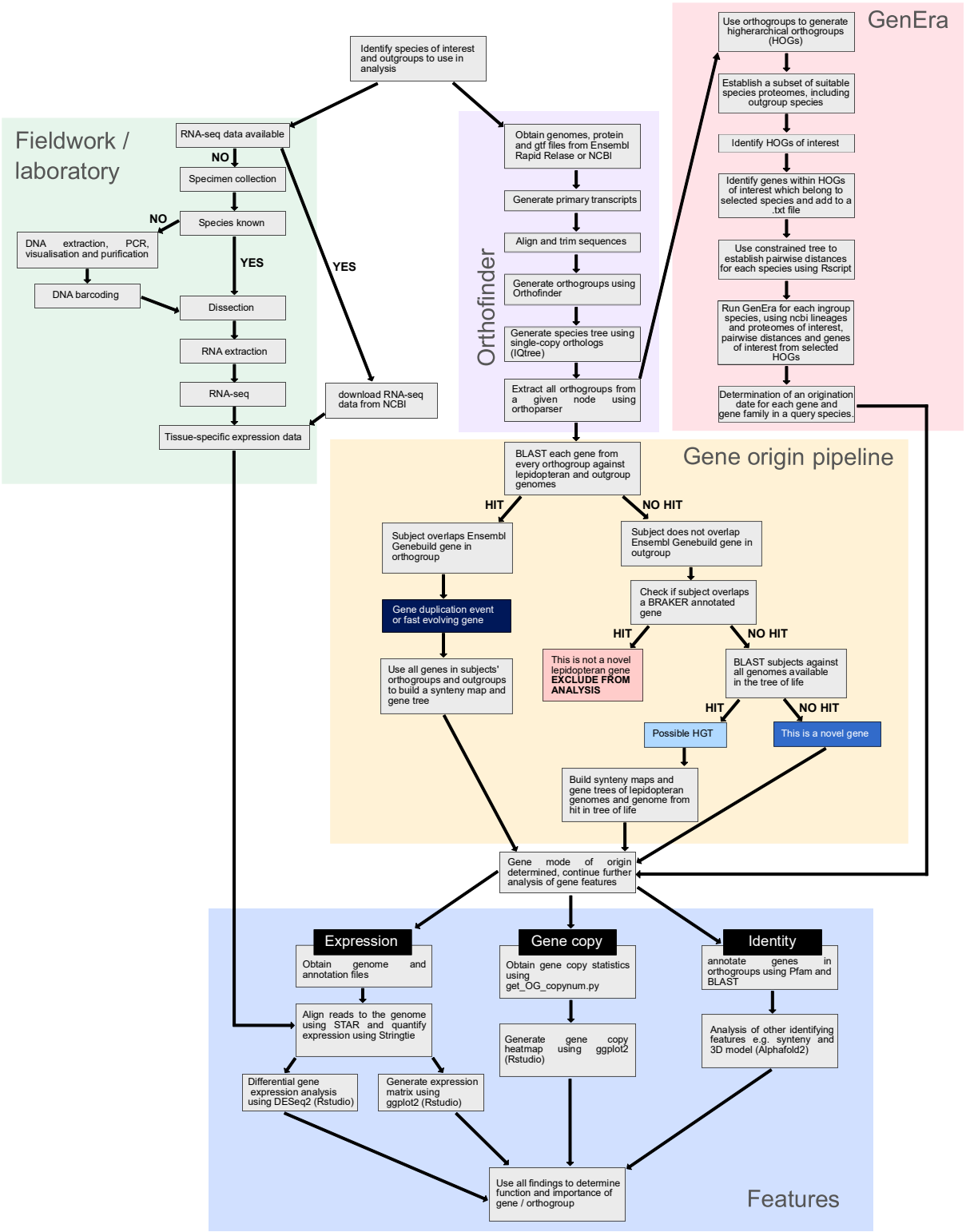


Figure 2.6 – Overall gene identification protocol involving laboratory and specimen collection to obtain tissue-specific expression data (green) , OrthoFinder (purple) and GenEra (red) novel gene identification methods, gene origin pipeline (orange) and how each of these steps generate data to conduct downstream analyses to further understand the features, and potentially identity and function of genes of interest (blue).

2.5 Conclusions

The overall aim of this chapter was to define a consistent strategy to identify, classify and define functions of evolutionary novel genes. Analyses have revealed that genes which duplicate and undergo extensive divergence are often placed in different orthogroups, meaning they do not always contain a set of genes from multiple species descended from a single gene in the last common ancestor (LCA) of that set of species.

With this knowledge in mind, a pipeline was developed to categorise novel genes into one of three categories: novel, arising from duplication or horizontal gene transfer. Additionally, it utilised BRAKER annotated genomes of outgroup species to avoid falsely determining genes present in outgroups as novel genes. This allows for overcoming of genomes which are annotated in different ways. Duplication novelty was identified within the Lepidoptera dataset used to refine this strategy. For this reason, duplication and divergence is used as a classification of gene novelty within this analysis. Due to the depth of the Bilaterian node, GenEra was used to assist in determining the mode of origin of gene families rather than using the developed pipeline.

From conducting an occupancy analysis, it is possible that false homologues may exist in orthogroups containing genes present in a small number of distantly related species, likely arising from poor annotation or truncated sequences. To overcome this, a cutoff threshold may be applied (where a specific percentage of species used in the analysis must be present to determine a true orthogroup). Additionally, focusing on genes in high copy number which are present in many species (as in Chapter 3) can also be used to overcome this.

This chapter therefore presents a strategy used in subsequent chapters which encompasses the identification of new genes, methods to determine mode of origin and downstream analyses to investigate functional analysis. It is also apparent that variation in the depth of nodes investigated and the evolutionary time period which they encompass requires variations in the approaches used to ensure the most appropriate prediction possible. As discussed, the methods are not without limitations, however allow for analysis which encompasses gene novelty.

Appendix

DNA BARCODING PROTOCOL (cytochrome oxidase subunit I); for use with insects*.

Prepared by Michał Jezierski.

N – number of DNA samples

SAFETY PRECAUTIONS: Wear lab coat, gloves and protective glasses at all times. Take special care during agarose gel preparation (see below).

MATERIALS:

DNA extraction*: use ZymoResearch Quick-DNA insect DNA extraction kit:

- 3 x N collection tube (provided with kit) or 1.5 mL microcentrifuge tube
- 1 x N 1.5 mL microcentrifuge tube – not part of the kit
- ZymoResearch (ZR) Bashing Bead Tubes (1 x N)
- Zymo-Spin IIIF filter (1 x N)
- Zymo-Spin IC column (1 x N)
- ZR Bashing Buffer (store at room T); 750 uL per reaction
- ZR Genomic Lysis Buffer (store at room T; no 2-mercaptoethanol necessary); 1.2 mL per reaction
- ZR Pre-Wash Buffer (store at room T); 200 uL per reaction
- ZR gDNA Wash Buffer (store at room T); 500 uL per reaction
- ZR DNA elution buffer (store at room T); 20-30 uL per reaction

PCR: 15 uL reactions (with DNA sample):

- GoTaq Green Buffer 5x (store at -20C)
- 10 mM dNTPs mix; 1.7 – 2.0 uL per reaction (store at -20 C)
- 10 uM primer solutions: LepF1/R1 or LepF1/R1-M13; 1 uL of each per reaction (store at -20C). Primer sequences at end of this document.
- molecular biology grade water
- approx. 0.05 uL of GoTaq polymerase per reaction (store at -20C)
- clean 1.5 mL tube (for master mix)
- clean PCR tube (one per reaction) + lids

Example volumes with 1 uL of DNA sample:

REAGENT	**1 x 15 uL reaction = 14 uL**	**10 x 15 uL reaction = 140 uL**
Water	7 uL	70 uL
5x GoTaq Green Buffer	3 uL	30 uL
10 uM Primers	1 uL of each primer = 2 uL	10 x 1 uL of each primer = 20 uL
10 mM dNTP mix	2 uL	20 uL
GoTaq	0.05 ul (negligible volume)	0.5 uL (negligible)

Thermocycler programme:

START

- 94.0 C for 60 seconds - 94.0 C for 30 seconds
- 94.0 C for 30 seconds - 54.0 C for 40 seconds x 35
- 50.0 C for 40 seconds x 5 - 72.0 C for 60 seconds
- 72.0 C for 60 seconds - 72.0 C for 10 **minutes** followed by ∞ at 4.0C

END

Gel electrophoresis quality control:

- 1% agarose solution
- 1x TBE buffer (approx. 1 litre)
- SYBR Safe; I used 3 uL per 50 mL and 8 uL per 200 mL of agarose
- HyperLadder 50 bp or other ladder with same resolution (fragment is 700 bp)

PCR product purification:

- Any PCR purification kit; follow specific protocol

Sequencing (BioSource):

- At least 5 uL of DNA sample (as concentrated as possible)
- At least 5 uL of primer per sample at 0.5x concentration of what is used in PCR

* Quick DNA works best with rather large samples (I would say at least 2 mg of tissue); for possibly (haven't tested) more efficient extraction methods as well as other thermocycler programmes check: Kress, W., & Erickson, D. (2012). *DNA Barcodes : Methods and Protocols* (Methods in Molecular Biology, Methods and Protocols, 858). Totowa, NJ (available on SOLO).

PROTOCOLS:

DNA EXTRACTION:

1. Add tissue sample (up to 10 mg according to manual but will work with more) to ZR bashing bead tube; add 750 uL of ZR Bashing Buffer.

2. Secure ZR bashing bead tubes in **TissueLyser 2** (equipment room) [*pull the small piston up to unscrew and release the box for tubes*]; once secured (screw tightly and balance for large sample sizes!) start the machine at **30 x 1/s** for **15 minutes**.

3. Take out ZR bashing bead tubes from the tissue lyser; centrifuge the ZR bashing tubes at ≥ 10,000g for 60s.

4. Transfer 400 uL of supernatant from centrifuged ZR bashing bead tubes to Zymo-Spin III-F filter in a collection tube; centrifuge at 8,000 g for 60s. Discard the filter afterwards.

5. Add 1.2 mL of ZR Genomic Lysis buffer [*2 times 600uL using P1000 works fine*] to the collection tube from step 4; mix by gently pipetting up and down. May get foamy.

6. Transfer 800 uL of mix from step 5 to a Zymo-Spin IC column in a **new** collection tube; centrifuge at 10,000 g for 60s. DO NOT DISCARD the remaining mix from step 5.

7. Discard flow through from the collection tube and repeat step 6.

8. Transfer Zymo-Spin IC column to a **new** collection tube. Add 200 ul of ZR Pre-Wash Buffer; centrifuge at 10,000g for 60s. Discard the flow through.

9. Add 500 uL of ZR gDNA Wash Buffer to Zymo-Spin IC column; centrifuge at 10,000 g for 60s. Discard the flow through.

10. Transfer IC column to a clean 1.5 mL microcentrifuge tube. Add 20 - 30 uL of ZR DNA Elution Buffer directly onto the column matrix (the white thing); centrifuge at 10,000g for 30s. Discard the column.

11. Store DNA on ice for use on the same day or at -20C.

PCR:

*All ingredients should be thawed at room temperature if necessary and stored on ice afterwards. Remember to include **positive control** [DNA extracted from a whole insect body usually works] and **negative control** [molecular biology grade water].*

1. Calculate the necessary volumes of each ingredient in the master mix and decide what and how much goes into PCR tubes before starting the preparation.

2. Prepare a clean 1.5 mL microcentrifuge tube for master mix.

3. Add water and then GoTaq Green Buffer to the tube. Vortex to mix.

4. Add dNTPs (and primers if you are to use the same ones for every reaction). Vortex to mix.

5. Add GoTaq polymerase. Vortex to mix. Store the master mix on ice.

6. Label the PCR tubes with sample names [*writing on the upper part of tubes prevents ink decay from thermocycling*]. Aliquot samples (and primers if not added to master mix) to PCR tubes.

7. Aliquot master mix to PCR tubes. Cover with lids and label the lid to identify the PCR run later.

8. Put in the block in thermocycler and remember to close it. Samples should be placed centrally if space permits [*e.g. with one strip, place horizontally in the middle*].

9. Use the programme described in **Materials** section.

10. After the run samples can be kept in the machine at 4.0 C if used on the same day or stored at -20C.

GEL ELECTROPHORESIS QUALITY CONTROL:

1. Prepare 1x TBE buffer; mix 100 mL of 10x stock TBE buffer with 900 mL of pure or ultra-pure water.

2. Prepare agarose gel solution. Measure out required amount of agarose on a weight using a plastic well. Pour the agarose into appropriate flask.

3. Add appropriate volume of TBE buffer to the flash with agarose to make 1% agarose solution [*0.5g agarose per 50 mL of TBE for 50 mL gel; 2g of agarose per 200 mL of TBE for 200 mL gel*].

4. Put the flash in a microwave and cover with a beaker to stop evaporation. Heat until the fluid is clear. **BE CAREFUL** – see step 6.

5. Prepare the gel setting assembly – you will need one plate, 2 plate holders and 1-2 combs. Enclose the plate with plate holders so that the gel can't flow out. Put the combs where you want them.

6. Gently take the beaker away. Grasp the flask using a glove/anything that will prevent you from burning yourself upon touching it. Agitate the flask gently with the opening **facing away** from you – agarose will get superheated in the process and start to boil upon agitation. You want to control when this happens. Take the flash out.

7. Cool down agarose solution to about 50C [*use hand to decide, should be hot but not burning*] by pouring cold tap water onto the flask [*swirl it gently under the tap – avoid air bubbles*].

8. Add appropriate volume of SYBR safe to the solution and mix it gently.

9. Gently pour the gel onto the assembly from step 5.

10. Leave to set for approx. 45 minutes.

11. Remove the comb by pulling it out gently. Remove the plate holders by pulling them away gently. Place the plate with gel in electrophoresis tank.

12. Pour 1x TBE buffer into the electrophoresis tank so that it covers the gel and the wells (approx. 1mm above the gel is fine).

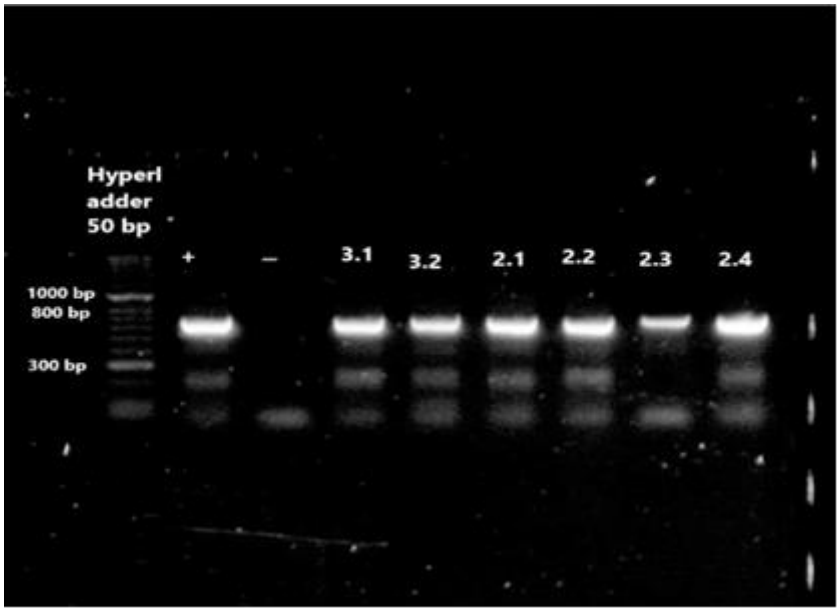
13. Load 2 uL of PCR products directly into the wells [*Green Buffer contains loading dye*]. Load 1 uL of HyperLadder 50 bp.

14. Put the cover on, attach the cables. Turn the electrophoresis machine on and set at 100V. Run for an hour.

15. Turn off the machine before removing the gel (you may electrocute yourself). Take the lid off. Put the gel in gel visualising unit.

16. Visualise the gel. Auto exposure may aid in seeing the band (select area with samples not including the ladder).

17. Example of a good gel result is shown on the next page. Bands like this should successfully be sequenced.



Example of a good gel result

PCR PURIFICATION AND SEQUENCING:

1. Follow the protocol for your PCR purification kit. Aim to obtain as concentrated DNA as possible.

2. For BioSource sequencing remember to send in primers (even with M13 add-on; they can't get it right without supplied primer). Send in samples and adequate amount of primers: that is at least 5 uL of DNA and 5 uL of 0.5x primer per each sample.

3. Barcode sequences can be verified using BLAST or preferentially BOLD Systems v4. FASTA search is available on BOLD Systems under "Identification". If voucher specimen is to be stored somewhere (OUMNH will be happy to do that!) then the barcode can be uploaded to BOLD.

PRIMER SEQUENCES

Primers synthesized by IDT.

LepF1 and LepR1 are standard insect CO1 primers, optimised for Lepidoptera but also tested here for Trichoptera.

LepF1-T1M13 and LepR1_t1RM13 are almost the same as the above but tailed with M13 and Reverse M13 sequences to make sequencing simpler**.

** For BioSource sequencing the M13 tails won't give as good results if during ordering you tell them that the primers are M13s, as they would if you send the primer with the sample. No idea why.

Either F primer can be used with either R primer.

In addition, there is a tRNAWFM13 primers which could be used in place of one of the F primers; this is based on a tRNA 5' to the CO1 gene. It has not been tested here.

Primers come dry and should be dissolved in molecular biology grade water to a concentration of 100uM (e.g. if 22 uMoles of primer synthesized, dissolve in 220 ul). This stock is then diluted 10X to make the 10uM solution used in the PCR reactions.

LepF1: ATTCAACCAATCATAAAGATATTGG

LepR1: TAAACTTCTGGATGTCCAAAAAATCA

LepF1_T1M13: TGTAAAACGACGGCCAGTATTCAACCAATCATAAAGATATTGG

LepR1_t1RM13: CAGGAAACAGCTATGACTAAACTTCTGGATGTCCAAAAAATCA

tRNAWF1M13: TGTAAAACGACGGCCAGTAAACTAATARCCTTCAAAG

References for origin of these primers

https://link.springer.com/protocol/10.1007%2F978-1-61779-591-6_3

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2996951/

https://link.springer.com/protocol/10.1007/978-1-61779-591-6_21

Chapter 3: Gene novelty and gene family expansion in the early evolution of Lepidoptera

Asia E. Hoile, Peter W. H. Holland, Peter O. Mulhair

Introduction to manuscript

This chapter has been published as a first author original research article in the Journal *BMC Genomics* (https://doi.org/10.1186/s12864-025-11338-x). It was first published online on 19th February 2025 and is reproduced here with minor changes as permitted as part of an integrated thesis. The authors credited on this manuscript, in order, are Asia E. Hoile, Peter W. H. Holland*, Peter O. Mulhair* (* - corresponding author).

Author contributions

P.W.H.H. and P.O.M. conceived the study and oversaw the research. A.E.H. and P.O.M. designed analyses and carried out the bioinformatic research presented. A.E.H., P.W.H.H. and P.O.M. interpreted all results. A.E.H. wrote the initial draft of the manuscript, and P.W.H.H. and P.O.M. edited versions. All authors read and approved the final manuscript.

3.1 Abstract

Almost 10% of all known animal species belong to Lepidoptera: moths and butterflies. To understand how this incredible diversity evolved we assess the role of gene gain in driving early lepidopteran evolution. Here, we compared the complete genomes of 115 insect species, including 99 Lepidoptera, to search for novel genes coincident with the emergence of Lepidoptera.

We find 217 orthogroups or gene families which emerged on the branch leading to Lepidoptera; of these 177 likely arose by gene duplication followed by extensive sequence divergence, 2 are candidates for origin by horizontal gene transfer, and 38 have no known homology outside of Lepidoptera and possibly arose via *de novo* gene genesis. We focus on two new gene families that are conserved across all lepidopteran species and underwent extensive duplication, suggesting important roles in lepidopteran biology. One encodes a family of sugar and ion transporter molecules, potentially involved in the evolution of diverse feeding behaviours in early Lepidoptera. The second encodes a family of unusual propeller-shaped proteins that likely originated by horizontal gene transfer from *Spiroplasma* bacteria; we name these the Lepidoptera *propellin* genes.

We provide the first insights into the role of genetic novelty in the early evolution of Lepidoptera. This gives new insight into the rate of gene gain during the evolution of the order as well as providing context on the likely mechanisms of origin. We describe examples of new genes which were retained and duplicated further in all lepidopteran species, suggesting their importance in Lepidoptera evolution.

3.2 Background

Diversification and adaptation depend on genetic change but associating genomic drivers underpinning phenotypic change is challenging. Many studies have approached this problem by starting with phenotypic polymorphisms within a species or differences between closely related species and then using genomic and experimental approaches to identify underlying causative mutations. Several of these studies have uncovered sequence changes in non-coding DNA affecting the expression of conserved genes (Carroll et al. 1994; Wucherpfennig et al. 2022; Livraghi et al. 2024; Tian et al. 2024). Other studies have identified coding sequence changes causing amino acid substitutions, or loss of function, as causative mutations that were subsequently fixed under selection (Hoekstra and Coyne 2007; Ota et al. 2007; Dutrow et al. 2022). It is clear, however, that changes in existing genes, whether they affect gene expression or protein sequence, cannot explain all adaptive evolution. Perhaps the best evidence lies in comparative genomics: when genome sequences are compared ample evidence is uncovered for the role of gene number variation, gene duplications, and gene novelty in driving evolution and adaptation (Paps and Holland 2018; Richter et al. 2018; Fernández and Gabaldón 2020; Guijarro-Clarke et al. 2020; Thomas et al. 2020; Cicconardi et al. 2023).

Gene novelty is a multi-faceted concept (Long et al. 2003; Kaessmann 2010; Haggerty et al. 2014; Van Oss and Carvunis 2019). We define novel genes as protein-coding loci that are lineage-specific (i.e. taxonomically restricted genes), without close homologues in other taxa (Rödelsperger et al. 2019). This is a pragmatic definition rather than a mechanistic one since we cannot always determine the mechanism by which a novel gene arose. The mode of

origin of taxonomically restricted genes might be gene duplication followed by extensive sequence divergence (Ohno 1970; Rastogi and Liberles 2005; Conrad and Antonarakis 2007; Sémon and Wolfe 2008; Holland et al. 2017; Dubose and De Roode 2024), fusion of distinct loci or a transposable element into a pre-existing locus (McClintock 1950; Long and Langley 1993; Leonard and Richards 2012; Bornberg-Bauer and Albà 2013; Cosby et al. 2021; Mulhair et al. 2023a), horizontal gene transfer (Husnik and McCutcheon 2018; Li et al. 2022; Keeling 2024) or *de novo* origin from non-coding DNA (Levine et al. 2006; McLysaght and Hurst 2016; Van Oss and Carvunis 2019; Zhao et al. 2024). Whatever the mode of origin, novel genes likely reflect novel biology as they will encode proteins with potentially distinct activity or function not present in the outgroup taxa. Examples in arthropods include horizontally acquired genes from bacteria underpinning adaptations to phytophagy (Wybouw et al. 2016) or male courtship behaviour in moths and butterflies (Li et al. 2022), and divergent gene duplicates recruited for limb patterning in water striders (Santos et al. 2017).

Here we investigate the origin of novel genes in the early evolution of the insect order Lepidoptera. Lepidoptera are a holometabolous order of insects consisting of the moths and butterflies and comprise nearly 160,000 described species or 8–10% of known animal species on the planet (Kawahara et al. 2019). The oldest members of the Lepidoptera crown group are estimated to have appeared in the Late Carboniferous (~ 300 mya) and were likely pollen feeders, with the evolution of a tube-like proboscis and nectar feeding occurring later in the Middle Triassic (~240 Ma). Today the Lepidoptera inhabit almost all terrestrial ecosystems, displaying a large variety of ecological adaptations relating to feeding, defence, and survival (Mitter et al. 2017; Kawahara et al. 2019; Kawahara et al. 2023). Larvae of the

earliest lineages were likely endophagous, feeding internally in the tissue of nonvascular land plants, with adults possessing mandibulate chewing mouthparts (as seen in extant members of the family Micropterigidae) suitable for pollen feeding (Krenn 2010; Bazinet et al. 2017). A period of diversification early in the evolution of Lepidoptera coincided with the development of the tube-like proboscis, used by adults to feed on nectar, and the expansion of angiosperms. The remarkable diversity present in Lepidoptera today can be attributed to continued co-evolution with diverse angiosperm lineages, major transitions in morphology and habitat, and the emergence of diverse feeding behaviours (Kawahara et al. 2023).

To assess whether novel genes arose in the early evolution of Lepidoptera, and whether any of these underwent further gene family expansion, we require complete genome sequences from a dense sampling of Lepidoptera and related insect orders. Previous studies have constructed deep-level phylogenies of Lepidoptera using a large density of species but relatively few loci (Kawahara et al. 2019), while other studies have studied specific gene families in depth (MacIas-Muñoz et al. 2019; Mulhair et al. 2023b; Mulhair et al. 2023c). Large genomic datasets have only recently become available through sequencing consortia such as the Darwin Tree of Life Project (Blaxter 2022) affiliated to the Earth Biogenome Project (Lewin et al. 2018). Here, we avail of this data by analysing 115 high quality insect genomes and identify 217 novel genes that arose on the stem lineage of Lepidoptera and 541 novel genes that arose on the stem lineage of the Ditrysia, a major clade encompassing most of lepidopteran diversity (Rota et al. 2022). We infer the likely modes of origin for these novel genes. We then focus attention on two gene families gained on the ancestral lepidopteran branch that were subsequently retained across all species, suggestive of recruitment to important roles in lepidopteran biology. One is a gene family encoding

divergent sugar transporter proteins; the other is a likely horizontal gene transfer from bacteria.

3.3 Materials and methods

3.3.1 Gene family construction and discovery of novel genes

Proteome data from 99 species of Lepidoptera and 16 other arthropod species (Supplementary Table S1) were obtained from Ensembl Rapid Release (rapid.ensembl.org; accessed February 2023); taxon sampling was based on obtaining robust phylogenetic coverage across Lepidoptera while also preferentially selecting species with proteome predictions based on the Ensembl genebuild annotation pipeline (i.e. annotation which incorporated RNA sequence data). Primary transcripts were obtained from the predicted proteome data and OrthoFinder v2.3.14 was run with default parameters to determine orthogroups within the dataset (Emms and Kelly 2019). To relate these to a species tree, amino acid sequences from 25 single copy orthologues present in all species, as obtained from the OrthoFinder output, were aligned using MAFFT v7.505 (Katoh and Standley 2013), trimmed using trimAl v1.4.rev15 build (Capella-Gutiérrez et al. 2009), and concatenated with PhyKIT (Steenwyk et al. 2021). This concatenated alignment was used to generate a species tree using IQ-TREE version 2.0-rc1 with 1000 bootstrap iterations, the given model LG + G4 and option -nt AUTO which automatically determines the best number of cores given the current data and computer capacity (Minh et al. 2020). Orthogroups gained at nodes of interest (i.e. the branch leading to Lepidoptera and the branch leading to Ditrysia) were extracted using Orthoparser (github.com/PeterMulhair/ortho_parser). To test further

whether orthogroups inferred by the analysis to be specific to Lepidoptera were actually present in outgroups but missing from predicted proteomes, Trichoptera genomes annotated by the alternative Augustus-Gaius pipeline (BRAKER) (Gabriel et al. 2024) were analysed. This was carried out using a BLASTp search of the orthogroups against the trichopteran BRAKER proteomes to find any potential missing homologues (using an e-value cutoff of 1e-5 and filtering hits above 25% sequence identity match along with query and subject coverage of 60% to remove hits due to partial homology). Downstream of these steps, genes within orthogroups were analysed by exploring gene copy number, conducting synteny analyses, and generating expression matrices using publicly available RNAseq data. Figures including phylogenetic trees and heatmaps were generated in R using ggtree v3.6.2 (Yu et al. 2017), ggplot2 v3.4.4 (Wickham 2016), and Pheatmap v1.0.12 (Kolde Ravio 2025). Protein models were predicted using AlphaFold (ColabFold v1.5.5: AlphaFold2) (Jumper et al. 2021) and imported into Chimera v1.18 (Pettersen et al. 2004). Molecular graphics and analyses of protein models were performed with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311. Chromosome plots showing gene positions were created using RIdeogram v0.2.2 (Hao et al. 2020).

3.3.2 Phylogenetic analysis of gene families

Phylogenetic trees of the two gene families of interest were built by aligning deduced protein sequences using MAFFT v7.505 followed by trimming using trimAl v1.4.rev15 build and tree building using maximum likelihood in IQ-TREE version 2.0-rc1. Trees were visualised using ggtree v3.6.2 in Rstudio. In the sugar transporter orthogroup analyses, PfamScan (command line tool pfam_scan.pl) was used to search each orthogroup against the Pfam-A.hmm database with cutoff –cut_ga and an e-value threshold of 1e-3 (Finn et al. 2014) to annotate functional domains in each gene. This was used to detect additional gene families labelled as belonging to sugar transporters (possessing Pfam domain Sugar_tr; PF00083), followed by phylogenetic analysis including Drosophila and other arthropod SLC sequences to infer the class of SLC each orthogroup belonged to (Denecke et al. 2020). In the propeller protein analyses, putative HGT was investigated using a BLASTp search (e-value threshold of 1e-3) (Altschul et al. 1990) against the BLAST nr database with all lepidopteran sequences removed (Supplementary Table S3). The source of the HGT was then inferred by building a gene tree from the BLAST hits. Additional orthogroups in our datasets possessing the *propellin* gene were discovered by running a BLASTp search of the initial orthogroup (OG0000175) against all orthogroups in our dataset, retaining only those with percent identity equal to or above 25% and query and subject equal to or above 60%. This uncovered 8 additional homologous orthogroups, each of which contained only lepidopteran species. To further test the likely mode of origin of each of the 9 orthogroups, we carried out sequence similarity searches against the non-redundant protein sequence database (nr) and the core nucleotide database (core_nt) using a set of 10 representative species from each of the orthogroups (Supplementary Table S4). In one of the orthogroups

(OG0008135), two of the species had hits against genes/proteins belonging to other insects. To test whether these BLAST hits represented true homologs, or the result of spurious homology, we aligned both insect and *Spiroplasma* proteins to a *Manduca sexta* propellin protein. This was carried out using the RCSB pairwise structure alignment tool (Bittrich et al. 2024).

3.3.3 Gene expression quantification

RNAseq data for *Bombyx mori* were obtained from NCBI datasets PRJDB8614 and PRJNA675719 (Yokoi et al. 2021; Xu et al. 2022), for *Danaus plexippus* from PRJNA663267 (Ranz et al. 2021), and for *Papilio machaon* from PRJNA270386 (Li et al. 2015). RNA reads were trimmed using Trimmomatic v0.39 (Bolger et al. 2014), and mapped to the reference genome using STAR 2.7.10b (Dobin et al. 2013). Stringtie v2.2.1 was used to quantify expression in each of the species datasets (Pertea et al. 2015) and expression matrices were generated in RStudio using Pheatmap. Where multiple samples were available for a given tissue of lifestage, these were averaged to give one value.

3.3.4 Gene synteny analysis

Synteny analyses were used to test orthology of genes within and beyond Lepidoptera. For genes of interest, the gene ID, chromosome number, and location were determined from the genome annotation and gene track browser on Ensembl Rapid Release (Harrison et al. 2024). Two conserved ‘marker genes’ either side of the gene of interest were chosen and BLASTp searches (using Ensembl default parameters) conducted against the genomes of four

Lepidoptera (*Danaus plexippus*, *Papilio machaon*, *Tinea trinotella* and *Micropterix aruncella*) and eight outgroups (*Limnephilus lunatus*, *Limnephilus marmoratus, Limnephilus rhombicus*, *Glyphotaelius pellucidus*, *Bibio marci*, *Drosophila melanogaster*, *Adalia bipunctata* and *Vespula vulgaris*). These data were used to compare chromosomal organisation and gene neighbourhoods surrounding the genes of interest, revealing if individual genes within lepidopteran orthology groups were 1:1 homologues between species and also whether highly divergent orthologues were present in outgroups.

3.4 Results

3.4.1 Novel genes emerging at the base of Lepidoptera

To build a framework for comparative analyses, a phylogenetic tree was built from 25 single copy genes from 115 species, comprising 99 Lepidoptera species representing 24 families, and 16 outgroup taxa (Figure 3.1, Supplementary Table S1). The tree is broadly consistent with previously hypothesized evolutionary relationships, including placing the Micropterigidae family (*Micropterix aruncella* and *Neomicropterix facetella* in our dataset) sister to the rest of the lepidopteran lineages, the presence of the large, established groups of Ditrysia, Apoditrysia, and Macroheterocera (Kawahara et al. 2019), and recovering monophyletic groups for all taxonomic families in the dataset (Kawahara et al. 2019; Rota et al. 2022) (Figure 3.1).

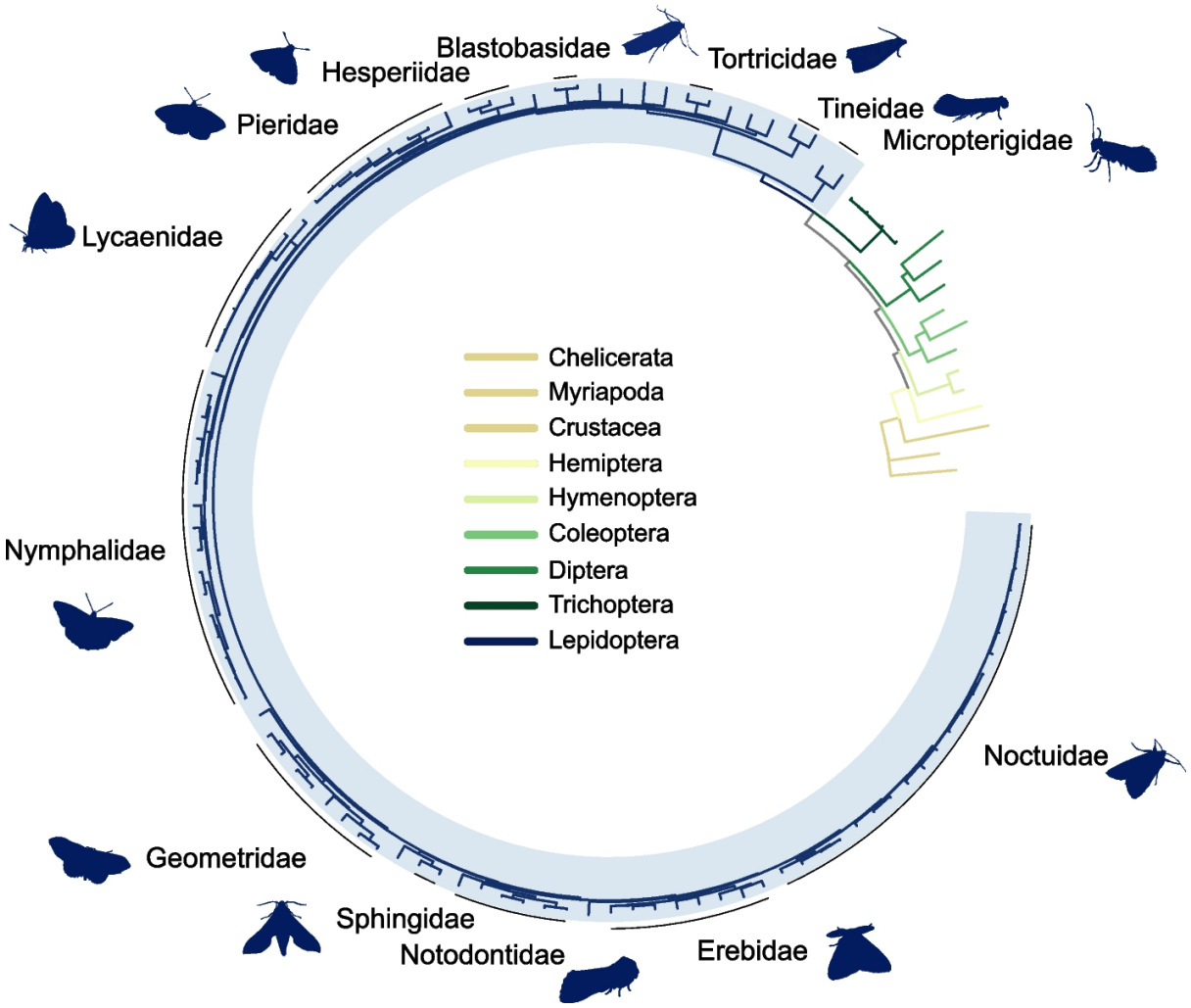


Figure 3.1 - Molecular phylogenetic tree of the 99 lepidopteran species from 24 families and 16 outgroup species inferred from 25 single-copy orthologues. Branches are coloured by insect order; species belonging to the named lepidopteran families are labelled with black lines on the outside of the tree

To identify novel genes or novel gene families that emerged early in lepidopteran evolution, we first constructed homologous gene groups ('orthogroups') using OrthoFinder (Emms and Kelly 2019). Novel gene families here are defined as orthogroups present in a clade but missing from all outgroup taxa i.e. taxonomically restricted genes. We filtered the complete set of orthogroups to only retain those present in greater than two species. To place each of these orthogroups onto the species tree, we took the parsimonious assumption that the common ancestor of all species present in each orthogroup represented the node of origin (Figure 3.2A). We identified 217 putative novel gene families originating on the branch leading to Lepidoptera (Figure 3.2A).

To assess the mode of origin for each gene family we applied Pfam annotations to search for protein domains (indicative of duplication and divergence from pre-existing genes) as well as carrying out sequence similarity searches against metazoan (excluding Lepidoptera; further suggestive of duplication) and non-metazoan sequences (suggestive of HGT) from the nr protein database. We deduce that the majority of novel gene families (177 orthogroups) which originated along the lepidopteran branch likely arose via duplication followed by extensive sequence divergence (Figure 3.2B). Putative HGTs accounted for only two orthogroups, as indicated by presence in Lepidoptera and non-metazoan proteomes but absent from animals other than Lepidoptera. We suggest that 38 orthogroups are potential orphan genes, candidates for origin by *de novo* gene genesis, although further analysis and additional data would be needed to test this hypothesis. We also detected 541 putative novel orthogroups on the branch leading to Ditrysia (representing all species outside of

Micropterigidae in our dataset) (Figure 3.2B). Of these, 398 likely arose from duplication, 13 via HGT, and 130 genes potentially originated *de novo*.

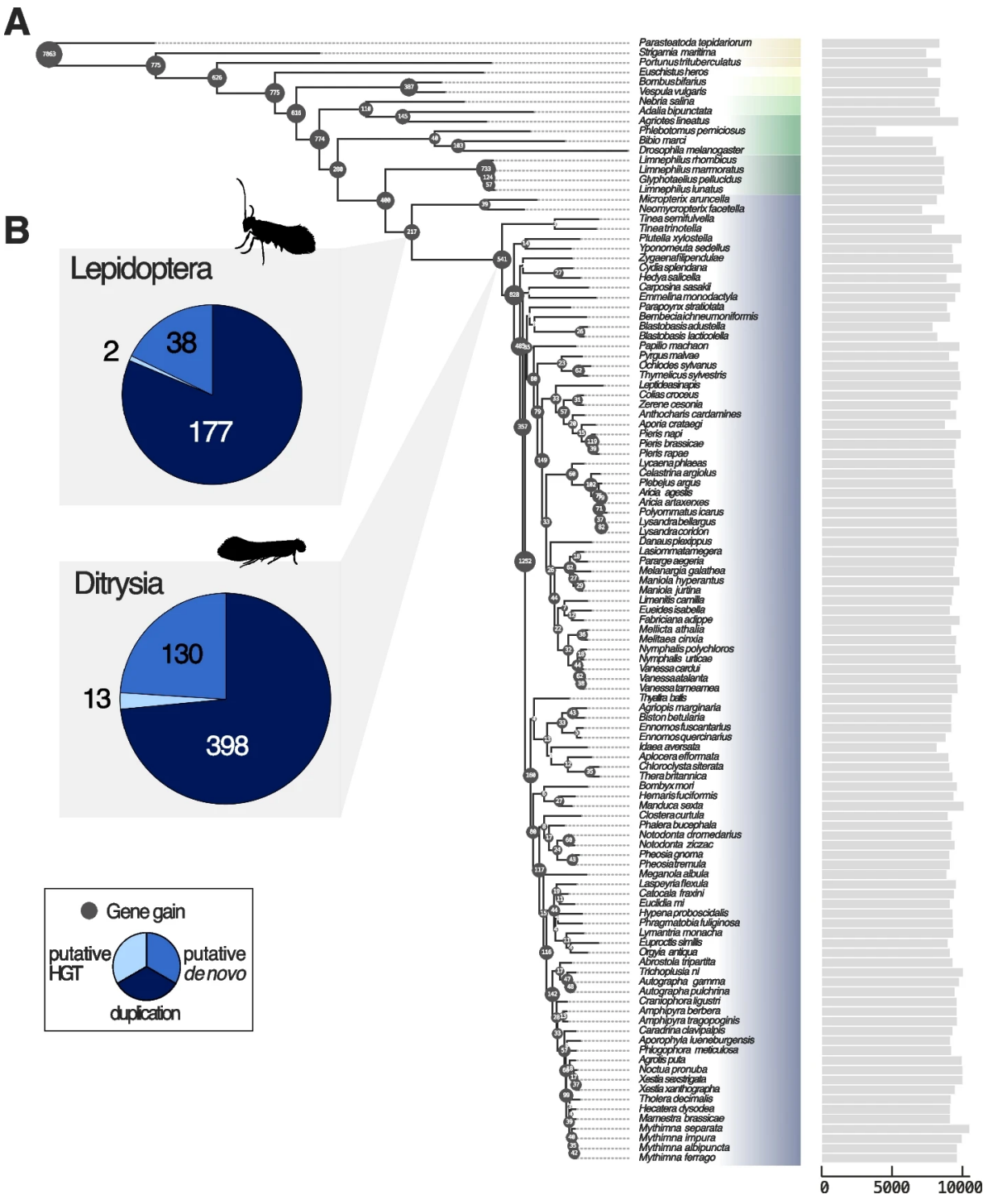


*Figure 3.2 - **A** Species tree showing numbers of orthogroups gained at each phylogenetic node. Insect orders are separated by colours. Bar chart to the right of the tree displays the total number of orthogroups identified in each species. **B** Pie charts show the number of orthogroups originating at the Lepidoptera and Ditrysia nodes and proportions of the putative modes of new gene origin*

We hypothesized that novel genes of particular importance to lepidopteran biology would be present in most species of Lepidoptera analysed, with little or no gene loss after gene emergence. Furthermore, some genes of functional importance may have undergone duplication and divergence since their emergence (Copley 2020). We therefore calculated gene copy number for every orthogroup originating at the Lepidoptera node and plotted these as a heatmap against a phylogenetic tree (3.3A). Approximately half of the 217 orthogroups showed a scattered phylogenetic distribution (present in a low number of species within Lepidoptera); these may represent genes that are frequently lost, or which underwent extensive sequence divergence within Lepidoptera complicating orthology assignment (right-hand columns in 3.3A). 87 orthogroups are present in 75% or more of the lepidopteran species in this dataset, with sporadic gene loss and occasional gene duplication on some internal branches (left-hand columns in Figure 3.3A).

To identify orthogroups with higher rates of duplication patterns, we first determined that the data does not follow a normal distribution (positive, non-symmetric, right skew) and is non-parametric (Anderson–Darling test, $p < 0.05$), and that at least one orthogroup has a gene copy distribution across species which differs from the mean number of gene copies per orthogroup per species (Kruskal–Wallis rank test, $p < 0.05$; mean number of gene copies = 1.1645). We found that two orthogroups deviate significantly from the mean number of gene copies within a given orthogroup gained at the lepidopteran node: OG0000164 and OG0000175 (Dunn test, $p < 0.0001$; Figure 3.3B). These two orthogroups have the highest variation in copy number, implying they have undergone extensive gene duplication within Lepidoptera, and they are also present in every lepidopteran species analysed. Sequence homology from BLASTp searches and domain annotation from Pfam

revealed that these proteins have a putative sugar transporter domain (OG0000164; MFS and Sugar/other transporter, PF00083.27, GO:0016020|GO:0022857|GO:0055085) and a 6-bladed beta propeller 3D structure (OG0000175; GO:0005515) (Figure 3.3C).

To determine whether there were any functions enriched in the full set of 217 orthogroups gained on the lepidopteran node, we analysed the functional domains of each to determine whether there were any categories which were significantly overrepresented. Although no functional categories were found to be enriched within this dataset, approximately 9% of the orthogroups (19 out of 217) were found to contain a zinc finger domain (Supplementary Table S2). We also discover that the Gloverin gene family (OG0000859) emerged on the branch leading to Lepidoptera (Figure 3.3C). The *gloverin* gene has previously been described as a lepidopteran novelty, and we confirm its emergence coincident with the evolution of Lepidoptera, where it has been retained in 86 of the 99 lepidopteran species in our dataset including *Micropterix aruncella* (3.3A). Gloverin, first purified from *Hyalophora gloveri* (Axén et al. 1997), is a glycine rich protein with no detectable homology outside of Lepidoptera. It functions as an antimicrobial peptide against a range of bacteria, with greater specificity to Gram-negative bacteria, and appears to be commonly and widely expressed across a range of life stages and tissues, with significant increases in expression observed following exposure to bacteria (Hwang and Kim 2011; Sparks et al. 2013).

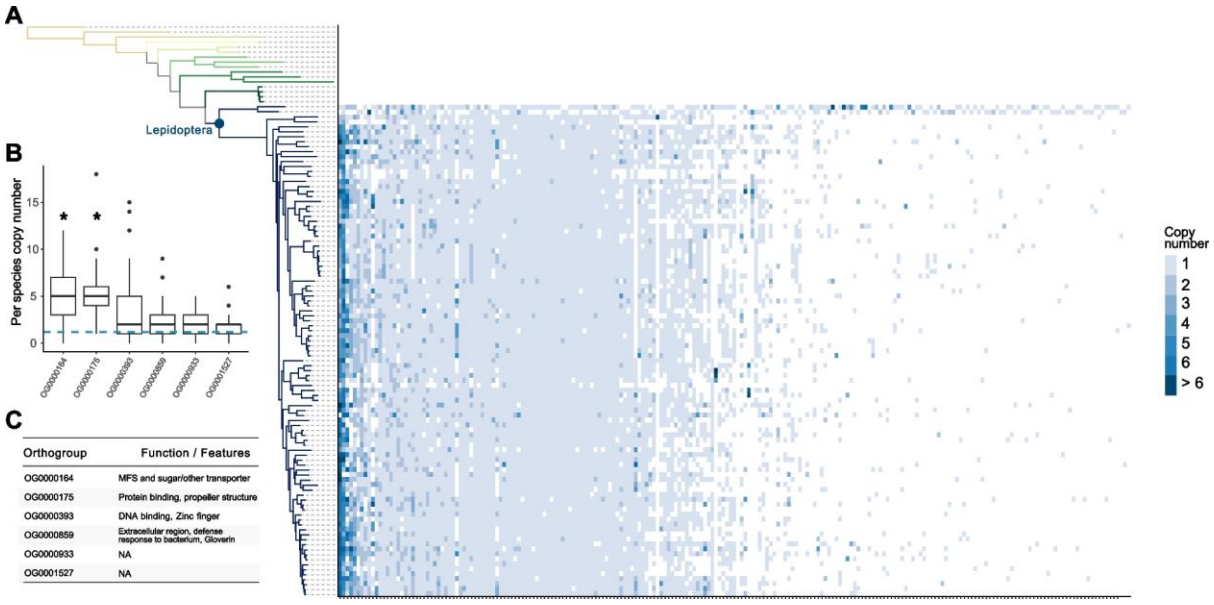


Figure 3.3 - Copy number of genes *gained on the ancestral node of Lepidoptera.* ***A*** *Heatmap (right) showing* gene *copy number for each orthogroup originating at the Lepidoptera node mapped to the species tree (left). Lepidoptera node is labelled with a blue circle. Orthogroups on the right-hand side of the figure have* genes *present in few species and may include spurious homologies.* ***B*** *Boxplots showing copy number variation per species in the top 6 orthogroups present in all or most lepidopteran species. Blue broken line signifies the mean copy number per species for all orthogroups. Orthogroups OG0000164 and OG0000175 have a mean copy number significantly different from the mean copy number of lepidopteran orthogroups, as signified by an asterisk ($p < 0.05$).* ***C*** *Table showing functions and features from six orthogroups deviating above the average copy number per orthogroup*

3.4.2 Gene expansion of lepidopteran sugar and solute transporters

The orthogroup originating on the node leading to the Lepidoptera with the highest mean copy number is a sugar transporter gene family (OG0000164) (Figure 3.3). Across the species analysed, the copy number for this lepidopteran-specific orthogroup ranged from one gene (*Micropterix aruncella*) to twelve genes (*Manduca sexta*). As the sugar transporter protein superfamily is large and diverse in animals (Denecke et al. 2020), and to understand the significance of this Lepidoptera-specific orthogroup, we extended our analysis to include all orthogroups containing a sugar transporter domain. We found 99 orthogroups with genes

possessing a sugar transporter domain present across all species in our dataset (3.4A), nine of which are annotated as emerging on the lepidopteran or ditrysian node; gene copy number for all nine orthogroups in each species shows varying rates of copy number expansion and gene loss (Figure 3.4A). Four of these orthogroups were single copy in all or most species, while five have undergone extensive gene duplication since their lepidopteran origins (Figure 3.4A and B).

Next, we wanted to determine whether these nine lepidopteran sugar transporter-containing orthogroups had a single evolutionary origin or whether they had evolved independently from separate ancestral sugar transporter genes. To resolve this, a phylogenetic tree of transporter proteins from representative lepidopteran and outgroup species was constructed using all sugar transporter orthogroups (Figure 3.4B). The tree topology suggests multiple origins of the lepidopteran- and ditrysian-specific sugar transporter genes, although not each of the nine orthogroups had independent origins. Notably, two lepidopteran-originating orthogroups (OG0000164 and OG0008208), and two ditrysian-originating orthogroups (OG0008344 and OG0000540) group close to each other in the phylogenetic tree (e–h). Another orthogroup (i; OG0008400) is located outside this clade, however each of these orthogroups are present in a larger clade consisting of members of the solute carrier 2 (SLC2) family (Figure 3.4B). In some taxa SLC2 genes have been shown to encode proteins that facilitate transport of small sugars across cell membranes (Denecke et al. 2020).

The remaining four sugar transporter orthogroups (a-d) are all ditrysian-specific, three of which form a single monophyletic group. The fourth orthogroup is located in a more

phylogenetically distinct group, however, all orthogroups belong to the SLC22 protein subfamily (Figure 3.4B). The SLC22 proteins are membrane transporters known to regulate metabolic functions, transporting a broader range of small molecules than SLC2 (Denecke et al. 2020). In all instances, the most closely related orthogroups in the gene tree contain both outgroups and lepidopteran species (Figure 3.4B). This implies that the lepidopteran- and ditrysian-specific transporter orthogroups originated from more ancient gene families that were present across all or most insects including Lepidoptera. These ancestral genes duplicated and underwent extensive amino acid substitutions specifically in the lineage leading to Lepidoptera or Ditrysia.

The four SLC2-like orthogroups [e–h] which group closely together in the gene tree (Figure 3.4B) are also co-located in the genome, found consistently in close association with one another across diverse lepidopteran species (3.4C). This suggests that these sugar transporter genes originated from a single ancestral duplication event at the base of the Lepidoptera and subsequently underwent tandem duplication in the ancestral lepidopteran and again in the branch leading to Ditrysia. In contrast, for the SLC22-like orthogroups gained on the ditrysian branch (a-d), we do not see the same close linkage and instead they are scattered on separate chromosomes (Supplementary Figure S1). If these originated from a single ancestral gene, as suggested by branching patterns in the gene tree, they dispersed around the genome after duplication.

To investigate possible functions of these lepidopteran-species transporter gene families, we assessed their patterns of expression across multiple time points and tissues from *Bombyx mori* (Yokoi et al. 2021; Xu et al. 2022). All genes from the nine lepidopteran and ditrysian-

specific gene families are expressed in at least one tissue or at one developmental time point (Figure 3.4D). The silk gland was the most frequent site of expression across the nine orthogroups, but some of the genes have wide and distributed expression (3.4D).

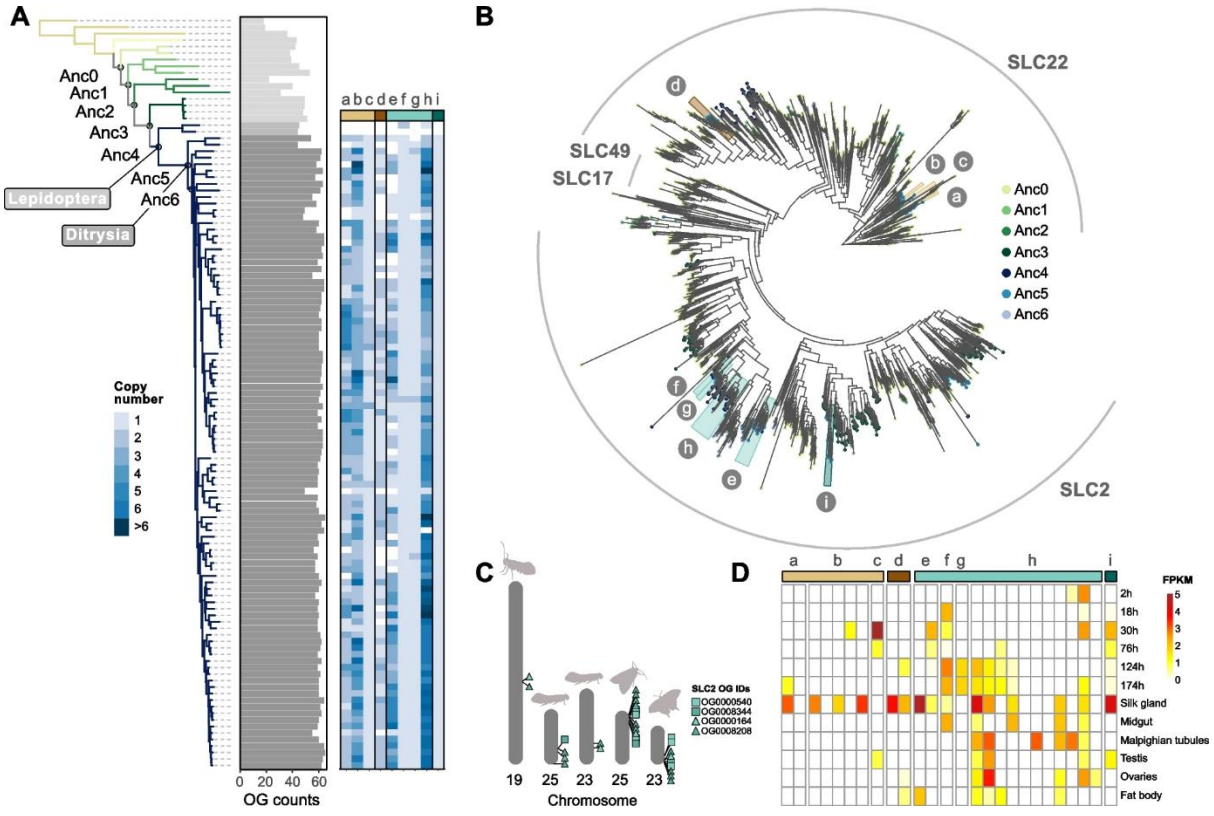


Figure 3.4 - Origins and evolution of lepidopteran and ditrysian-specific sugar transporter genes. Orthogroups are identified as follows: a—OG0000700, b—OG0000319, c—OG0007512, d—OG0001801, e—OG0000540, f—OG0008208, g—OG0008344, h—OG0000164, i—OG000840 ***A*** *Species tree on the left is coloured by taxonomic group, with the Lepidoptera and Ditrysia nodes labelled. The numbered ancestral nodes correlate to the node of origin for the orthogroups shown in the gene tree (part B). The grey bar chart (middle) shows that Lepidoptera (darker grey bars) have a higher total number of sugar transporter orthogroups compared to outgroup species. Copy number of lepidopteran and ditrysian-specific orthogroups varies greatly (heatmap, right): SLC22 transporters are below the light and dark brown bars (labelled a-d); SLC2 transporters are below the light and dark blue/green bars (labelled with letters e-i).* ***B*** *Phylogenetic tree built using a representative sample of outgroup sugar transporters, combined with sugar transporters identified in the Lepidoptera. SLC2 transporters are highlighted in light and dark blue/green along with letters e-i, while SLC22 transporters are highlighted in light and dark brown with letters a-d. Tip colours represent the node of origin (as shown in the species tree in part A) for each orthogroup.* ***C*** *The four closely related SLC2 transporters were mapped to a selection of lepidopteran chromosomes (left to right: Micropterix aruncella, Tinea trinotella, Tinea semifulvella, Papilio machaon and Autographa gamma). Sugar transporters of lepidopteran origin are represented by a triangle, while those of ditrysian origin are represented by a square. All four transporter orthogroups group in close physical proximity, on the same chromosome.* ***D*** *Heatmap of expression patterns of nine lepidopteran/ditrysian originating sugar transporters in Bombyx mori tissues*

3.4.3 Lepidoptera propellin genes arose through horizontal gene transfer

The second gene family which emerged at the base of Lepidoptera and is maintained in significantly high copy number across all butterfly and moth species analysed is orthogroup OG0000175 (Figure 3.3). The genes in this family were previously undescribed in insects. Below we show they encode proteins with a beta-propeller structure; we therefore name this the *propellin* gene family. Intriguingly, this group of genes is conserved across Lepidoptera yet does not have detectable sequence identity to any orthogroups in the arthropod outgroups included in our initial analysis. Furthermore, iterative BLAST searching revealed that OG000175 is not the only set of *propellin* genes in Lepidoptera; the genes are split into eight distinct orthogroups, including the original group OG000175 with the highest copy number. All eight propellin orthogroups are specific to Lepidoptera (Supplementary Figure S2). Combining all propellin orthogroups together, we find Lepidoptera genomes have an average of 11 gene copies, ranging from 3 copies in *Neofacetella micropterix* (Micropterigidae) to 25 copies in *Phragmatobia fuliginosa* (Erebidae) (Supplementary Table S5).

To assess the likely origin of the *propellin* gene in Lepidoptera, we carried out a BLASTp search using all proteins in orthogroup OG0000175 against the NCBI nr protein database excluding Lepidoptera sequences. This revealed significant sequence similarity matches to proteins from bacterial species. The most frequent bacterial genus in the set of matches was *Spiroplasma*, with additional matches in *Macrococcus* and *Escherichia* (Supplementary Table S3). Using iterative rounds of BLAST searching, we found very few matches outside bacteria; we identified a potentially related unnamed gene in the genome of the plant *Picea sitchensis*

(spruce; ABK22491.1) and a fungus gnat *Bradysia coprophila* (30% identity to a *Spiroplasma* homologue of Lepidoptera *propellin* genes over 16% query cover). To confirm the likely bacterial origins of the different *propellin* orthogroups, we also carried out BLASTp and tBLASTn searches against the nr and core nt databases, respectively, for each of the *propellin* orthogroups. We searched protein sequences from 10 representative species in each of the 8 orthogroups using both methods and found that bacterial, specifically *Spiroplasma*, sequences represented the majority of sequence similarity hits (Supplementary Table S4). While two genes from one orthogroup showed sequence similarity to other insect proteins (E3 ubiquitin ligases), we deduce that these hits are likely a result of spurious homology, with low query coverage (34–40%) and sequence similarity likely resulting from convergent amino acid residues in repetitive regions. In addition to this, all other hits from other species in the same orthogroup showed sequence similarity to *Spiroplasma* and other bacterial proteins. These *Spiroplasma* proteins were deduced to have similar tertiary structures to propellin (i.e. 6-bladed beta propeller; RMSD value of 3.73); in contrast, the spurious insect protein hits possessed multiple alpha helices and no structural similarity (RMSD value of 5.57).

Next, we constructed a phylogenetic tree of all *propellin* copies and their putatively homologous sequences. This shows that all lepidopteran *propellin* sequences are closely related in the gene tree (Figure 3.5A, Supplementary Figure S3). The most closely related branches to the Lepidoptera clade are *Spiroplasma* sequences which, along with a larger sister clade dominated by *Spiroplasma* sequences, suggests there has been a putative horizontal gene transfer from bacteria to Lepidoptera (Figure 3.5A). Based on the gene tree topology, we cannot exclude the possibility of multiple horizontal transfer events into

Lepidoptera. Although there are clear sequence similarity matches between Lepidoptera and *Spiroplasma*, the level of primary sequence identity is low. The highest percentage identity found between a lepidopteran protein and a *Spiroplasma* protein had only 35% identity over a sequence alignment of 134aa. This represents 45% coverage of a lepidopteran propellin protein (ENSAGMG00005008917.1) and 18% of a *Spiroplasma* protein (WP_164028422.1). To further assess legitimacy of the homologous relationships between these genes with low sequence identity, we predicted 3D structures of the deduced proteins from *Spiroplasma*, Macrococcus, and Lepidoptera (using six genes from Manduca sexta as representative of Lepidoptera) with AlphaFold (Jumper et al. 2021) (Figure 3.5B; Supplementary Figure S4). We find clear similarity in predicted protein structure with all lepidopteran and bacterial sequences having a 6-bladed propeller structure (Figure 3.5B). Further support for homology between lepidopteran *propellin* proteins and bacterial proteins was found when we aligned representative protein structures, which showed an RMSD value of 3.83 and TM-score of 0.69 (Supplementary Figure S5). Each structured propeller region within a propellin protein is approximately 221aa long consisting of blades of 30aa in length. There is variability outside of the beta-propeller domain including regions of varying length and structure, most notably in the additional domains in the *Spiroplasma* protein model (Figure 3.5B). To reflect this protein structure, we name the Lepidoptera genes the *propellin* gene family.

To further investigate the evolution of *propellin* genes, we focussed attention on the high-copy number propellin orthogroup (OG0000175; Supplementary Figure S2). Phylogenetic analysis divides this orthogroup into six clades within Lepidoptera, which we refer to as gene subfamilies (Figure 3.5C). The early diverging lineage of Lepidoptera, represented by the family Micropterigidae, has a *propellin* gene in clade 1 in the gene tree (Figure 3.5C). Next,

we examined the local gene synteny for these six subfamilies across representative lepidopteran species. Genes from each subfamily, excluding clade four, were found in a microsyntenic cluster of genes ('marker genes') which are homologous across most or all species (Figure 3.5C). This confirms that each of these 6 *propellin* subfamilies are one-to-one orthologues across Lepidoptera. Importantly, many of the marker genes also exist in microsyntenic blocks in the arthropod outgroups, consistent with these being the genomic sites where the Lepidoptera-specific *propellin* gene was integrated (Figure 3.5C). Since the six *propellin* subfamilies are at distinct chromosomal locations, yet form a monophyletic group in molecular phylogenetic analysis, we propose that this orthogroup emerged through a single HGT event from *Spiroplasma* or another bacterial source to Lepidoptera, followed by duplication and transposition around the genome. These duplications generated not only the six subfamilies analysed in detail, but also likely the additional *propellin* genes referred to above. We note that genes in subfamily 1 are intronless (or have one intron), while the remaining subfamilies and additional *propellin* orthogroups have between 0 and 8 introns, with the median count being 1 intron. This could reflect transposition via an RNA intermediate or could be a legacy of the gene's bacterial origin (Supplementary Table S5).

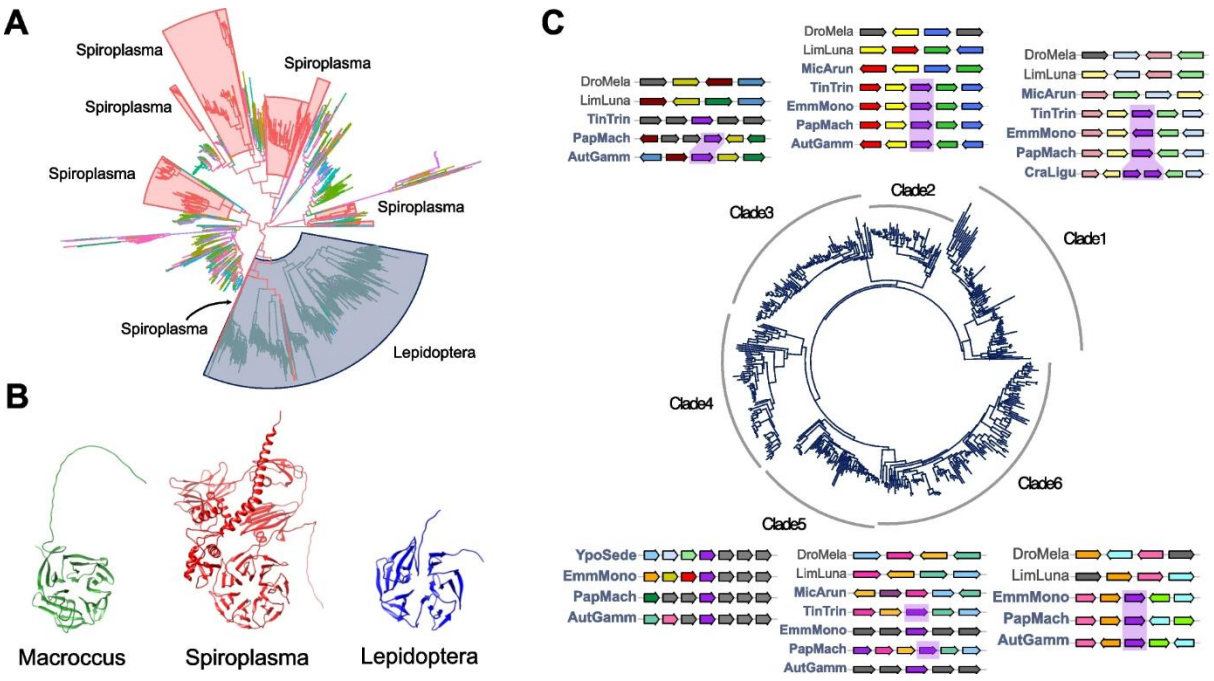


Figure 3.5 - Lepidoptera-specific genes encoding proteins with sequence identity and structural similarity to bacterial 6-bladed propeller proteins. ***A*** *Gene tree of propellin and putative bacterial homologs. The Lepidoptera clade (blue) and Spiroplasma clades (red) are labelled with coloured boxes and text. All other branches represent a range of bacterial species which are shown in Supplementary Figure S3. Molecular phylogenetic analysis indicates the propellin genes of Lepidoptera are monophyletic, whose most closely branching lineages are Spiroplasma genes, and sister group to a clade dominated by Spiroplasma genes (highlighted in red).* ***B*** *AlphaFold predictions suggest lepidopteran propellin proteins form 6-bladed propeller structures similar to bacterial homologues; examples shown from Macrococcus (green), Spiroplasma (red) and the lepidopteran M. sexta (blue). Additional protein structure predictions in Supplementary Figure S4.* ***C*** *Molecular phylogenetic analysis indicates that the largest orthogroup of lepidopteran propellin genes divides into 6 clades, each gene (purple) located at a different chromosomal location, most of which show conserved synteny between lepidopteran species (synteny indicated by shaded purple regions). The Micropterigidae species M. aruncella only has a gene in clades 1. Marker genes are shown by various colours*

For a first insight into the possible functional role of the lepidopteran *propellin* genes, we analysed the expression of all copies of *propellin* using transcriptomic data sets from three species: *Bombyx mori*, *Danaus plexippus*, and *Papilio machaon* (Supplementary Figure S6). While we find evidence for expression of all gene copies in each species, the patterns are complex and variable within and between species. In *Danaus plexippus* for example, while most *propellin* copies show some expression in larval or pupal stages (8 of the 11 genes), levels of expression are highest in the adult life stage, with particularly high expression found

in the thorax, compared to the head or abdomen (Supplementary Figure S6). In *Papilio machaon* most copies are restricted to one or two life stages, while others are strongly expressed throughout the life cycle of the butterfly. Expression in *Bombyx mori* shows wider coverage across life stages and tissue types, with most gene copies expressed in early developmental and adult life stages. While expression is common across most adult tissue types in *Bombyx mori*, there is little, or no expression found in the midgut or silk glands (Supplementary Figure S6). While there is little correlation in expression between homologous copies of *propellin* across all three species, we note that transcriptomic datasets available are not comprehensive. However, such pervasive expression across life stages and tissues in multiple species provides support to the fact that these genes are functional across a wide range of lepidopteran species.

We noted above that there were some sequence similarity matches outside bacteria and Lepidoptera. The putatively homologous gene from Diptera is an uncharacterised locus (LOC119081672, encoding putative protein XP_037046651) on an unplaced scaffold in the genome assembly of a fungus gnat *Bradysia coprophila* (Urban et al. 2021). We find this gene is present in two species of *Bradysia*. It is unlikely that the fungus gnat scaffold is a contaminating sequence since it is present in two species, and because it is adjacent in the genomes to recognisable insect genes (Supplementary Figure S6). Analysis of the unplaced scaffold reveals clearly dipteran genes immediately 3’ (LOC119081668) and relatively close 5’ (LOC119081673 and LOC119081675) to the gene of interest. Intriguingly, a locus immediately 5’ (LOC119081585) has high similarity to springtail (*Collembola*) tyrosine kinases, and the next neighbouring gene (LOC119081673) is *Bradysia*-specific (Supplementary Figure S6). We therefore suggest the Diptera gene LOC119081672 arose by

an independent HGT from *Spiroplasma* in the *Bradysia* fungus gnat genus, which has likely also acquired other genes by HGT. We have not deduced the origin of the loci with a sequence match in *Picea sitchensis* (spruce).

3.5 Discussion

In this study we identified 217 'novel' genes arising on the evolutionary lineage leading Lepidoptera, after it had diverged from outgroups including the closest related order Trichoptera (caddisflies). We caution, however, against this as a quantitative measure of genomic novelty. First, we are using a pragmatic definition of novelty that includes *de novo* genes, horizontally transferred genes, and gene duplication followed by sequence divergence; altering parameters relating to sequence divergence could increase or decrease the gene count (Weisman et al. 2020). To improve inference of new genes in early lepidopteran evolution, we employed a phylogenetically informed approach to construct gene families, minimising the effects of bias resulting from rapid sequence divergence (Emms and Kelly 2019). Second, novelty at the Lepidoptera node could be 'undercounted' if some genome annotations are incomplete, particularly those of early diverging lepidopteran taxa. Third, there are factors that could spuriously 'overcount' novelty. For example, in our study around half the novel orthogroups were found sporadically in a small number of distantly related Lepidoptera species. This could indicate repeated gene loss following the origin of the novel gene but could also include 'noise' as a result of some proteins being grouped incorrectly due to spurious sequence identity. Secondary loss of genes from caddisfly genomes could theoretically cause overcounting of genes on the Lepidoptera node, but we have minimized this risk through use of four caddisfly genomes. We also noted a

small degree of overcounting (< 2%) emerging from alternative genome annotation methods (Weisman et al. 2022), but we accounted for this (see Methods). Specifically, the initial input data consisted of proteomes predicted from the Ensembl Genebuild annotation which incorporates RNA sequence data and filters poorly supported potential coding transcript proteins. A second method of genome annotation, the BRAKER method, is potentially less stringent and found some genes that had been missed by Genebuild. The difference amounted to just two orthogroups. The same caveats apply to counts of novel genes at other similarly deep phylogenetic nodes. Despite this caveat, we find it interesting that even more apparent gene novelty (541 gene families) dates to the node leading to Ditrysia. These genes require further analysis, but the observation suggests that the evolution of new biological traits continued during the early evolutionary radiation of moths. More important than an absolute number of novel genes, the analysis gives us a first look into the relative importance of different modes of gene origin during the emergence of Lepidoptera. We find the majority of novel gene families gained on the ancestral lepidopteran branch arose via gene duplication and divergence (~ 82%) while around ~ 18% genes had no sequence matches or any recognisable domains. These are putative candidates for genes arising *de novo* from non-coding genes. Just two genes (< 1%) are candidates for having arisen via HGT (including the *propellin* gene), with hits to bacterial or fungal species.

One of the genes that likely arose via HGT was highlighted in our analysis as a novel gene that underwent extensive gene duplication in Lepidoptera to generate a large gene family. This gene family, which we name the *propellin* genes, is potentially functional as evidenced by the extensive retention through evolution and conserved domain structure. Currently, however, its precise role in lepidopteran biology is unclear. Phylogenetic analyses suggest

that the progenitor of the *propellin* gene family was transferred to an insect from *Spiroplasma* bacteria, some time on the Lepidoptera stem lineage. *Spiroplasma* is a well-known intracellular symbiont in arthropods. Furthermore, *Spiroplasma* is known to colonise reproductive tissues, which in turn impacts upon the host's reproduction, and indeed this genus is one of two bacterial symbionts in Lepidoptera for which maternal transmission has been demonstrated (Duplouy and Hornett 2018). In some cases, transmission is enhanced by manipulation of host physiology, such as male-killing which increases the number of female offspring as observed in *Danaus chrysippus* (Jiggins et al. 2000). Clearly, persistent association with reproductive tissues gives opportunity for horizontal gene transfer, as the symbiont DNA is in close physical proximity to the DNA of the host germline. This has been seen in the relationship between a mealybug and two endosymbiont species *Tremblaya* and *Moranella* (Husnik and McCutcheon 2018). Interestingly, we also found a putatively homologous gene in two species of Diptera (genus *Bradysia*), possibly reflecting an independent HGT event. This is consistent with previous findings that some types of gene are more prone to HGT than others, perhaps those encoding proteins with few interaction partners (Cohen et al. 2011). The evolutionary retention of the likely HGT-derived *propellin* gene, plus its extensive gene duplication in Lepidoptera, suggest this gene family likely evolved to perform functions that are important for the biology of moths and butterflies. We do not know the biological role, or roles, of *propellin* genes in Lepidoptera, but note that their bacterial homologues have diverse functions including ligand-binding proteins, signalling proteins, lysases, structural proteins, isomerases and hydrolases (Chen et al. 2011). It is worth noting that the *Spiroplasma* genes which group closest to the *propellin* genes in the gene tree are annotated as hypothetical proteins without known function, suggesting more work is needed to understand the functional context of this gene.

The only other novel lepidopteran gene to show such widespread retention and extensive gene duplication encodes a family of SLC2-like sugar transporter proteins. In other animals, members of the SLC2 sugar transporter superfamily encode glucose-uptake proteins, ribose transport proteins, and several putative membrane proteins probably involved in sugar transport (Fiegler et al. 1999; Denecke et al. 2020; Ioannidis et al. 2022). The functional link to sugars is particularly intriguing since the ecological association between Lepidoptera and sugar-feeding changed markedly in early lepidopteran evolution. Specifically, adult moths in the basal family Micropterigidae primitively lack a proboscis and are pollen feeders, whereas adult moths and butterflies in the Ditrysia use a proboscis to access sugar-rich nectar in flowers. Our wider comparative survey of sugar transporter gene families picks up potentially interesting co-evolution between this ecological shift and the sugar transporter genes. We find that although OG0000164 (and one other sugar transporter gene family) are present in pollen-feeding Micropterigidae, it is not until the evolution of the nectar-feeding Ditrysia that we see extensive gene duplication, widespread gene retention and the emergence of additional SLC-like sugar transporter gene families (Denecke et al. 2020). We suggest, therefore, that novel sugar transporter gene families emerged at the base of Lepidoptera, but it was only later in lepidopteran evolution that massive gene duplication and functional divergence of sugar-transporter genes took place, in association with nectar feeding. The causal link between these genetic changes and the evolution of novel feeding behaviour in the early evolution of Lepidoptera warrants further study.

3.6 Conclusion

We have demonstrated the emergence of 217 novel gene families (orthogroups) on the node leading to Lepidoptera and 541 novel gene families emerging on the node leading to the Ditrysia. Two orthogroups have significantly higher gene copy per species across Lepidoptera indicative of extensive gene duplication following their origins. One likely originated by horizontal gene transfer from the endosymbiont bacterium *Spiroplasma* and then duplicated to generate a diverse group of '*propellin*' genes encoding a 6-bladed propeller domain. The other encodes a large set of sugar transporter proteins and is part of a diverse set of sugar and solute transporter genes that duplicated and diverged extensively in early lepidopteran evolution.

3.7 Data availability

Genome data associated with this study is listed in Supplementary Table S1 along with accession numbers. Data and code generated in this study can be found on figshare; figshare.com/s/32d1b9055257dad1892f.

Chapter 4: Gene expression in the reduced first thoracic legs of a Nymphalid butterfly

Asia E. Hoile , Peter O. Mulhair , Peter W.H. Holland

Introduction to manuscript

This chapter has been submitted as a first author original research article to the Journal Of *Insect Molecular Biology* and is currently under review. The manuscript is reproduced here with minor changes as permitted as part of an integrated thesis. The authors credited on this manuscript, in order, are Asia E. Hoile, Peter O. Mulhair, Peter W. H. Holland* (* - corresponding author).

Author contributions

P.W.H.H. and A.E.H. conceived the study. A.E.H. designed analyses and carried out the bioinformatic and laboratory research presented. A.E.H., and P.O.M. interpreted all results. A.E.H. wrote the initial draft of the manuscript, and P.W.H.H. and P.O.M. edited versions. All authors read and approved the final manuscript.

4.1 Abstract

Nymphalid butterflies have unique leg morphology amongst Lepidoptera: they are the only family with greatly reduced forelegs (T1) in adults of both sexes, which are not used for walking. Previous studies have suggested that T1 legs may have chemosensory functions. To investigate which genes underpin this biology, we undertook a differential gene expression analysis in female *Maniola jurtina*. We find that nymphalid T1 legs have a distinct transcriptomic profile to T2 and T3 legs, and also to palps; sensory appendages in Lepidoptera. We find over 250 genes commonly expressed in nymphalid T1 legs and palps, but not in T2 legs. Despite this overlap, T1 legs are still more similar to T2 and T3 legs in their gene expression profiles than they are to palps. Genes expressed in common between palps and T1 legs include one encoding a trypsin-domain protein descendent from a nymphalid-specific duplication and several involved in sensory functions including an extraocular blue opsin. Our findings indicate clear transcriptomic differences between T1 legs and walking legs in *Maniola jurtina*, pointing to the functional basis of these differences, including minor acquisition of palp-like gene expression.

4.2 Introduction

Although the majority of adult Lepidoptera typically walk on all six of their legs, two taxonomic families are characterised by greatly reduced forelegs (T1 or prothoracic legs), meaning they only walk using the legs of the second and third thoracic segments (T2 and T3 respectively) (Wolfe et al. 2011). These two families, Nymphalidae and Riodinidae, likely evolved reduced forelimbs in parallel, and it is notable that reduced T1 legs are observed in both sexes in Nymphalidae, but only in male riodinids (Wolfe et al. 2011). Despite this, the T1 legs in nymphalid females retain all five tarsal segments, while in males the post-tarsus is never present (Fox 1966).

The Riodinidae are a sister lineage to the Lycaenidae, and together these lineages are sister to Nymphalidae (Wolfe et al. 2011; Kawahara et al. 2023). A slight reduction has been observed in other butterfly species, and the degree of reduction of T1 legs and the differences in reduction between sexes have traditionally been among the main characteristics used to define major butterfly taxa (Fox 1966). Hesperidae and Papilionoidea have minimal reduction, with T1 legs being only slightly shorter than T2 and T3 legs. Lycaenidae have forelegs distinctly smaller than T2 and T3 legs, however all three families use all three sets of legs for walking (Fox 1966).

It has been suggested that the T1 legs of nymphalids evolved novel sensory functions, subsequently losing their walking function (Wolfe et al. 2011). For example, tarsal sensillae on T1 are implicated in host plant recognition prior to oviposition (Calvert and Hanson 1983; Baur et al. 1998) and the physiological reactions of sensillae to plants differ between T1 and

T2 or T3. Furthermore, nymphalid T1 legs demonstrate a negative reaction to sugar solution, whereas a positive reaction is observed from T2 and T3 legs (Fox 1966). This difference is not observed in the Pieridae family where both T1 and T2 legs demonstrate a positive reaction to a sugar solution (Fox 1966). It is relevant that the number of putative sugar transport genes expanded throughout the evolution of Lepidoptera, giving potential for specialisation of roles (Hoile et al. 2025). Similarly, previous work found a link between gene duplication in key sensory genes, the gustatory receptors responsible primarily for taste perception, and the localised expression of these paralogs in the legs of *Heliconius* butterflies (Briscoe et al. 2013). This leg specific expression was more pronounced in females, which, accompanied by sexual dimorphism in the number of gustatory sensilla present on the T1 legs, is thought to play an important role in oviposition behaviour (Renwick and Chew 1994).

Legs are not the only appendages with sensory roles in insects. Palps are small, segmented appendages near the mouthparts of insects (Scoble 1995; Kristensen and Skalski 1998). In Lepidoptera, the term ‘palps’ typically refer to labial palps, paired structures attached to the labium and hence part of the last (most posterior) head segment. Lepidopteran labial palps have roles in chemoreception, in tactile sensing to guide feeding behaviour and, in some cases, protection of the coiled proboscis (Myers 1969; Diiak et al. 2023). Maxillary palps are typically reduced or vestigial in Lepidoptera due to the evolution of a proboscis which is specialised for nectar feeding; an exception being the families Micropterigidae, Agathiphagidae and Heterobathmiidae which lack the proboscis (Krenn 2010; Diiak et al. 2023).

Arthropod legs and labial palps are considered to be serially homologous structures, meaning they have evolved from a common ancestral appendage type (Panganiban et al. 1997; Hughes and Kaufman 2002). Consistent with proposed serial homology, genes necessary for appendage development, and typically expressed in arthropod legs, are also expressed in palps. For example, *in situ* hybridisation revealed that the *Distal-less* (*Dll*) gene is expressed in a 'sock-and-ring' pattern in the developing walking legs of the butterfly *Precis coenia*, and in a subtly different 'sock-only' pattern in developing labial palps (Panganiban et al. 1997). Gene expression similarities are observed in other arthropods, including expression of *Distal-less* (*Dll*), *homothorax* (*hth*) and *dachshund* (*dac*) in *Porcellio scaber* (Crustacea) and *Steatoda triangulosa* (Chelicerata) (Hughes and Kaufman 2002; Chen et al. 2016). Studies in *Drosophila melanogaster* have identified Hox genes as responsible for the specification of these different head structures; in particular the *proboscipedia* (*pb*) and *Sex combs reduced* (*Scr*) genes (Aplin and Kaufman 1997; Hughes and Kaufman 2002).

The above studies have demonstrated shared evolutionary origin and some degree of developmental similarity between legs and palps, but they clearly evolved into very distinct structures in insects: in essence, legs are primarily for walking and palps are for sensing. In this context, the dramatic size reduction of T1 legs in nymphalid butterflies is intriguing, they are no longer used for walking and there is some evidence for a novel sensory role. However, morphologically they do not resemble palps, and appear instead as small leg-like structures suspended between the T1 segment. We wished to investigate whether nymphalid T1 legs resemble T2 and T3 legs (waking legs) in gene expression profile, despite their morphological reduction, or whether they express a distinctive set of genes. We also ask whether nymphalid T1 legs have similarity in transcriptomic expression to labial palps in nymphalids,

thereby testing if they have co-opted biochemical or physiological characteristics of palps. Finally, we asked whether any transcriptomic expression differences between T1 and walking legs were related to genes that had arisen through genetic novelty originating during nymphalid evolution; for example whether new genes have evolved specifically for roles in the T1 legs. We chose to focus on adults of *Maniola jurtina* (Meadow Brown), an abundant European grass-feeding butterfly (family Nymphalidae, subfamily Satyrinae).

4.3 Experimental procedures

4.3.1 Sample collection and RNA sequencing

33 female and 18 male *Maniola jurtina* (Meadow Brown) individuals were collected between 05 July 2024 and 22 August 2024 at Wytham Woods, Oxford (UK grid reference SP 468085). Specimens were stored at -70 °C and dissected on dry ice whilst frozen. Palps, T1, T2 and T3 legs were removed and pooled with the corresponding tissue of between 6 and 15 other individuals, however it is worth noting that not all individuals collected were utilised for RNAseq (Hoile et al. 2025b: Supplementary 1). RNA was extracted from pooled samples using the RNAeasy micro kit (Qiagen) and eluted in 20µl of ultrapure water (Hoile et al. 2025b: Supplementary 1). RNA sequencing was performed on replicates using the Illumina NovaSeq X Plus Series (PE150) Sequencing System (Novogene) at a coverage of 20M paired-end 150bp reads per sample.

4.3.2 Differential gene expression analyses

Reads were mapped to the *Maniola jurtina* genome (Lohse and Weir 2021) using STAR version STAR_2.4.0g1 (Dobin et al. 2013). Differential gene expression analysis was conducted using the DESeq2 (version 1.38.3) (Love et al. 2014) package in RStudio accounting for batch effects within the data. The dataset was filtered to only include genes which had a base mean > 5 , $\log_2$ fold change > 1 and adjusted p-values < 0.05. The base mean is defined as the average of the normalized FPKM count values, divided by size factors, taken over all samples. Each of the four tissues were compared to the other tissues to identify genes significantly differentially expressed (upregulated or downregulated). Plots were created using RStudio packages as follows: heatmaps of the top 50 genes with the highest and top lowest fold change were generated using ComplexHeatmap 2.14.0 (Gu 2022); Volcano plots and PCA plot were generated using ggplot 3.4.4 (Wickham 2016); the Venn diagram was created using venneuler 1.1-4 (Gao et al. 2021). Expression analysis of candidate sensory genes used FPKM values averaged across replicates; these were then normalized as Z-scores.

4.3.3 Discovery of novel genes

Proteome data from 51 species (17 Nymphalidae, 33 other Lepidoptera, 1 Trichoptera) were obtained from Ensembl Rapid Release http://rapid.ensembl.org (accessed August 2024) (Hoile et al. 2025b: Supplementary 1). Taxon sampling was designed to achieve robust phylogenetic coverage across Lepidoptera while also preferentially selecting species with proteome predictions-based RNA sequence data. Proteomes were filtered to retain the

longest transcript for each gene and OrthoFinder v2.3.14 was run to determine orthogroups within the dataset (Emms and Kelly 2019). To relate these to a species tree, amino acid sequences from 1,528 single copy orthologues from the OrthoFinder output and present in all species, were aligned using MAFFT v7.505 (Katoh and Standley 2013), trimmed using trimAl v1.4.rev15 build (Capella-Gutiérrez et al. 2009), and all alignments were then concatenated with PhyKIT (Steenwyk et al. 2021). The concatenated alignment was used to generate a species tree using IQ-TREE version 2.0-rc1 (Minh et al. 2020) using 1000 bootstrap iterations, the LG + G4 model and option -nt AUTO which automatically determines the best number of cores given the current data and computer capacity. Orthogroups gained on the branch leading to the Nymphalidae were extracted using Orthoparser (github.com/PeterMulhair/ortho_parser).

Genes within orthogroups were analysed to generate expression matrices, explore gene copy number, and conduct synteny analyses. Figures including phylogenetic trees and heatmaps generated in R used ggtree (Yu et al. 2017), ggplot2 (Wickham 2016), and Pheatmap (Kolde Ravio 2025), and were edited using Affinity Designer 2 (Serif 2024). *Maniola jurtina* chemosensory genes were annotated using a python script (Hoile et al. 2025b: Supplementary 2) which extracted functional domains using Pfam (Mistry et al. 2021) from the proteomes of all species to annotate and classify proteins which chemosensory function.

4.4 Results

We hypothesized that any divergent transcriptomic profiles underpinning the difference observed in nymphalid T1 legs may result from either novel genes arising on the node leading to Nymphalidae or from shifts in expression of preexisting genes due to co-option of new regulatory networks. To investigate this we divided our analyses into three main parts: (i) a comparison of T1 to T2 and T3 legs, (ii) a comparison of T1 to palps, and (iii) a detailed investigation into the molecular basis of expression similarities arising between T1 and palps. We also assessed the contribution of novel genes versus existing genes which have gained distinct expression profiles coincident with the evolution of the reduced T1 legs. It should be noted that the differential gene expression analysis was only conducted in adult females, and no other life stages were investigated. Only genes which were upregulated in both T1 legs and palps in comparison to T2 legs were investigated, however it should be acknowledged that genes downregulated in these two tissues may also be important.

4.4.1 Transcriptomic differences between T1 legs and walking legs

To obtain a dataset suitable for conducting a differential gene expression analysis, 51 *Maniola jurtina* butterflies were collected, dissected to remove palps, T1, T2 and T3 legs respectively (Figure 4.1A), and pooled to produce 3 replicates consisting of between 6 and 15 individuals, with 15 males pooled to form the single male sample described in Supplementary 1. We focused on samples consisting of thirty-three female *M. jurtina* specimens for differential gene expression analyses, conducted using DESeq2 with the percentage of uniquely mapped reads ranging between 50-74% across samples (Hoile et al.

2025: Supplementary 1). Gene expression quantification was measured with the gene models of *M. jurtina* using StringTie (Pertea et al. 2015). First, to determine whether there were broad expression profile differences between tissues, we constructed a principal component analysis (PCA) plot from these gene expression quantifications of palps, T1, T2 and T3 legs. Despite a consistent batch effect, the analysis reveals that palp gene expression has the greatest dissimilarity to the other tissue types, while T2 and T3 legs have the smallest difference between them, indicated by the close grouping of these data points (Figure 4.1B). T1 legs do not cluster with other tissues but do show more apparent similarity to T2 and T3 legs than to palps. This suggests that T1 legs have a gene expression profile which is not typically found in palps or in the other legs (Figure 4.1B).

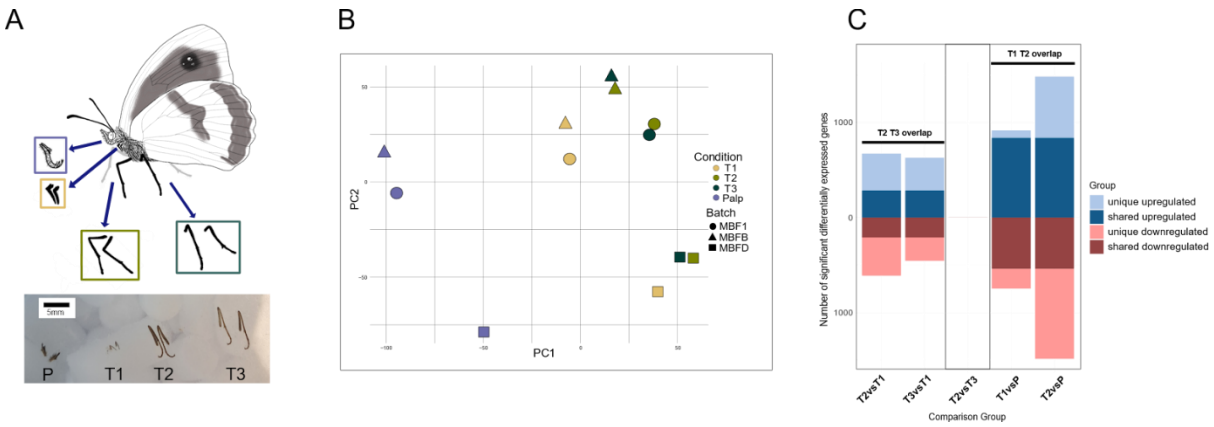


*Figure 4.1 – **A** Maniola jurtina illustration demonstrating dissection of palps and T1, T2 and T3 legs. Scale bar 5mm. **B** PCA analysis of transcriptomes from T1, T2 and T3 legs, and palps. **C** (Left) Number of genes significantly 'upregulated' (higher expression) and 'downregulated' (lower expression) between, from left to right: T2 vs T1, T3 vs T1, T2 vs T3, T1 vs palps, T2 vs palps. The darker shading shows the number of upregulated and downregulated genes in common between T2 vs T1 and T3 vs T1 (left hand bars), or in common between T1 vs palps and T2 vs palps (right hand bars). Note that gene expression is essentially identical between T2 and T3, that gene expression is T1 legs is very different to T2 and T3, T1 gene expression is more similar to palps than T2 gene expression is to palps, and T1 gene expression is more similar to other legs than it is to palps.*

4.4.2 Nymphalid reduced legs and walking legs have distinct gene expression

To quantify the differences in gene expression between tissues, pairwise comparisons were conducted across all four tissue types (P, T1, T2 and T3), evaluating each dataset against the others in a like-for-like manner (e.g. T1 vs T2 etc). Differentially expressed genes underwent filtering to only include genes which had a base mean > 5 , log2 fold change > 1 and adjusted p-values < 0.05 (Hoile et al. 2025b: Supplementary 3).

When comparing T2 walking legs against T3 walking legs, only one gene was found to be differentially expressed and has putatively lower expression in T2 in comparison to T3 (ENSMJUG00000000611, an olfactory receptor) (Figure 4.1C). This indicates there is very little, if any, transcriptomic difference between T2 and T3 legs. We therefore used T2 as a representative of the walking legs in further analyses. When comparing the reduced legs of T1 vs the walking legs of T2, however, 791 differentially expressed genes were identified. Of these 402 showed higher expression in T1 and 389 had lower expression in T1. This indicates a clear transcriptomic difference in the reduced legs compared with the walking legs (Figure 4.1B and C).

4.4.3 Larger differences in expression profile observed between legs and palps

In a comparison of T1 legs vs palps, we found 1,718 differentially expressed genes. Of these, 972 showed higher expression in T1 and 746 had lower expression in T1. In a comparison of palps vs T2 legs, we found 2,864 differentially expressed genes. Of these, 1,380 had higher expression in the palps and 1,484 had lower expression in the palps. This suggests that palps

have a very different transcriptomic profile to legs, but the difference is greatest when compared to walking legs. In addition, although the reduced legs of T1 are transcriptomically very different to the rear legs (T2 and T3), all 'legs' are more transcriptomically similar to each other than to palps (Figure 4.1C).

To investigate the nature of the gene expression differences between the reduced legs (T1) and walking legs (T2, T3), we first searched for gene expression differences unique to T1 legs. Of the genes differentially expressed between T1 vs palps and T1 vs T2, 106 genes were shared across both analyses (Figure 4.1C). These represent gene expression differences unique to T1 legs. Of these, 47 were more highly expressed in T1 and 59 were down-regulated in T1 in both analyses. This indicates that of the suite of genes commonly expressed in lepidopteran appendages, many are either up-regulated or down-regulated in T1 legs. T1 legs may have acquired biochemical functions different to either of the other analysed tissues (Hoile et al. 2025b: Supplementary 3).

4.4.4 Nymphalid reduced legs have acquired aspects of palp-like gene expression

Although there are large gene expression differences between palps and legs, it is possible that the reduced legs of Nymphalidae acquired some palp-like characters. To test for this, we searched for genes up-regulated in palps (compared to T2 legs) that are also up-regulated in T1 (compared to T2 legs). We found 277 such 'palp' genes expressed in T1 legs (Figure 4.2A). This indicates that T1 legs may have acquired some of the biochemical functions typically associated with palps. To delve deeper into the *M. jurtina* genes identified, Pfam annotation and a BLAST search was conducted against the *Drosophila melanogaster* proteome (Hoile et

al. 2025: Supplementary 4) to examine their putative functions. A nested pie chart was generated to visualise the overall functional category, combined with specific function for each of the genes identified in the 277 T1/palp up-regulated genes (Figure 4.2B). The majority of genes were classed as metabolic (21.3%), while a large proportion of genes were uncharacterised or lncRNA (13% and 11% respectively). Genes in the sensory and developmental functional categories each comprised only 1% of these genes (Figure 4.2B).

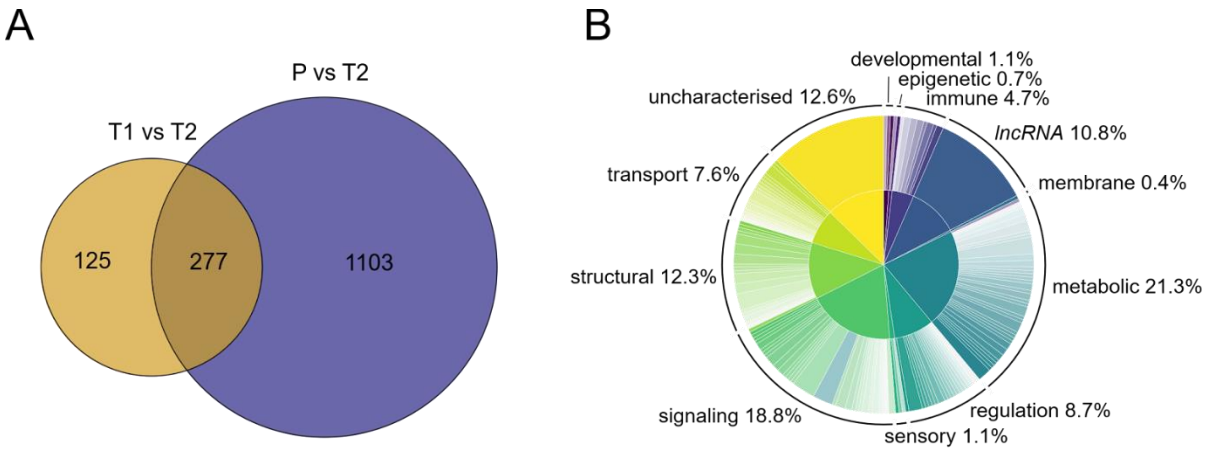


*Figure 4.2 - Genes of interest explored from the 277 genes upregulated in both T1 legs and palps **A** Common upregulated genes between T1 and Palps showing a crossover of 277 genes common and upregulated in both T1 legs and palps. **B** Functional category and in-depth functions of the 277 genes which were upregulated in both T1 legs and palps.*

As a complementary approach, we also examined the expression of predicted sensory genes in the RNAseq datasets to ensure physiologically relevant genes which did not meet the statistical threshold for differential expression were not overlooked. We examined genes encoding proteins involved in chemosensation, notably olfactory binding proteins (OBP), chemosensory proteins (CSP), sensory neuron membrane proteins (SNMP), ionotropic receptors (IR), gustatory receptors (GR) and olfactory receptors (OR), and also visual sense,

notably visual opsins and genes in the visual sensing pathway. For each gene family, we identified homologues in the *M. jurtina* proteome and calculated FPKM values of gene expression in palps, T1, T2, and T3 legs. Expression values were averaged across female replicates for all tissues and Z-score calculated to allow for comparison across tissues for each gene (Figure 4.3B). In total we extracted 60 OR genes, 28 GR, 32 IR, 29 OBP, 18 CSP, and 15 SNMP chemosensory-related genes from the *M. jurtina* proteome. Of the genes analysed, we found 40 genes with a positive Z-score in palps suggesting they have a specialised sensory function in palps only. We also found 32 sensory-related genes with a positive Z-score only in T1 legs; these could represent sensory functions unique to the reduced legs. Finally, we found 16 sensory-related genes with a positive Z-score in both palps and T1 legs, suggestive of a shared sensory function. Notably, this includes a gene encoding blue-sensitive opsin with strongest expression in T1, suggesting extraocular opsin activity in T1 legs. We also note expression in palps and T1 legs of a putative diacylglycerol kinase gene (retinal degeneration A), possibly encoding a component of the signalling pathway downstream of opsin activation.

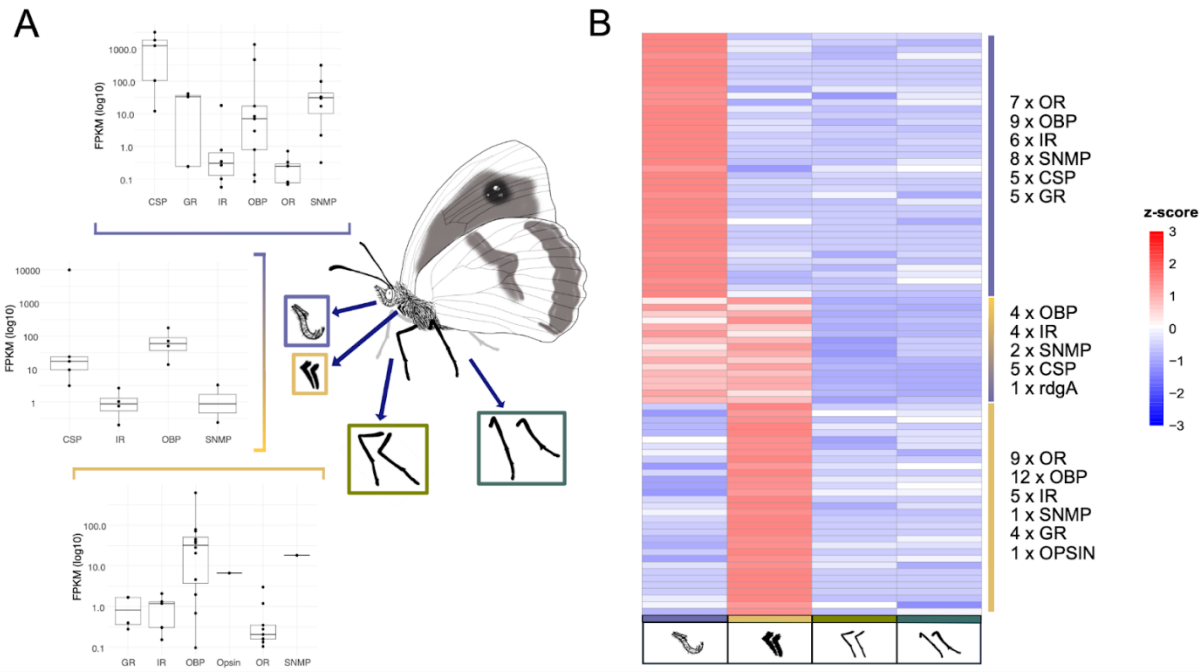


*Figure 4.3 – **A** Putative sensory genes upregulated in palps (purple), T1 legs (beige) or both tissues (purple to beige gradient). FPKM values (in log10) for each gene type are shown as boxplots per tissue sample. **B** Z-score expression plots of sensory genes of interest; only shown are genes with higher expression in T1, palps, or both.*

4.4.5 A Nymphalid-specific gene duplication up-regulated in palps and T1 legs

For each of the 277 genes identified as up-regulated in both T1 legs and palps (and subsequently down-regulated in T2 and T3 legs), gene copy number was investigated to determine whether any showed variation indicative of gene duplication events during the evolution of the Nymphalidae lineage (Hoile et al. 2025: Supplementary 5). We found 8 orthogroups had duplicated, or were only identified in the Nymphalidae subfamily Satyrinae plus *Danaus plexippus* i.e. not present in any species outside this clade. One orthogroup was notable as present, typically in single copy, in nearly all species used in this analysis (all Lepidoptera species and Trichoptera outgroup), but with clear patterns of nymphalid-specific

gene duplication events (OG0000217) (Figure 4.4A and B). The lowest gene copy number in the Nymphalidae is observed in *Eueides isabella* (1 copy) with the highest found in *Bicyclus anynana* (11 copies) and a mean of 6 gene copies across all nymphalid species. Phylogenetic analysis suggests that within the Nymphalidae, duplications of this gene generated four clades (gene subfamilies), with some subfamilies undergoing further duplication (Figure 4.4A).

Although eight *Maniola jurtina* genes were identified within this orthogroup, only one of these genes was present in the up-regulated group of 277 L1/palp-expressed genes (ENSMJUG00000013800) (Figure 4.4C). Six genes were expressed in some or all of the tissues investigated (ENSMJUG00000012716, ENSMJUG00000014532, ENSMJUG00000014814 ENSMJUG00000013800, ENSMJUG00000015669 and ENSMJUG00000004577). Two gene copies have a higher expression in the palps (ENSMJUG00000015669 and ENSMJUG00000004577), which had an average FPKM value of 35 and 38 respectively in palps and 21 and 17 respectively across all leg tissues, while the remaining three genes had relatively consistent expression patterns across all four tissue types (Figure 4.4C). This suggests the paralogous gene copies have diverged in their expression domains. For all *M. jurtina* sequences in this orthogroups, a BLASTp search against *Drosophila melanogaster* resulted in hits to uncharacterised proteins, with the top hit for ENSMJUG00000013800 being CAL85485 (gene ID: CG9649) with a percentage identity of 29.930 and e-value of $8.69e^{-27}$. All proteins were annotated as possessing a trypsin domain (Pfam domain PF00089.29), placing these genes as part of the S1 family of peptidases (Rawlings and Barrett 1994).

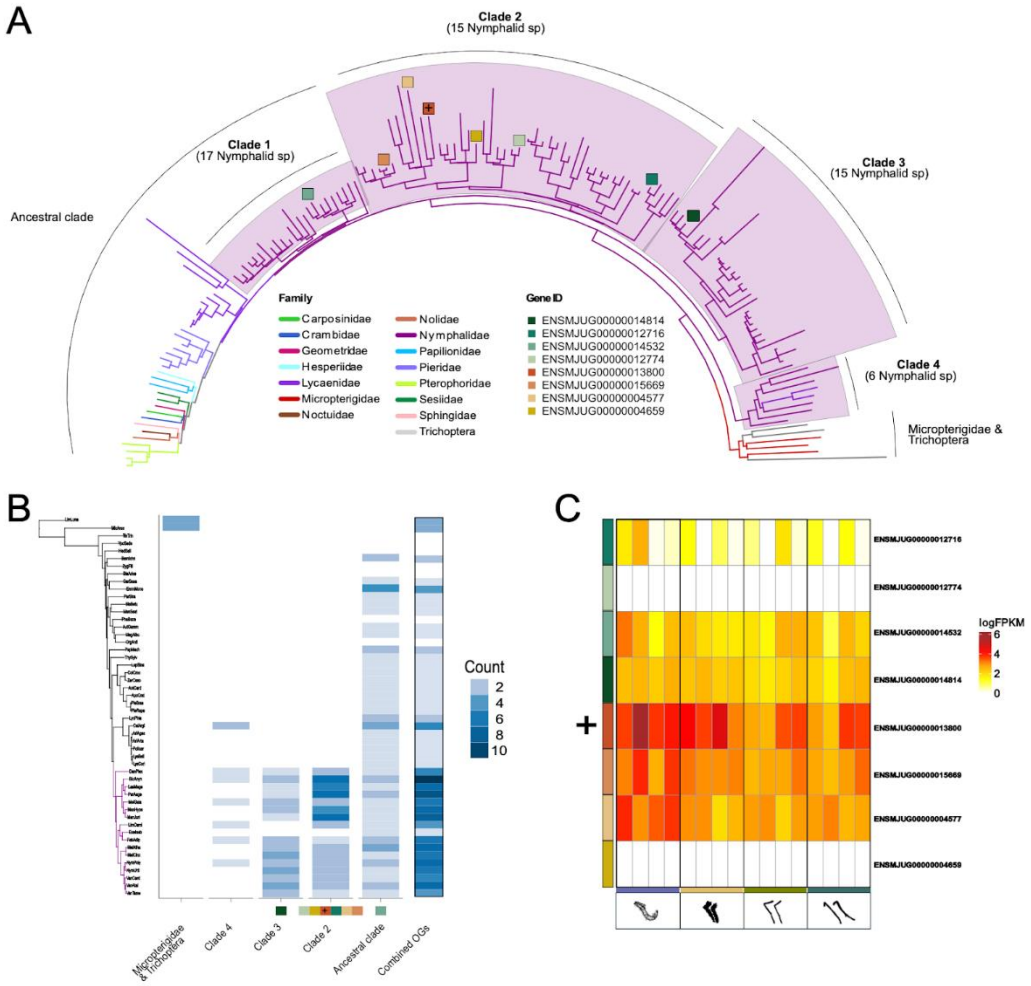


Figure 4.4 – In-depth analysis of a nymphalid-specific gene duplication from a gene upregulated in both T1 legs and palps. ***A*** *Phylogenetic tree (coloured by lepidopteran family) of the orthogroups containing putative Nymphalid-specific gene duplication. The specific gene copy which is upregulated in Maniola jurtina T1 legs and palps is denoted by + . The number of unique Nymphalid species per clade is noted.* ***B*** *Gene count for each species used in this analysis in the orthogroups of interest, split into clades with overall gene copy in the orthogroups on the right hand side.* ***C*** *logFPKM expression across T1, T2 and T3 legs, and palps for each of the Maniola jurtina genes in these orthogroups of interest. The gene which is upregulated in T1 legs and palps is denoted by +. Only three of these gene copies demonstrate expression in any of palps, T1, T2 or T3.*

4.4.6 Novel genes do not underpin transcriptomic differences in T1 legs

To investigate whether any transcriptomic diversity observed in the T1 legs arose from novel genes arising on the node leading to the Nymphalidae, we first built a phylogenetic tree using 1,528 single copy genes across 51 species onto which gene gain could be mapped. The species selected were representative of 21 lepidopteran families and one non-lepidopteran insect (*Limnephilus lunatus*, order Trichoptera) to give strong taxonomic coverage across Lepidoptera.

The analyses above included some examples of pre-existing genes in Lepidoptera (trypsin peptidases and chemosensory genes) showing patterns of duplication and/or differential gene expression in the palps and T1 legs of *Maniola jurtina*. To assess whether additional novel genes or novel gene families emerged in Nymphalidae we first constructed homologous gene groups ('orthogroups') from all 51 species in our dataset using OrthoFinder (Emms & Kelly, 2019). Novel gene families here are defined as orthogroups present in a clade but absent from all outgroup taxa i.e. taxonomically restricted genes (Hoile, et al., 2025). We identified 57 gene families originating on the branch leading to the Nymphalidae (Figure 4.5). Interestingly, even larger numbers of gene gain events are observed on nodes subsequent to the emergence of Nymphalidae: 104 genes originating on the node containing non-Danainae nymphalids and 129 genes on the branch leading to the subfamily Satyrinae.

Of the 57 Nymphalidae-specific orthogroups, we deduce that the majority (35 orthogroups) likely arose via duplication followed by extensive sequence divergence. Putative HGTs account for only three orthogroups, as indicated by presence in non-metazoan proteomes but absent from animals other than nymphalid butterflies. We suggest that 19 orthogroups are orphan genes, genes which have either emerged by de novo gene genesis or have diverged in their amino acid content to such an extent to not be detectable by standard sequence identity searches. None of the 57 genes were determined as differentially expressed in the palps or T1 legs compared to T2 and T3 legs (Hoile et al. 2025b: Supplementary 6). This suggests that while new genes did arise during the evolution of Nymphalidae, it is unlikely they made a major contribution to the novel biology of nymphalid T1 legs.

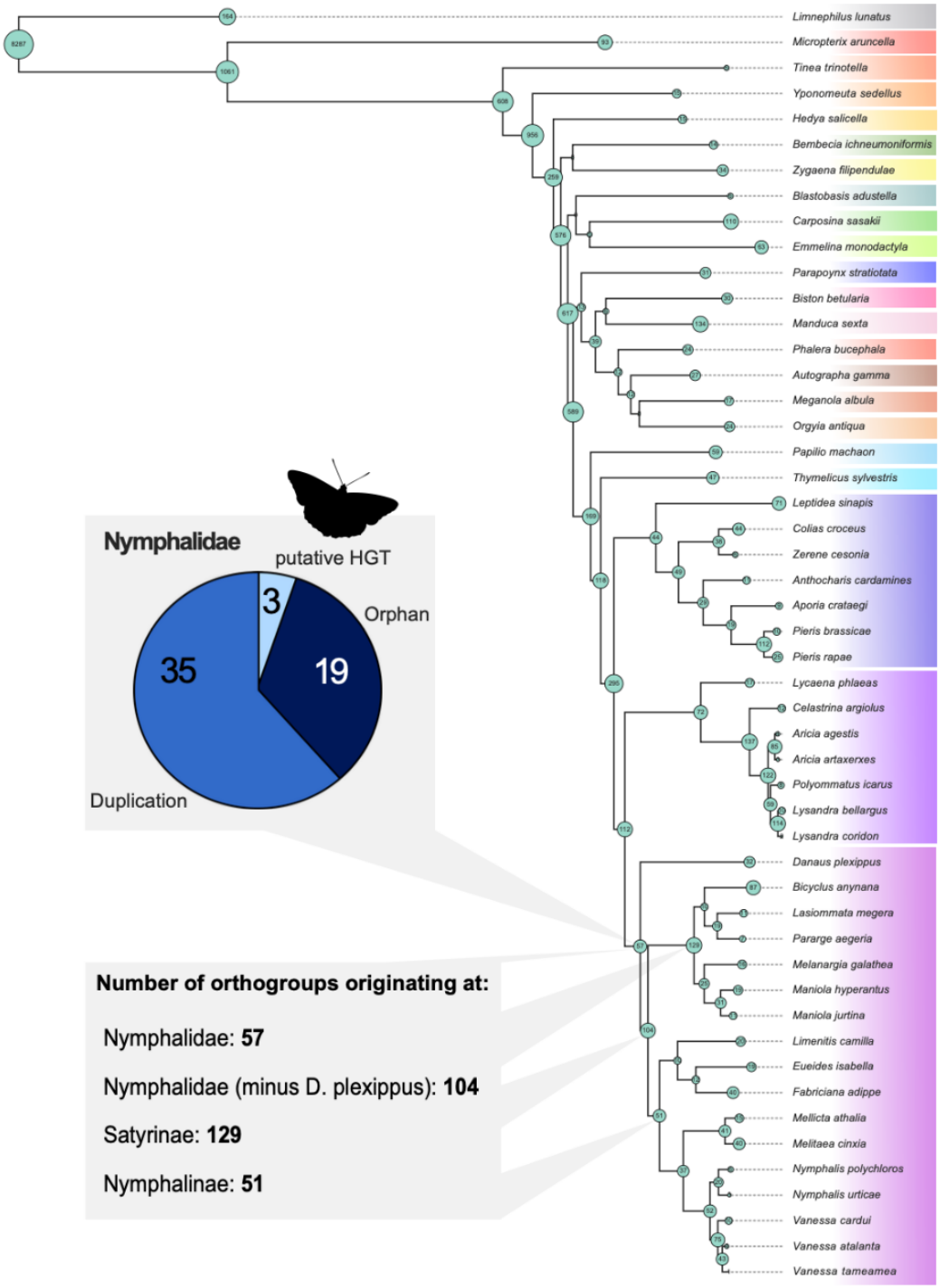


Figure 4.5 - Genes emerging on the Nymphalid butterfly node. Species tree showing numbers of orthogroups gained at each phylogenetic node. Lepidopteran families are labelled by coloured boxes. Pie charts show the number of orthogroups originating at the Nymphalidae, Nymphalidae minus D. plexippus, Satyrinae and Nymphalinae nodes and proportions of the putative modes of new gene origin.

4.5 Discussion

Previous research has established serial homology between legs and labial palps, with evidence from similar developmental gene expression and homeotic transformations. This serial homology extends to the legs of the T1 segment. However, serial homologues can diverge in structure and function in evolution, presumably underpinned by gene expression differences (Monteiro, 2021; Monteiro et al., 2025). While there has been research assessing differences in gene expression between sensory tissues, including legs, in butterflies (Briscoe et al., 2013; van Schooten et al., 2020; Wu et al., 2022) we know of no prior research that has examined the transcriptomic similarities and differences between the highly diverged T1 legs of nymphalid butterflies and their serial homologues.

We first established that there are very little, if any, transcriptomic differences between T2 and T3 legs. Only one gene was identified as differentially expressed between T2 and T3 legs, emphasising a high degree of transcriptomic similarity. Similarity was also observed in the PCA analysis, demonstrating a low proportion of variance between T2 and T3 (Figure 4.1B). Although a higher proportion of variance was observed between T1 and walking legs (T2 and T3), PCA analysis concluded that T1 legs have a higher degree of transcriptomic similarity to walking legs than they do to palps (Figure 4.1B). This is also supported by a greater number of differentially expressed genes observed between T1 vs palps in comparison to T1 vs T2 (Figure 4.1C). This confirms the hypothesis that T1 legs bear some degree of genetic differences to walking legs (T2 and T3) (Figure 4.1C).

Although T1 legs share a closer gene expression profile to the walking legs than they do to palps, they have still acquired some gene expression patterns similar to that of the palps distinct from that found in walking legs (Figure 4.1B and C). We find 277 genes which are significantly up-regulated in both T1 legs and palps, indicating that T1 may have acquired some of the expression and, by consequence, function typically associated with the palps. Additionally, T1 legs have some genes which are not upregulated in any of palps, T2 or T3 legs (Figure 4.1C and Supplementary 1). We also identified over 80 sensory-related genes unique to either palps, T1 legs or both tissues, although not all of these were identified in the differential gene expression analysis (Figure 4.3). This suggests that, along with diverging in their expression profiles where they are closer to palps than the walking legs are, T1 legs may have also acquired expression and potential functions not found in either of these three tissues; exploring the specific function of these in the context of T1 leg biology or physiology will be an interesting area to explore in future research.

Having established that T1 legs have acquired some transcriptomic similarity to palps, we were interested to know if some of the transcriptomic differences characterising reduced T1 legs could be related to evolution of new genes or expanded gene families. From the genes which were found to be up-regulated in T1 and palps, when compared to walking legs, eight orthogroups to which some of these genes belong were present in *Danaus plexippus* and the Nymphalidae subfamily Satyrinae only. Interestingly, the Satyrinae have the highest number of novel genes emerging on their node of origin (129) in comparison to the subfamily Nymphalinae (51) and the Nymphalidae family itself (57) (Figure 4.5). Research suggests that Satyrinae often exhibit greater host plant specialisation in comparison to other subfamilies within Nymphalidae e.g. Nymphalinae (Nylin et al., 2014). An example of this is many

Satyrinae predominantly utilising plants from the order Poales (grasses), while Nymphalinae typically have a broader range of host plants across multiple plant orders. As such, it is possible that the orthogroups which only contain Satyrinae and Danainae genes assist in this greater host plant specificity (Nylin et al., 2014). One orthogroup was identified as having genes present in almost every species used in the study (average of one gene copy per non-Nymphalidae species), with a marked increase in copy number occurring in 16 out of 17 Nymphalid species in our dataset (OG0000217) (Figure 4.4A and B). We annotated this gene family as consisting of trypsin-related proteins based on their functional domain content. Trypsins are a family of serine protease enzymes which play a role in protein digestion by breaking down polypeptide chains into smaller peptides and amino acids. They are typically involved in digestion but may have other roles (Brenner, 1988; Rawlings & Barrett, 1994). Literature investigating trypsin-like gene expression in any arthropod legs is very limited, however several serine proteases and protease inhibitors have been identified in the crustacean olfactory organ, and it suggested that they may play a role in perireception (e.g., odour activation or inactivation) or in the development or survival of olfactory receptor neurons (Johns et al., 2004). The diverse expression pattern observed for *M. jurtina* genes in this orthogroup suggests that duplication and divergence may have taken place within this orthogroup and therefore the genes may have adopted new functions in nymphalid butterflies. It is possible that these trypsin domain-containing genes may play a role in sensory perception in Nymphalidae T1 legs and palps, however further research is required to test this hypothesis.

As palps are known to have some sensory function, and we found that T1 legs have some transcriptomic similarity to palps, we also explored the potential role of sensory genes in nymphalid T1 legs. Over 30 sensory genes showed highest expression in T1 legs compared to palps and walking legs, with the majority of these being olfactory binding proteins and olfactory receptors (Figure 4.3B). This gene set also included extraocular expression of blue opsin. Non-visual and extraocular opsin expression has been observed in *Drosophila melanogaster*, crustaceans, cephalopods and fish. Retinal non-visual receptors and associated opsins have been identified across vertebrate tissues and are thought to utilise the phototransduction pathway to detect light for non-visual purposes (Feuda et al., 2022; Kingston & Cronin, 2016). Extraocular colour sensing has also been observed in *Biston betularia* (peppered moth) caterpillars, allowing them to choose to rest on colour-matching twigs for camouflage, even when ocelli are obscured (Eacock et al., 2019). Additionally, there is a potential role of extraocular photoreceptors in colour change, resulting from pigment production in response to background colour across multiple species in the animal kingdom (Poulton, 1890). This suggests that T1 legs have acquired some sensory function which is not observed in walking legs.

Additionally, 16 sensory genes were most highly expressed in T1 legs and palps, implying that T1 legs have acquired some sensory function also observed in palps (Figure 4.3B). Notably, this includes a retinal degeneration A or diacylglycerol kinase gene (retinal is a chromophore which binds to opsins and can be involved in visual transduction (Wald, 1934). Diacylglycerol (DAG) kinase in *Drosophila melanogaster* is thought to have a role in the development of the embryonic nervous system and function of the adult nervous system, muscle and regulation of signal transduction in neurons (Harden et al., 1993; Masai et al.,

1992). It is possible that the opsin and retinal degeneration A genes identified are being utilised for a similar purpose, perhaps relating to light sensing in the T1 legs of *Maniola jurtina*.

Previous research has identified gustatory and olfactory receptors as mediators of insect-plant reactions through identification of secondary plant compounds as deterrents or attractants of insect oviposition and feeding, however the exact role of reduced T1 legs in this is not yet clear (Baur et al., 1998; Calvert & Hanson, 1983; Fox, 1966; Renwick & Chew, 1994; Wolfe et al., 2011). It is possible however, that specialised receptors on the T1 legs, which may be unique among specific Nymphalidae species, may allow for host-specific interactions with certain plants.

We therefore conclude that T1 legs in Nymphalidae butterflies have sensory function, and that some transcriptomic similarity is observed between T1 legs and palps. However, it is notable that T1 legs still bear more gene expression similarity to walking legs than they do to palps.

4.6 Data Availability

RNAseq data generated in this study are deposited under NCBI BioProject PRJNA1303389. All code and processed data required to reproduce results are deposited on the Figshare open research repository http://figshare.com/s/2b41aaf69f613a990bff (Hoile et al. 2025b).

4.7 Acknowledgements

A.E.H. was supported by the Oxford Interdisciplinary DTP and funding from the Biotechnology and Biological Sciences Research Council (UKRI-BBSRC), grant number BB/T008784/1. P.O.M. was supported by a BBSRC Fellowship, grant number UKRI893. We thank Yuanzhen Zhu for his advice on Z-scores for comparison of tissue expression. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Chapter 5: Functional enrichment and gene novelty in bilaterian-specific tissues

5.1 Abstract

The Cambrian explosion was a period during which the bilaterian animals experienced a rapid diversification, marking a significant burst of evolutionary activity for this lineage. To understand the genetic mechanisms underpinning such a significant event, we identify and assess gene gain as a driver of tissue-specific bilaterian evolution. We undertake this investigation by utilising two phylogenetic tree topologies: the former placing the Xenacoelomorpha as sister to the Nephrozoa (Nephrozoa hypothesis) and the latter placing Xenacoelomorpha within Deuterostomia (Xenambulacraria hypothesis). Publicly available RNAseq data was obtained for *Octopus bimaculoides* and *Branchiostoma lanceolatum*, representing the Protostomia and Deuterostomia respectively, and used to determine enrichment of Bilateria-specific hierarchical orthogroups (HOGs) in gut, nervous system and muscle tissue. Results were supported using RNAseq data from four further bilaterian species. We focussed on 13 HOGs that emerged on the bilaterian node of the tree containing Xenacoelomorpha within Deuterostomia. In the tree topology supporting the Nephrozoa hypothesis, 9 of these HOGs were observed on the node leading to Nephrozoa. It is likely that 9/13 HOGs arose by gene duplication followed by extensive sequence divergence and 4/13 have no known homology outside of Bilateria and possibly arose via de novo gene genesis. We provide an insight into the role of gene novelty in a tissue-specific context during the early evolution of the Bilateria, suggest their likely mode of origin, their putative function and importance in bilaterian evolution.

5.2 Introduction

The emergence and diversification of the Bilateria represents a critical evolutionary milestone in the history of life on Earth. Bilateria form a major clade within the animal kingdom and can be distinguished by the presence of bilateral symmetry during embryonic development (Namigai et al. 2014). Bilateria can be identified by several features that distinguish them from non-bilaterian animal lineages. These include bilateral symmetry along the anterior-posterior axes, the development of a complete through-gut with a distinct mouth and anus, the formation of mesodermal tissues including muscle blocks, and cephalisation, resulting in a concentration of sensory structures and neural tissues at the anterior end of the organism (Minelli, 2009; Brusca and Shuster, 2016). Collectively, each of these defining characteristics allowed for the exploration and exploitation of the environment in three dimensions (Holland 2015). The evolution of these features contributed to complexity and ecological versatility of the Bilateria and has likely resulted in the exploitation of the wide range of ecological niches that these organisms came to occupy (Holland 2015). Bilateral symmetry is significant as it allowed for directed movement through the environment through coordinated left-right muscle contractions, giving rise to active predation, new methods of locomotion such as burrowing, and the emergence of complex food webs (Holland 2015).

Although the emergence of the first bilaterian organism took place during the Ediacaran period (represented by *Ikaria warioota* (Evans et al. 2020), the Cambrian period which began around 541 million years ago witnessed an extraordinary expansion of bilaterian diversity in an event widely known as the Cambrian Explosion (Zhuravlev and Wood 2018). This

evolutionary burst encompassed the rapid appearance of numerous animal body plans, significant increases in body size, the emergence of complex ecosystems and the sudden proliferation of mineralized skeletal remains in the fossil record, a pattern first noted by William Buckland in the 19th century (Conway Morris 2000; Chen 2009).

The Cambrian explosion has been described as a "three-phase explosion" (Zhang and Shu 2021) as it likely resulted from a culmination of a series of intertwined ecological, developmental, and genetic changes. Among these, molecular innovations such as genetic novelty are believed to have played a fundamental role (Wu and Lambert 2023). Genetic innovations underpinning bilaterian complexity provide some explanation for their evolutionary success (Erwin et al. 2011). For example, the expansion and diversification of ANTP-class homeobox genes during the bilaterian stem lineage are thought to have been crucial in enabling the development of novel body architectures (Holland 2015; Holland et al. 2017). ANTP-class homeobox genes include the NK, Hox, and ParaHox clusters. NK genes are primarily associated with mesodermal patterning, Hox genes are associated with positional identity along the anterior-posterior axis (especially in the central nervous system), and ParaHox genes are primarily associated with patterning the digestive tract (Holland 2015). These genes, and others associated with developmental control of morphology and physiology likely contributed to the evolution of increasingly complex body plans capable of active, high-energy locomotion, which is a trait thought to be central to diversification during the Cambrian (Carbone and Narbonne 2014).

Current research suggests that both phenotypic and molecular evolution proceeded at accelerated rates during the Cambrian Explosion (Marshall 2006; Lee et al. 2013a). Rates of

phenotypic evolution were approximately four times higher, while molecular evolution occurred at a rate 5.5 times greater during the Cambrian explosion in comparison to the rest of the Phanerozoic eon, which extends from ~538.8 million years ago to the present (Cohen et al. 2013; Lee et al. 2013a). These inferred rates are based on robust evolutionary assumptions about the precise age of arthropods obtained through Bayesian and maximum likelihood phylogenetic clock methods on an extensive anatomical and genomic data set for arthropods (the most diverse Phylum during the Cambrian and today) and are consistent with evolution by natural selection and data from living organisms (Lee et al. 2013a). These findings highlight the importance of genetic novelty in driving rapid diversification and underline the unique evolutionary dynamics of the Cambrian period.

Although genomic novelty is undoubtedly a contributor to the diversity observed during the Cambrian explosion, it is also apparent that reductive evolution observed at a protein-coding level has also been a contributor to genome composition and evolution (Guijarro-Clarke et al. 2020). Previous studies have highlighted the prevalence of both gene and protein gains and losses during the emergence of different animal groups, and the association of the latter in the loss of anatomical structures in evolution (Moore and Bornberg-Bauer 2012; Tsai et al. 2013). In this context, we define novelty as protein-coding loci that are lineage-specific and do not have close homologues in other taxa (Paps and Holland 2018; Rödelsperger et al. 2019; Hoile et al. 2025). This definition reflects the fact that we cannot always determine the mechanism by which a novel gene arose, but also that they likely reflect novel biology from encoding proteins which may have distinct activity and/or function which is absent from outgroup taxa (Hoile et al. 2025). We categorise genetic novelty into three categories: gene duplication and subsequent extensive sequence divergence (Ohno 1970; Rastogi and

Liberles 2005; Conrad and Antonarakis 2007; Sémon and Wolfe 2008; Holland et al. 2017; Dubose and De Roode 2024), horizontal gene transfer (Husnik and McCutcheon 2018; Li et al. 2022; Keeling 2024) and originating from non-coding DNA indicative of de novo origin (Levine et al. 2006; McLysaght and Hurst 2016; Van Oss and Carvunis 2019; Zhao et al. 2024).

One consideration when analysing new genes within the Bilateria is the debate concerning which animal groups reside within the Bilateria and which are considered to be a sister taxon. The phylogenetic relationships between Bilateria and associated outgroups continues to be a subject of debate (Dunn et al. 2014). One group of interest is the phylum Xenacoelomorpha, which includes *Xenoturbella* and Acoelomorpha (although even relationships within Xenacoelomorpha have been recently contested (Redmond 2024)). These marine worms display bilateral symmetry but lack several typical bilaterian features, such as an anus, a circulatory system, and nephridia (Hejnol and Pang 2016). Phylogenetic analyses have resulted in conflicting hypotheses regarding their placement. Some studies suggest that Xenacoelomorpha are placed within the deuterostomes (forming a clade with Ambulacraria), while others propose that they are the earliest branching bilaterians, forming a sister group to the Nephrozoa (comprising both protostomes and deuterostomes) (Mulhair et al. 2022; Juravel et al. 2023; Álvarez-Presas et al. 2024).

In addition to the Xenacoelomorpha, the placement of the non-bilaterian phyla: Ctenophora, Cnidaria, Placozoa, and Porifera remains a matter of discussion (Dunn et al. 2014). Ctenophores (comb jellies) are known for their ciliary locomotion and bi-radial symmetry (Malakhov and Gantsevich 2022). The Cnidaria, such as jellyfish, sea anemones, and corals, possess a diffuse nerve net and specialized stinging cells known as cnidocytes (Schierwater

and DeSalle 2021). Placozoans are often regarded as the simplest free-living animals. They are small, flattened organisms which lack true tissues and organs (Pennisi 2021). The Porifera (more commonly known as sponges) are characterized by a porous body plan and the absence of nervous, digestive, or circulatory systems (Wörheide et al. 2012).

Several phylogenetic topologies have been proposed to resolve the relationships among these groups. One widely supported hypothesis places Cnidaria and Placozoa as sister taxa, together forming a clade that is sister to Bilateria (Laumer et al. 2018). Ctenophora is sometimes positioned as the earliest-branching metazoan lineage, while alternative hypotheses place Porifera in this position (Halanych and Anderson 2015). Although the placement of these outgroups is important for reconstructing early metazoan evolution, this research focusses specifically on the genetic mechanisms which have resulted in bilaterian innovation. The precise relationships among non-Bilateria do not directly impact the identification of gene novelties at the bilaterian node, however acknowledging these phylogenetic complexities provides context for interpreting evolutionary patterns.

This chapter aims to identify putative novel bilaterian gene families which may have contributed to the development of Bilateria-specific tissues, along with function, domain and suggest a mode of origin for each. By focusing on gene novelty and its role in body plan evolution we aim to better understand how the genomic architecture of bilaterians enabled the remarkable biological diversity observed both during the Cambrian period and throughout the subsequent history of animal life. Although previous studies have explored rapid gene expansion and novelty within the Bilateria, until now the breadth of high-quality Bilaterian genomes has not been available (Heger et al. 2020), and although previous studies

have focussed on tissue-specific gene enrichment, they often focus on phyla or sub-taxa within the Bilateria rather than the Bilateria as a whole (Mantica et al. 2024),

This study utilises large genomic datasets which have been made available through sequencing consortia such as the Darwin Tree of Life Project (Blaxter 2022) affiliated to the Earth Biogenome Project (Lewin et al. 2018). Here, we utilise available data by analysing proteomes from 113 Metazoa, including 87 Bilateria. Following data filtering we identify 13 novel hierarchical orthogroups which are enriched in bilaterian-specific tissues of *Octopus bimaculoides* (Protostomia) and *Branchiostoma lanceolatum* (Deuterostomia). We demonstrate evidence of expression in the respective tissues of four further bilaterian species, encompassing six phyla in total. We infer the likely modes of origin for these novel genes and speculate on their importance in bilaterian evolution.

5.3 Materials and methods

5.3.1 Construction of constrained trees

Two species trees were built using guide trees with constrained topology, established from published literature, from BUSCO genes which were identified in 90% of the species used in this analysis, as no universal single copy orthologs were identified in the OrthoFinder outputs. In one constrained tree, Xenacoelomorpha (represented by *Symsagittifera roscoffensis*) was placed as a sister taxon to the Nephrozoa, while in the second constrained tree Xenacoelomorpha was placed within Deuterostomia (Supplementary 2).

To construct these trees, sequences were aligned using MAFFT v7.505 (Katoh and Standley 2013), trimmed using trimAl v1.4.rev15 build (Capella-Gutiérrez et al. 2009) and concatenated with PhyKIT (Steenwyk et al. 2021). The concatenated alignment was used to generate a species tree using IQ-TREE version 2.0-rc1 (Nguyen et al. 2015) with an applied constrained topology. The tree was built using 1000 bootstrap iterations, the LG + G4 model and option -nt AUTO which automatically determines the best number of cores given the current data and computer capacity.

5.3.2 Discovery of novel genes

Proteome data from 113 species (87 Bilateria, 12 Porifera, 1 Placozoan, 2 Ctenophora, 2 Choanoflagellates, 1 Filasterea and 8 Cnidaria) were obtained from Ensembl Rapid Release http://rapid.ensembl.org (accessed August 2024) or NCBI Datasets https://www.ncbi.nlm.nih.gov/datasets (accessed January 2025) (Supplementary 1). Proteomes accessed from Ensembl Rapid Release are proteome predictions based on the Ensembl genebuild annotation pipeline, while those from NCBI Datasets may have a variety of annotation methods such as BRAKER. Taxon sampling was designed to achieve robust phylogenetic coverage across the Bilateria while also preferentially selecting species with proteome predictions-based RNA sequence data. Primary transcripts were obtained from the predicted proteome data and OrthoFinder v3.0.1b1 (Emms & Kelly, 2019) was run to determine orthogroups within the dataset.

Hierarchical orthogroups (HOGs) were extracted for each of the two respective OrthoFinder runs. HOGs gained on the branch leading to the Bilateria from the constrained tree placing

Xenacoelomorpha within Deuterostomia, and on both nodes including and excluding Xenacoelomorpha where they were placed as sister to the Nephrozoa, were extracted using Orthoparser (github.com/PeterMulhair/ortho_parser). HOGs are "sets of genes that are inferred to have descended from a single ancestral gene within a specific clade of species" (Train et al. 2018). This allows for the organisation and understanding of gene families, considering both orthologs and paralogs across different species. Larger, more ancient HOGs may contain multiple smaller, more recent HOGs which reflects the evolutionary history of gene families (Train et al. 2018).

Due to the high rate of gene loss across the Bilateria (Guijarro-Clarke et al. 2020), a threshold value was set to reduce the likelihood of spurious homology. For a HOG to be considered real, rather than a result of spurious sequence homology and incorrect clustering, we impose a rule that it must be present in 25% of species descendent from the selected node, in addition to being present in both *Branchiostoma lanceolatum* and *Octopus bimaculoides*. The latter two species are specified to facilitate later gene expression analyses. HOGs which fit this threshold value were extracted and gene copy plots with phylogenetic trees were created using ggplot2 (Wickham 2016) and ggtree (Yu et al. 2017) in Rstudio (Posit team 2025).

5.3.3 Extracting enriched genes and identifying HOG crossover between species

RNAseq data was obtained from NCBI SRA https://www.ncbi.nlm.nih.gov/sra/ (accessed May 2025) for *Branchiostoma lanceolatum* and *Octopus bimaculoides* (Supplementary 2). Reads for each species were mapped to its respective reference genome

using STAR version STAR_2.4.0g1 (Dobin et al. 2013) and Stringtie was used to quantify expression (Pertea et al. 2015). Genes which met the threshold values above were tested for enrichment in bilaterian-specific tissues of interest using Rstudio (Posit team 2025). For a gene to be considered enriched, its expression had to be twice the value of the median expression of all tissues for that gene and have an FPKM >5. Expression matrices were generated in RStudio using Pheatmap (Kolde Ravio 2025).

Putative functions for genes of interest were identified using BLASTp (Mahram and Herbordt 2015) searches against the standard nr database in addition to Pfam annotations (Mistry et al. 2021). Pie charts displaying putative functions of enriched genes in bilaterian-specific tissues of interest were created using ggplot2 (Wickham 2016)in Rstudio and figures were edited using Affinity Designer 2 (Serif 2024). Venn diagrams displaying shared Hierarchical Orthogroups between tissues and across species were created using VennDiagram in Rstudio (Chen 2022).

5.3.4 Searching for sequence homology beyond Bilateria

GenEra (Barrera-Redondo et al. 2023) was used as an independent test of the mode of origin of genes in HOGs of interest, through the detection of sequence homology across proteins in the NCBI nr database while accounting for homology detection failure. GenEra estimates gene-family founder events and therefore could indicate the likelihood of a gene arising through duplication and divergence, horizontal gene transfer or de novo (Domazet-Lošo et al. 2007).

Proteins for HOGs of interest were extracted for the nine following species across the Bilateria: *Symsagittifera roscoffensis* , *Branchiostoma lanceolatum*, *Gallus gallus*, *Homo sapiens*, *Portunus trituberculatus*, *Drosophila melanogaster*, *Octopus bimaculoides*, *Lingula anatina* and *Lumbricus rubellus*. GenEra was run with standard parameters, in addition to use of the -a parameter to include the proteomes of the nine species above and four representatives of outgroup species: *Aurelia aurita*, *Bolinopsis microptera*, *Sycon ciliatum* and *Monosiga brevicollis* (Supplementary 1).

Where relevant, 3D models were generated using AlphaFold 2 (ColabFold v1.5.5) (Jumper et al. 2021) and visualised using Chimera 1.18 (Pettersen et al. 2004).

5.4 Results

This research investigates the role of gene novelty in tissues that are characteristic of the Bilateria, by addressing two main questions. Firstly, we ask whether novel gene families are being utilised for novel bilaterian tissues, and secondly, what are the modes of origin of novel gene families in the Bilateria? For operational reasons we recognise gene families through a sequence similarity method which defines hierarchical orthogroups (HOGs) inferred to have descended from a single ancestral gene within a specific clade of species (Train et al., 2018). This aims to allow for better understanding of these gene families by considering both orthologs and paralogs across different species. These questions are applied to two phylogenetic tree topologies: the first placing the Xenacoelomorpha within Deuterostomia (hereafter referred to as **tree 1**) and the second placing Xenacoelomorpha as

sister to the Nephrozoa (hereafter referred to as **tree 2**). Full phylogenetic trees are available in Supplementary 2.

5.4.1 Placement of Xenacoelomorpha has a small impact on bilaterian gene family identification

Initial analyses revealed the origin of 1,757 Hierarchical orthogroups (HOGs) emerging on the stem lineage of Bilateria - the internode below the Bilateria node of tree 1. Only 172 novel HOGs were identified on the internode leading to Xenacoelomorpha plus Nephrozoa in tree 2, however 1,619 novel HOGs were identified on the internode below this which excluded Xenacoelomorpha (Figure 5.1).

This small difference in total number of new HOGs between tree 1 and tree 2 may reflect slight differences in how the gene clustering algorithms treat the two data sets. However, the key finding is that both analyses suggest that around 1,700 'new' gene families (HOGs) emerged during the evolution of Bilateria. As both trees gained a similar number of novel HOGs, further analyses were only conducted on the Bilateria node for tree 1, and the node leading to Nephrozoa for tree 2. It was important to determine a cutoff value which was considerate of the extensive gene loss across the Bilateria while still removing HOGs containing genes which may have resulted from spurious homology (Guijarro-Clarke et al. 2020). Therefore, to be considered in further detailed analyses, HOGs were filtered to contain genes in a minimum of 25% of species used in this analysis, in addition to being present in *Branchiostoma lanceolatum* (European amphioxus, deuterostome) *Octopus bimaculoides* (Octopus, protostome) as these species were used in subsequent analyses.

As a result of this filtering step, 93 novel HOGs were retained on the Bilateria node of tree 1 and 73 novel HOGs were identified on the Nephrozoa node of tree 2: a difference of 20 genes between the two trees (Figure 5.1). It was expected that a greater number of HOGs would be identified in tree 1 in comparison to tree 2 as a result of the inclusion of *Symsagittifera roscoffensis* as a representative of the Xenacoelomorpha. With the exception of a very small number of HOGs, the majority of genes are in low copy number (1-5) across both tree configurations.

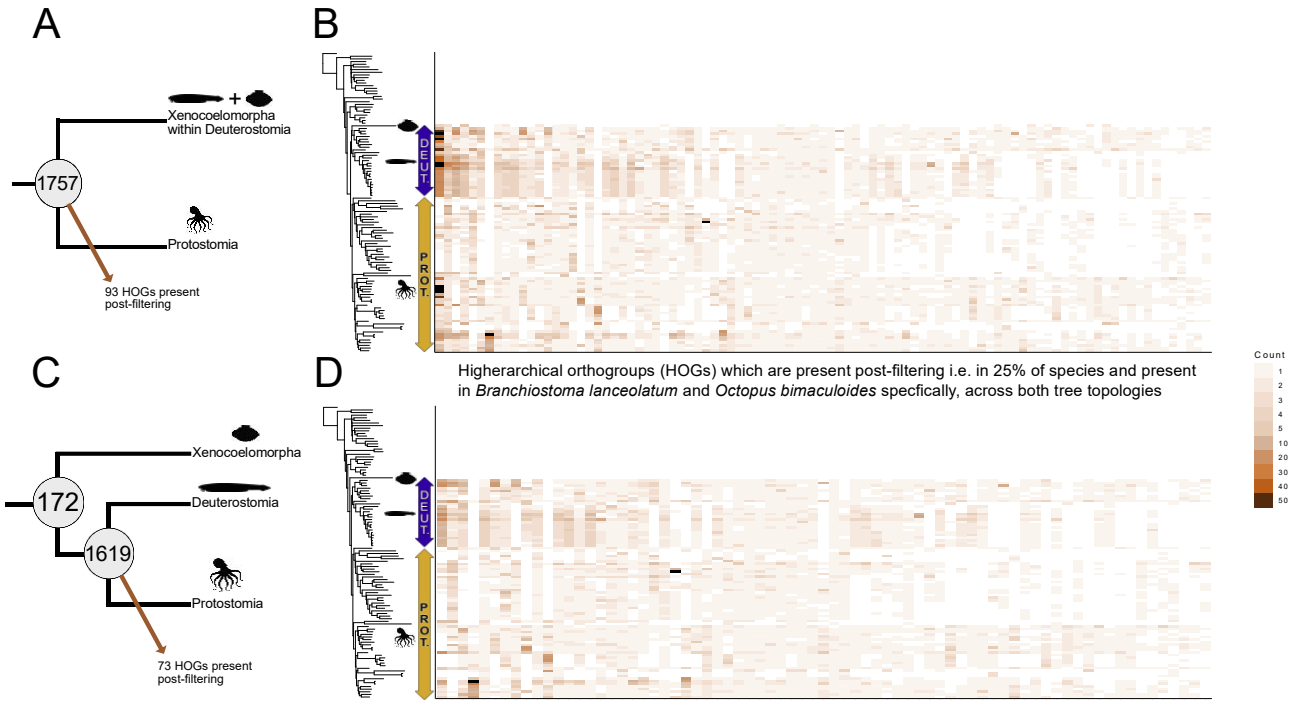


Figure 5.1 – Copy number of genes and the number of Hierarchical Orthogroups (HOGs) gained on the ancestral Bilateria node for two phylogenetic tree topologies: placement of Xenacoelomorpha within Deuterostomes and placement of Xenacoelomorpha as a sister taxon to Bilateria. ***A*** *number of HOGs gained on the Bilateria node for a tree which places Xenacoelomorpha within Deuterostomes (tree 1) and the number of HOGs post filtering which include a minimum of 25% of species in the tree in addition to the presence of at least one gene in Branchiostoma lanceolatum (Deuterostome) and Octopus Bimaculoides (Protostome).* ***B*** *Copy number of genes within the filtered HOGS on the Bilateria node for tree 1.* ***C*** *number of HOGs gained on the Bilateria node for a tree which places Xenacoelomorpha as a sister taxon to Bilateria (tree 2) and the number of HOGs post filtering which include a minimum of 25% of species in the tree in addition to the presence of at least one gene in Branchiostoma lanceolatum (Deuterostome) and Octopus Bimaculoides (Protostome).* ***D*** *Copy number of genes within the filtered HOGS on the node leading to the Nephrozoa for tree 2.*

5.4.2 Bilateria-specific genes have varied functions across tissues

We then asked whether some of the Bilateria-specific gene families showed enriched expression in bilaterian tissues or structures such as through-gut, central nervous system or axial muscle. Genes from the 93 and 73 respective gene families (HOGs) were investigated for evidence of expression enrichment in the above tissues. Genes were determined to be enriched in Bilateria-specific tissues if their expression was >5 FPKM and twice the median expression of that gene across all tissues used in the dataset. As the tissues observed from the two species used (*Branchiostoma lanceolatum* and *Octopus bimaculoides*) varied slightly, they were categorised under four main groups to represent Bilaterian-specific tissues and features as follows: nerve cord/nervous system, through gut, muscle and sensory structures on the head. In amphioxus 218 expression-enriched novel genes were identified for tree 1 and 143 genes were identified in tree 2. In octopus, 126 expression-enriched novel genes were identified for tree 1 and 101 genes were identified for tree 2. Across these, between 15 and 26% of novel genes are enriched in gut tissue, between 25 and 32% enriched in nervous system tissue and 15 and 20% enriched in muscle tissue (Figure 5.2 Supplementary).

Pfam annotation and a BLASTp search against the standard nr database was conducted for each 'expression-enriched' gene to infer a potential function or functional domain, and function was displayed as a pie chart for each tissue for both *B. lanceolatum* and *O. bimaculoides*. In *O. bimaculoides*, the greatest diversity of enriched gene functions was observed in gut tissue and axial nerve cord, while the most specialised functions appeared to be in muscle tissue (Figure 5.2). Results were relatively consistent across both tree topologies, with a slightly greater diversity of function observed in tree 1 (Figure 5.2A and

B). Across all tissues, the majority of 'expression-enriched' genes had the function 'cell adhesion / junction proteins'. It is also interesting to observe a high number of shared gene functions across all tissues; however, these were observed in different proportions according to tissue (Figure 5.2A and B).

A similar pattern of expression-enriched gene function was observed for *B. lanceolatum,* displaying the greatest diversity of enriched gene function in gut tissue and neural tube (Figure 5.2C and D). Despite this, overall *B. lanceolatum* appeared to have a greater diversity of function across all tissues in comparison to *O. bimaculoides* (Figure 5.2). As observed in *O. bimaculoides*, *B. lanceolatum* gene functions were consistent across tree topologies, with a slightly higher diversity of functions observed in tree 1. The greatest difference in tissue novel gene function between species was observed in the muscle. Although it has the least diversity of function of all tissues in both species, novel genes enriched in *B. lanceolatum* muscle tissue have a greater functional diversity than those in *O. bimaculoides*. Despite this, genes with the function of cell adhesion / junction proteins dominate in both cases (Figure 5.2C and D).

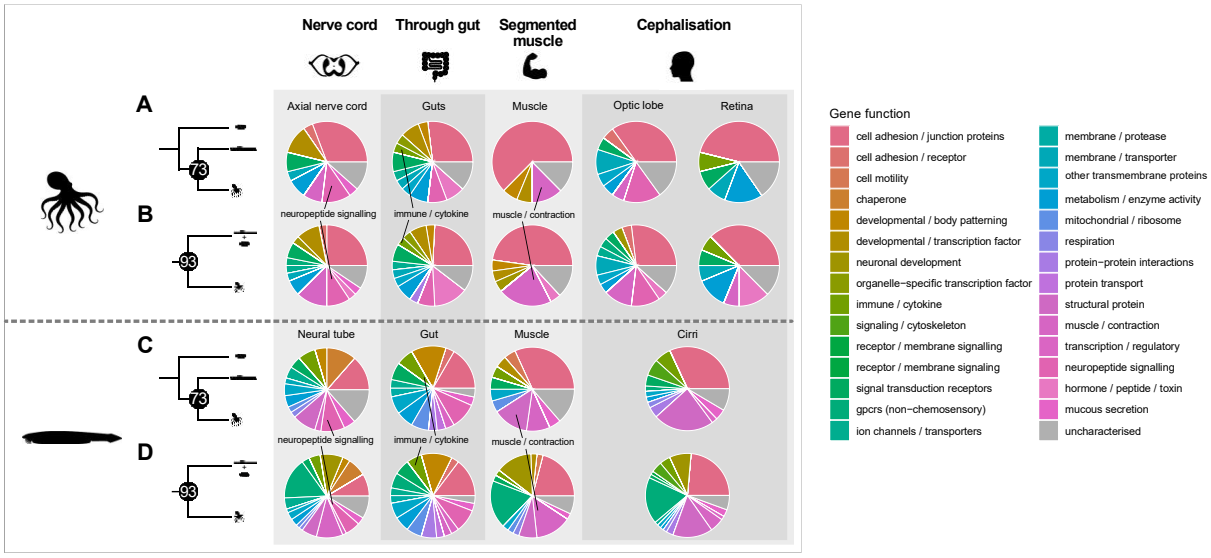


*Figure 5.2 – Enriched gene function in Bilateria-specific tissues observed across two tree topologies and two species. **A** and **C** Enriched genes identified on the phylogenetic tree which places Xenacoelomorpha within Deuterostomes and **B** and **D** Enriched genes identified on the phylogenetic tree which places Xenacoelomorpha as a sister taxon to Bilateria. **A** and **B** represent enriched genes in Octopus bimaculoides while **C** and **D** represent enriched genes in Branchiostoma lanceolatum.*

5.4.3 Novel bilaterian gene families reveal tissue-specific and diverse expression patterns

Novel gene families (HOGs) could comprise sets of paralogous genes with similar roles (e.g. all in gut), or they could comprise paralogues with diverse roles. To assess whether these expression-enriched HOGs in Bilateria were tissue-specific or included genes expressed across tissues, we considered the expression of all genes within each relevant hierarchical orthogroup (HOGs) between tissues of *Octopus bimaculoides* or *Branchiostoma lanceolatum*. Venn diagrams were used to display the overlap between tissues within a species.

In *O. bimaculoides*, the gut contained the highest number of tissue-specific HOGs (10 and 9 HOGs in tree 1 and 2), followed by muscle (6 and 4 in tree 1 and 2). In contrast, the optic

lobe and axial nerve cord showed the greatest degree of exclusive HOG overlap, sharing 6 HOGs in tree 1 and four in tree 2. Only a very small number of HOGs were consistently shared across all bilaterian-specific tissues: two in tree 1 (HOG0000811, an inhibitor of apoptosis, and HOG0011860, Laminin_G_2) and one in tree 2 (HOG0011860). These functional categories are associated with cell survival, differentiation, migration, and adhesion, suggesting a conserved role for these genes in maintaining cellular integrity across tissues (Figure 5.3).

In *Branchiostoma lanceolatum*, neural tube was the tissue with the highest number of tissue-specific HOGs across both trees (13 and 11 HOGs respectively). The tissues with the highest number of exclusive shared HOGs were the neural tube and cirri (4 HOGs in both cases). A higher number of HOGs shared across all Bilaterian-specific tissues was observed in *Branchiostoma lanceolatum* in comparison to *Octopus bimaculoides* (2 and 1 HOGs for *O. bimaculoides* and 5 and 4 HOGs respectively for *B. lanceolatum* (Figure 5.3). Notably, there was no overlap in the set of HOGs shared across all tissues between *O. bimaculoides* and *B. lanceolatum*. This absence of common genes suggests that the specific HOGs recruited for cross-tissue functions are lineage-specific, however it is clear that both species do exhibit tissue-level conservation of certain gene families too.

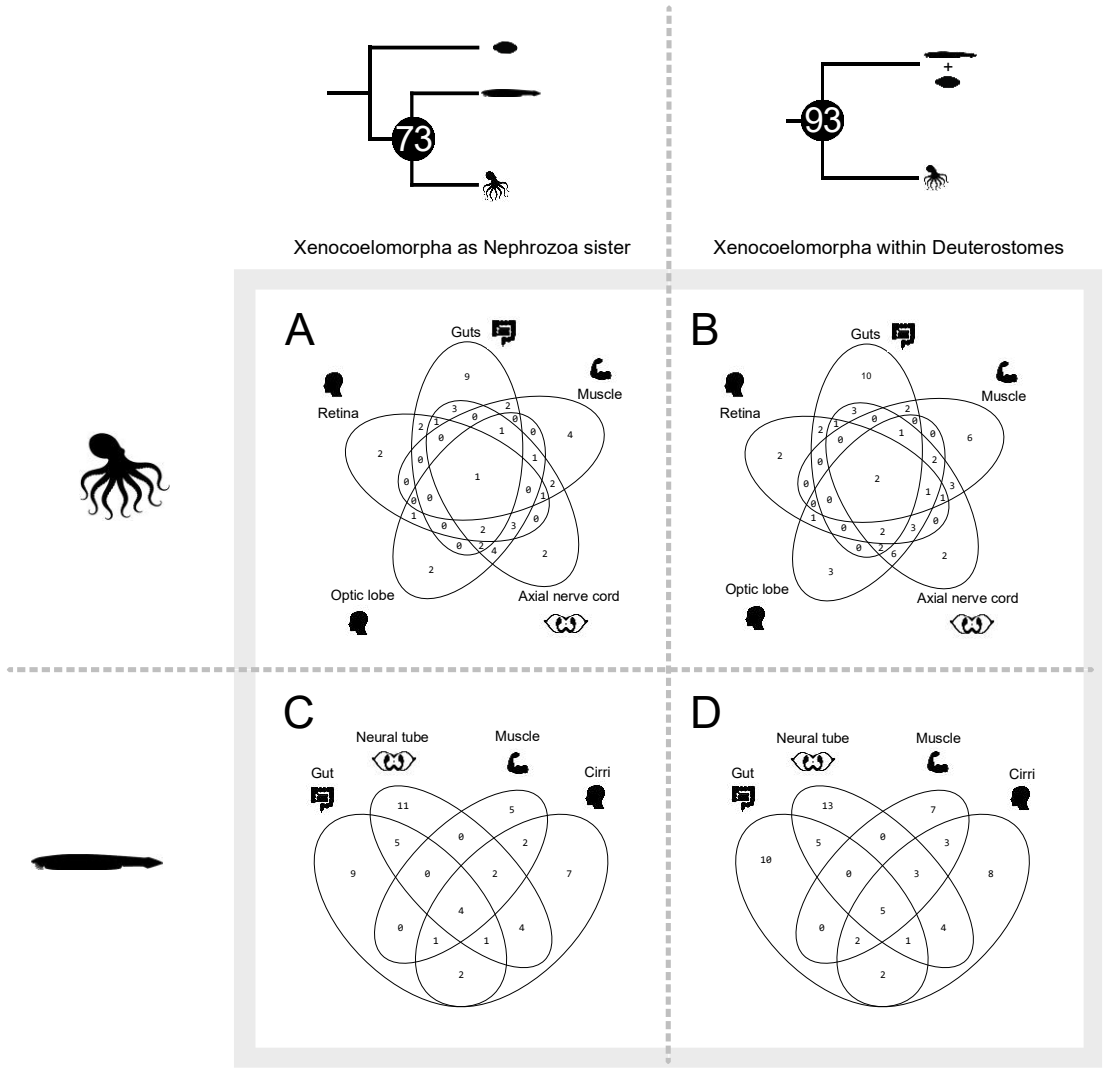


*Figure 5.3 – Venn diagrams of Hierarchical Orthogroups (HOGs) enriched in Bilateria-specific tissues observed across two tree topologies and two species. **A** and **C** Enriched genes identified on the phylogenetic tree which places Xenacoelomorpha as a sister taxon to Bilateria (tree 2). **B** and **D** Enriched genes identified on the phylogenetic tree which places Xenacoelomorpha within Deuterostomes (tree 1). **A** and **B** represent enriched genes in Octopus bimaculoides while **C** and **D** represent enriched genes in Branchiostoma lanceolatum.*

5.4.4 Enriched HOGs are expressed across the Bilateria and have diverse functions

Having compared tissue-specific HOG expression patterns within *B. lanceolatum* or *O. bimaculoides*, we asked whether these gene families had similar expression across both species. In essence, we asked if a 'gut' gene family in octopus is also a 'gut' gene family in amphioxus, and similarly for muscle and central nervous system.

To obtain HOGs of interest, HOGs which were enriched in the same tissues in both *O. bimaculoides* and *B. lanceolatum* were identified i.e. had expression overlap between the two species. The tissues of focus in these analyses were: gut, nerve cord/central nervous system and muscle. As cephalisation can be defined as "the concentration of sense organs, nervous control, at the anterior end of the body, forming a head and brain" (Hombría et al. 2021) and we did not specifically have a whole head tissue sample for either organism, only varied sensory organs, it was deemed unsuitable to proceed with "cephalisation" as a category of one of the enriched tissues.

Overall, 13 HOGs with tissue-specific enrichment of expression conserved between *O. bimaculoides* and *B. lanceolatum* were identified in tree 1, and 9 were identified in tree 2 (Figure 5.4A and B). Of these HOGs which were identified in tree 1; 4 were identified in gut only, 2 in both gut and nerve cord/nervous system, 3 in nerve cord/nervous system only, 3 in muscle only and 1 in nerve cord/nervous system and muscle (Figure 5.4A). In tree 2; 3 HOGs were identified in gut only, 2 in both gut and nerve cord/nervous system, 2 in nerve cord/nervous system only and 2 in muscle only (Figure 5.4B).

To ensure that the identified enriched HOGs weren't an anomaly specific to *B*. *lanceolatum* and *O. bimaculoides*, RNAseq data was obtained for a subset of species representing additional bilaterian phyla and sub-phyla to identify expression patterns. The species selected were *Gallus gallus* (Chordata), *Portunus trituberculatus* (Arthropoda), *Lingula anatina* (Brachiopoda) and *Lumbricus rubellus* (Annelida). As tissue-specific RNAseq data was limited and varied for each species, this analysis searched for evidence of expression in the specific enriched tissue. For example, if a HOG was identified as enriched in gut across both *B*. *lanceolatum* and *O. bimaculoides*, and had evidence of expression in *G. gallus*, this would satisfy the criteria for expression as illustrated in Figure 5.4C. Excluding instances where tissues were unavailable for a given species (denoted NA), all HOGs had evidence of expression in at least half of all species used in this analysis suggesting that tissue-specific enrichment of these bilaterian HOGs occurs across the Bilateria. It should be noted that it appears that some species have undergone gene loss in the instance of some HOGs (Figure 5.4C). HOGs enriched in the gut only, were least likely to have evidence of expression across all species. In contrast, HOGs enriched in more than one tissue were more likely to have evidence of expression across all species, even if this was only in one of the two tissues (Figure 5.4C). It is also possible that errors in RNA sequencing could be responsible for instances where expression is not observed. Figure 5.4D illustrates the findings from Figure 5.4C in the context of location of HOG expression on a model bilaterian.

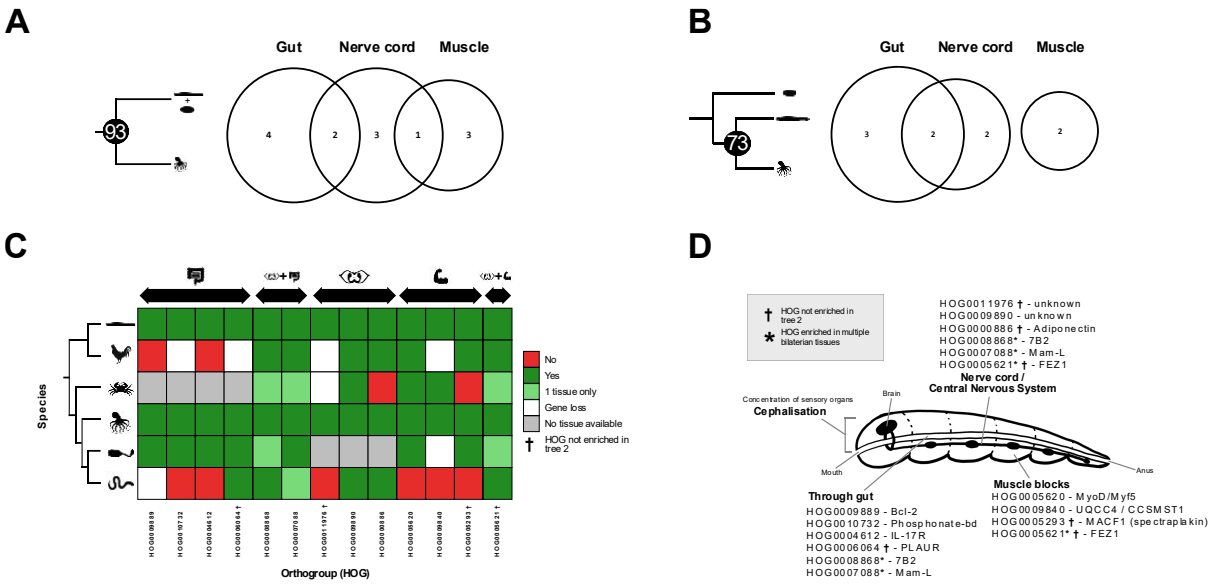


Figure 5.4 – Shared hierarchical orthogroups (HOGs) which were enriched in Bilaterian-specific tissues in both Branchiostoma lanceolatum and Octopus bimaculoides were investigated further to determine whether they were expressed in the same tissues across all Bilateria. ***A*** *HOGs enriched in Bilaterian-specific tissues (gut, nerve cord and muscle) from 'within deuts' tree in both B. lanceolatum and O. bimaculoides showing an overlap of some HOGs between gut & nerve cord and muscle & nerve cord.* ***B*** *HOGs enriched in Bilaterian-specific tissues (gut, nerve cord and muscle) from 'Bilateria sister' tree in both B. lanceolatum and O. bimaculoides showing an overlap of some HOGs between gut & nerve cord only.* ***C*** *Evidence of expression of enriched HOGs specifically in the enriched tissues across additional bilaterian species (top to bottom) B. lanceolatum, Gallus gallus, Portunus Trituberculatus, O. bimaculoides, Lingula anatina and Lumbricus Rubellus. Tissue categories (left to right) gut only, gut and nerve cord, nerve cord only, muscle only, muscle and nerve cord.* ***D*** *Labelled canonical bilaterian displaying bilaterian-specific traits. HOGs enriched in specific tissues are noted.*

Functions of the HOGs which were enriched in bilaterian-specific tissues of *O. bimaculoides* and *B. lanceolatum* and which were evidenced by expression across the Bilateria, were explored further. To glean further information about the identities of the HOGs of interest, sequences within a HOG of interest for *Homo sapiens* and *Drosophila melanogaster* underwent BLASTp searches as these sequences were most likely to have annotation available. These were compared to the Pfam annotation available for *B. lanceolatum* and *O. bimaculoides* to ensure consistency. Once a putative domain or gene family identity was established for a HOG, potential function was noted (Table 1).

In summary, we identified 13 bilaterian-specific gene families which had evolutionarily consistent and specific enriched expression in Bilateria-specific tissues, and where possible identified putative functional domains.

Table 1 - Top BLAST hits for HOGs of interest in Homo sapiens, Drosophila melanogaster and Pfam annotations for Branchiostoma lanceolatum and Octopus bimaculoides for each of the hierarchical orthogroups (HOGs) which are enriched in gut, nerve cord/ central nervous system and muscle across the Bilateria. Where function is marked NA, this indicates the absence of any gene for the given species in the HOG of interest.

Orthogroup	Human	Drosophila	Octopus & Amphioxus	Is functional domain Bilateria-specific?	Enriched tissue	Description
HOG0009889	Bcl-2	NA	Bcl-2	✘	Gut	Apoptosis inhibitor
HOG0010732	NA	NA	Phosphonate-bd	✘	Gut	Not well established
HOG0004612	IL-17R	NA	IL-17R	✔	Gut	Controls gut microbiota, maintains gut homeostasis and prevents excessive inflammation by stimulating production of antimicrobial peptides
HOG0006064	NA	PLAUR	PLAUR	✘	Gut	Cell surface receptor which is involved in tissue remodelling
HOG0008868	7B2	7B2	7B2	✔	Gut + CNS	Molecular chaperone and transcriptional activator for PC2 - implicated in hormone secretion and as an anti-aggregation chaperon to prevent neurodegeneration
HOG0007088	Mam-L	Mam-L	Mam-L1	✘	Gut + CNS	Transcriptional co-activator involved in the Notch, Hippo, Wnt and Shh pathways
HOG0011976	NA	NA	DUF4653	?	CNS	Unknown
HOG0009890	cDNA g20	NA	No Pfam ID	?	CNS	Unknown
HOG0000886	Collagen and C1q	NA	Collagen and C1q	✘	CNS	Adiponectin consists of a collagen and C1q domain. It regulates homeostasis, glucose metabolism, neurogenesis, synaptic plasticity and neuroprotection
HOG0005620	MyoD	MyoD / MyoG	Myf5	✔	Muscle	Myogenic regulatory factors control skeletal muscle development and regeneration. MyoD is involved in differentiation
HOG0009840	UQCC4 / CCSMST1	Dmel_CG30373	CCSMST1	✔	Muscle	Crucial in cellular energy production, especially in tissues which have high energy demants e.g. skeletal muscle
HOG0005293	MACF1	Dmel_CG18304	DUF4482	✘	Muscle	Large cytoskeletal protein involved in microtuble and actin filament coordination in cells. Plays a crucial role in cell proliferation, migration and polarity
HOG0005621	FEZ1	Dmel_CG15365	FEZ1	✔	Muscle + CNS	Multifunctional protein adaptor involved in intracellular transport, neuronal development and regulation of gene expression. It also plays a critical role in brain development by influencing neuronal progenitor specification , migration and axon outgrowth

5.4.5 Bilateria-specific genes enriched in Bilateria-specific tissues have diverse modes of origin

To determine the likely mode of origin for these 13 HOGs of interest, sequence alignments, phylogenetic trees and 3D protein models were generated to search for structural homology against known sequences.

Sequences for *Homo sapiens*, *Drosophila melanogaster*, *Octopus bimaculoides*, *Branchiostoma lanceolatum*, *Aurelia aurita* and *Sycon ciliatum* were extracted for all HOGs in the dataset obtained in the initial OrthoFinder run which contained a Pfam annotation that matched one of the 13 HOGs of interest. Following this, a phylogenetic tree was built by combining each of the 13 HOGs of interest with any homologous HOGs, determined by a

matching Pfam ID (e.g. the HOG of interest and its homologous HOGs all shared a homeodomain). Additionally, where possible known sequences containing the relevant Pfam domain was obtained from NCBI and were added to the relevant trees. This was to determine firstly whether the HOGs of interest had known homologs in other HOGs which matched the putative gene domain and secondly to determine whether the HOG was truly Bilateria-specific. In addition to this, phylogenetic trees gave an indication of the likely mode of origin of HOGs of interest (Supplementary 3).

Where no Pfam ID or known BLAST hits were available, a reciprocal BLAST was performed using genes from all HOGs generated in the OrthoFinder runs. This was to determine whether sequences from these HOGs shared homology to any other HOGs. For example, sequences from HOG0009890 hit CenScul_LOC111634175 from HOG0051786 in a BLAST search, however, the reciprocal BLAST did not yield a significant result. The top hit was LytVari_LOC121410909 with a 58.33% identify, an alignment length of 12, an e-value of 3.6 and bitscore of 17.3. This was a similar case with HOG0011976 where sequences hit SchCali_SC17215.1 from HOG0066939 where the reciprocal BLAST hit SchCali_SC01154.1 with a percentage identity of 50%, alignment length of 18, an e-value of 1.5 and bitscore of 19.6. In both instances, these putatively homologous HOGs contained only two sequences from a single species. In summary, neither HOG0009890 nor HOG0011976 yielded significant BLAST hits to any HOGs in the dataset and therefore their function remains unknown.

As an alternative approach to OrthoFinder, GenEra was run using the nr database and added proteomes of *Symsagittifera roscoffensis*, *Branchiostoma lanceolatum*, *Gallus gallus*, *Homo sapiens*, *Portunus trituberculatus*, *Drosophila melanogaster*, *Octopus bimaculoides*, *Lingula*

anatina, *Lumbricus rubellus*, *Aurelia aurita*, *Sycon ciliatum*, *Bolinopsis microptera* and *Monosiga brevicolli*s. Respective runs were completed for each of the bilaterian species listed above and results were used to determine the most recent common ancestor node for each HOG of interest (Figure 5.5A). This point of origin was considered in combination with the phylogenetic trees generated and known origins of the identified domains in each HOG, to determine the putative mode of origin for each HOG (Figure 5.5A). Although OrthoFinder identified the nodes of origin of HOGs, GenEra was used to identify the inferred mechanism of gene emergence, to understand whether genes with similar domains or identities from outside these HOGs of interest can be identified deeper in evolutionary time.

Three HOGs of interest originated at the node of cellular organism evolution according to GenEra, which suggests that these bilaterian-specific gene families likely arose via duplication followed by extensive sequence divergence (Figure 5.5A). These are HOG0000886, HOG0010732 and HOG0004612. Domains identified within the former two of these are known to be part of large, ancestral gene superfamilies. HOG0000886 contains both collagen and C1q domains and is hypothesized to be adiponectin. Adiponectin is thought to be vertebrate-specific, with some adiponectin-like proteins existing in bilaterian invertebrates. Collagen and C1q domains are much older than this lineage, possibly explaining the point of origin of this HOG (Fujimoto et al., 2023). It is therefore likely that this domain family arose prior to the Bilateria, however as a result of duplication and divergence, this HOG or gene family is Bilateria-specific.

Pfam annotations of *B. lanceolatum* and *O. bimaculoides* suggested that genes within HOG0010732 may contain a phosphonate binding domain which is found in bacteria, fungi and across the Metazoa (George, 2023). Construction of a phylogenetic tree containing genes from this HOG and other phosphonate-binding genes suggested that this HOG has high sequence divergence from other phosphonate binding genes (Supplementary 3). Despite this, the 3D structure constructed by AlphaFold bears a high degree of similarity to the 3D structure of a phosphonate-binding or ABC transporter protein (Figure 5.5B). It is therefore possible that this Bilateria-specific gene family or HOG has arisen by duplication followed by extensive sequence divergence but retained structural homology to a more ancient set of genes. This is recognised in the literature, with these proteins commonly exhibiting structural conservation despite extensive sequence divergence (Locher 2008). HOG0004612 was identified as an interleukin receptor 17 family which is typically vertebrate-specific (Saco et al. 2021). Despite this, other Toll/interleukin receptors have been identified in early Metazoa. It is therefore possible that this HOG is a different interleukin receptor family number and has arisen via duplication and divergence within the Bilateria (Kubick et al. 2021).

10 HOGs of interest were deduced by GenEra to have originated on the branch leading to Bilateria (Figure 5.5A). Six of these are either previously reported as Bilateria-specific gene families or are Bilateria-specific duplications of more ancestral gene families. AlphaFold2 was used to generate 3D models for four HOGs of interest (Figure 5.5C). HOG0005293 Is hypothesised to be Microtubule-Actin Crosslinking Family (MACF) based on a BLASTp search of the human sequence within this HOG. MACF is hypothesised to be Bilateria-specific but part of the larger ancestral Spectraplakin family. HOG00007088 is believed to contain

mastermind protein genes. Although mastermind-like proteins exist in earlier-diverging Metazoa such as the Cnidaria, Mastermind proteins are Bilateria-specific.

The functions and identities of HOG0009890 and HOG0011976 were unknown and therefore 3D models were built to try and understand further about their structure. A human gene from HOG0009890 has previously been identified in studies and although the function is unclear, it has been found to be highly expressed in the cerebellum and cerebellar hemisphere in the human brain (Mudge et al. 2025)(https://shorturl.at/WKv5G). This supports the findings in Figure 5.4C which suggests that this orthogroup is enriched in the nerve cord/nervous system of Bilateria. Both HOGs contain an alpha helix, however the remainder of the sequence is largely disordered and therefore does not provide further information with regards to identity or function.

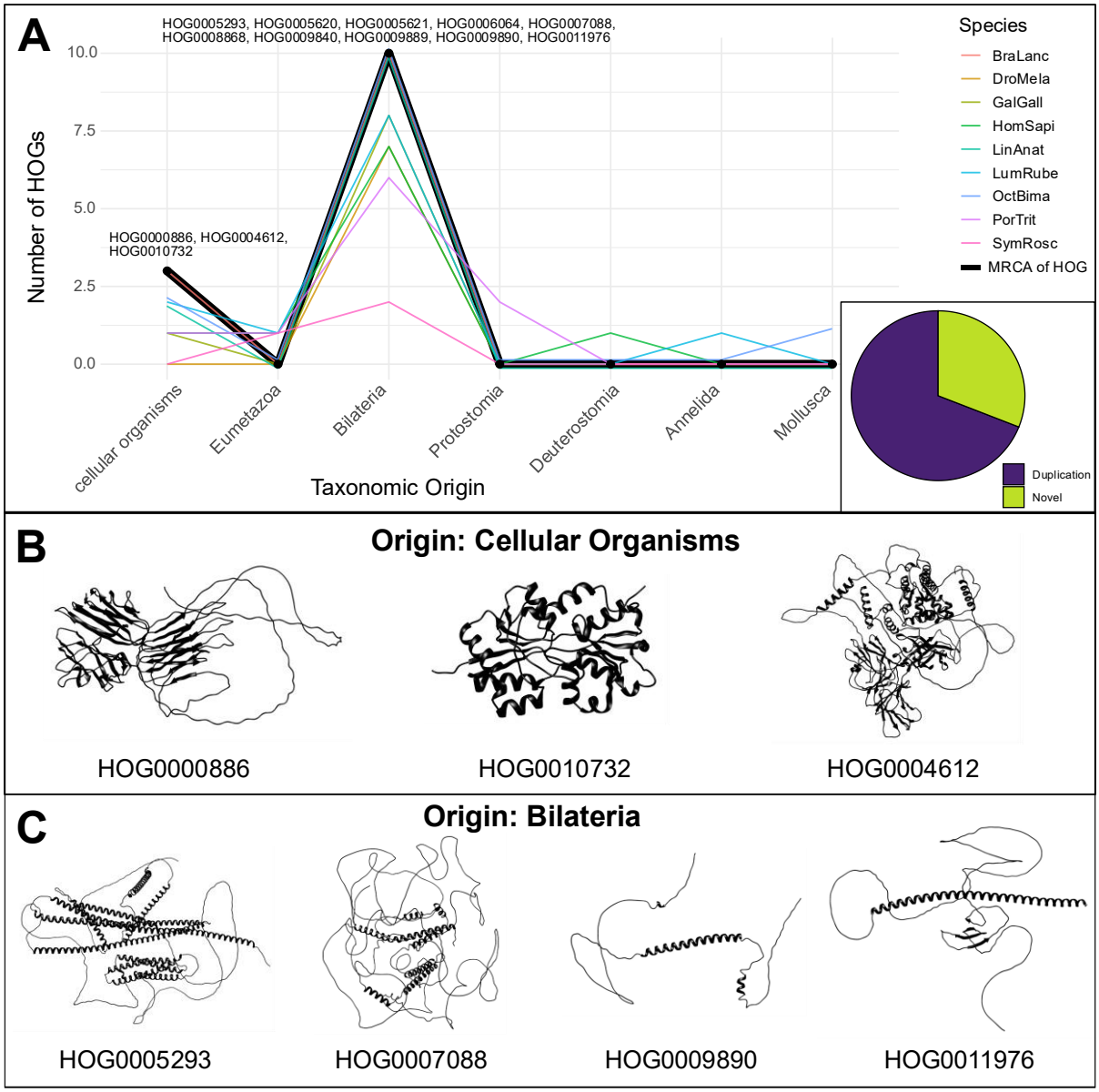


Figure 5.5 - Exploration of putative modes of origin of the 13 hierarchical orthogroups (HOGs) of interest. **A** GenEra putative modes of origin of genes within the 13 HOGs of interest for nine species across the Bilateria: Branchiostoma lanceolatum, Drosophila melanogaster, Gallus gallus, Homo sapiens, Lingula anatina, Lumbricus rubellus, Octopus bimaculoides, Portunus trituberculatus and Symsagittifera roscoffensis shown by coloured lines. A thick black line represents the most recent common ancestor (MRCA) at which a HOG has been recorded to have emerged, according to GenEra. **A** inset: Putative modes of origin for all 13 HOGs are shown on the pie chart. **B** AlphaFold-generated 3D models for each of the three HOGs determined to have emerged within cellular organisms, according to GenEra. **C** AlphaFold-generated 3D models for four of the 10 HOGs determined to have emerged within cellular organisms, according to GenEra. Models were generated for HOG0011976 and HOG0009890 as their identity and function was unknown, a model was generated for HOG0005293 as this is thought to be a bilaterian-specific duplication of the large Spectraplakin family, while HOG0007088 is thought to be a Bilaterian-specific duplication of the mastermind-like gene family. The remaining six HOGs of interest were known to either be documented Bilaterian-specific or Bilaterian-specific duplications of gene families and therefore building 3D models were not necessary to confirm this. All 3D models were generated using sequences from Branchiostoma lanceolatum.

5.5 Discussion

This study identified approximately 1,700 hierarchical orthogroups (HOGs) which emerged on the stem lineage leading to the Bilateria node. It should be noted that the definition of 'novelty' in this study refers to genes which have arisen via mechanisms such as gene duplication followed by divergence, horizontal gene transfer and genes that have arisen de novo, and that altering parameters which relate to sequence divergence, or quality of proteome annotation may influence the count of gene novelty (Weisman et al. 2020).

To mitigate some of these factors, a filtering step was employed to ensure confidence in the number of HOGs which were identified as novel. This meant that only HOGs which were present in at least 25% of species descendent from the node of interest, in addition to the requirement of at least one gene from *Branchiostoma lanceolatum* and *Octopus bimaculoides* being present in the HOG. Although we aimed to identify genes which might play a role in Bilateria-specific tissues across the Bilateria, this filtering step could lead to 'undercounting' of HOGs being identified on this node (Guijarro-Clarke et al. 2020). Another factor to consider is that both constrained phylogenetic trees generated were differentiated by the placement of *Symsagittifera roscoffensis* as a representative of Xenacoelomorpha, however gene loss in this species specifically may falsely undercount gene novelty and could be strengthened by adding additional Xenacoelomorph species when they become available. Following this filtering step, 93 HOGs from tree 1 and 73 HOGs from tree 2 passed the threshold value set above, from the original 1,700.

Of the novel HOGs identified across both phylogenetic topologies, we specifically aimed to identify those whose expression was enriched in tissues that were characteristic of the Bilateria, across bilaterian species. This was achieved by restricting attention to genes which met the criteria of expression in Bilateria-specific tissues with a value of >5 FPKM and twice the mean expression of that gene across all tissues used in the dataset. HOGs which had genes that met the above thresholds, and which were present in both *B. lanceolatum* and *O. bimaculoides* were used in subsequent analyses. 13 HOGs were expression-enriched in Bilateria-specific tissues from the 93 HOGs which met the filtering threshold from tree 1, and 9 of these expression-enriched HOGs were also identified from the 73 filtered HOGs from tree 2.

Firstly, it appears that although novel HOGs which are enriched in bilaterian-specific tissues (gut, nerve cord/nervous system and muscle) may have overlapping functions between tissues, tissues have different proportions of gene functions (Figure 5.2). Enriched novel genes in gut and nerve cord/nervous system have the greatest variety of gene functions in both *Branchiostoma lanceolatum* and *Octopus bimaculoides*, while muscle has the least variety of gene function (Figure 5.2). This result provided reassurance that identified novel genes which were enriched in bilaterian-specific tissues were likely to be real and the functions observed in each tissue were consistent with previous findings. To enforce these findings, evidence of expression was identified in the tissues of interest (gut, nerve cord/nervous system and muscle) across four additional bilaterian species (Figure 5.4C). This suggests that some novel bilaterian gene families may play a role in tissue novelty across the bilaterian lineage and is therefore not just specific to one species.

Gut tissue typically has a multilayered structure and a diversity of cell types which work to perform a variety of functions such as digestion, secretion, absorption and is involved in the immune response (Rao and Wang 2010). Nerve cord tissue is also structurally complex and contains many constituent cells with specialised functions which work together to create neuronal circuits which transmit electrical and chemical signals through the body of an organism, which are essential for processing of information and coordination of responses (National Research Council (US) Committee on Research Opportunities in Biology 1989). In contrast, the primary function of muscle tissue is contraction which is important for movement and force and mainly consists of muscle fibers and contractile proteins (Dave et al. 2023).

When including only HOGs which meet the filtering threshold specified in the methods section, approximately 14% of novel HOGs arising on the bilaterian node in tree 1 and approximately 11% of novel genes arising on the Nephrozoa node in tree 2 are enriched in Bilateria-specific tissues across species. HOGs enriched in these tissues had different types of functions across tissues. Enriched HOGs in the gut typically fall under the category of apoptosis and immune signalling regulators (Lee et al. 2013b). This is surprising given the main function of the gut is digestion or absorption and might suggest that genes related to digestion might be species specific due to the variety of diets observed in the Bilateria (Vianello et al. 2025).

HOGs which are enriched in both gut and nerve cord/nervous system fall under the category of intracellular regulators of hormone processing and gene expression (Watanabe et al., 2013). HOGs which are enriched in nerve cord/nervous system are mostly uncharacterised,

however one is a regulator of homeostasis (Esfahani et al. 2015). HOGs enriched in muscle only fall under the category of differentiation, metabolism and cytoskeletal regulators (Ghasemizadeh et al. 2021), and finally the HOG which is enriched in both muscle and nerve cord/nervous system plays a role in neuronal development and axonal transport regulators (Alhesain et al. 2023).

Similar studies have previously identified four of these 13 HOGs on the node leading to Bilateria (HOG0008868 containing a 7B2 domain, HOG0005620 containing a MyoD domain, HOG0000886 containing Collagen and C1Q domains and HOG0005621 which contains a FEZ1 domain) (Heger et al. 2020). Despite this, it appears that this study has identified nine HOGs which have not previously been identified as related to bilaterian tissue novelty. Although a human gene from HOG0009890 has previously been identified in studies and was also found to be enriched in nervous system tissue, until now this has not been explored in the context of Bilaterian novelty (Mudge et al., 2025). GenEra was used to attempt to identify the origin of each of these HOGs and confirm whether they truly were Bilateria-specific (Barrera-Redondo et al., 2023).

For HOGs where the Pfam ID and BLAST hits indicated a gene family which may have originated outside the Bilateria, phylogenetic trees were built using a subset of species and contained genes from every HOG with a matching Pfam ID to that of the HOG of interest. Where available, known NCBI sequences which matched the Pfam ID were added to the tree, to confirm the identity of sequences within the HOG. Where identity of the HOG was unknown, a reciprocal BLAST search was performed using all HOGs in the dataset to identify any functional domains in the unknown HOGs. This method did not yield any significant

results for HOG0011976 or HOG0009890 and therefore their function and identity remain unknown and requires further analysis beyond the scope of this study (Table 1). This means that it was not possible to identify the function or functional domain of all genes of interest in this study.

GenEra was used as a complementary method to OrthoFinder and used to assist in the understanding of mode of origin of HOGs and not just the node at which they arise. Results largely agreed with the above findings, however placed HOG0000886 on the node at which cellular organisms arise, potentially indicating a Bilateria-specific gene duplication of this HOG. Although the phylogenetic tree built for HOG0010732 and phosphonate binding proteins indicated a large sequence divergence between this HOG and known phosphonate binding proteins, the AlphaFold 3D model generated suggested structural homology, possibly indicating duplication followed by extensive sequence divergence (Supplementary 3, Figure 5.5B).

It is important to acknowledge that OrthoFinder can over or under-group gene families which may skew understanding of their mode of origin. An example of this is HOG0007088 which contains a Mastermind protein domain. In a phylogenetic tree containing Mastermind-like cnidarian sequences from two different orthogroups, these cnidarian sequences sit in the middle of HOG0007088, potentially suggesting that OrthoFinder has split this gene family too stringently. Despite this observation, GenEra results still point towards HOG0007088 being a bilaterian-specific HOG which has arisen on the bilaterian node (Supplementary 3, Figure 5.5A). In a similar context, a phylogenetic tree containing HOG0006064 and corresponding orthogroups of a similar function, splits HOG0006064

within another orthogroup, suggesting it may have been split from another HOG by OrthoFinder. Despite this, GenEra places it as a bilaterian-specific HOG which would indicate it has arisen via duplication and extensive sequence divergence (Supplementary 3, Figure 5.5A).

5.6 Conclusion

Following the filtering threshold set to reduce the likelihood of spurious homology, we demonstrate the emergence of 93 novel hierarchical orthogroups (HOGs) on the node leading to the Bilateria in a constrained phylogenetic tree placing Xenacoelomorpha within Deuterostomia (tree 1) and also demonstrate the emergence of 73 novel HOGs on the node leading to the Nephrozoa in a constrained phylogenetic tree placing Xenacoelomorpha as sister to the Nephrozoa (tree 2), from an original number of approximately 1,700 HOGs. 13 HOGs are enriched in one or more tissue which are characteristic of the Bilateria (gut, nerve cord/nervous system and muscle) in tree 1 and 9 of these are also observed in tree 2. Results were compared with the GenEra output, and it was concluded that nine HOGs arose via duplication followed by extensive sequence divergence, and four appear to have arisen do novo and bear no similarity to any sequences in the tree of life. We demonstrate that some novel bilaterian genes are used for bilaterian tissues and suggest putative identities and functions for genes within each HOG and their likely mode of origin.

5.7 Data availability

Data and code generated in this study can be found on figshare;

10.6084/m9.figshare.30245557

Chapter 6: General Discussion and conclusions

Phenotypic change, whether fixed or not, allows for adaptation to a new environment and the potential for species evolution. As discussed in the introduction, there are multiple drivers by which phenotypic novelty may arise, such as developmental and regulatory mechanisms, ecological and evolutionary influences, and genetic novelty. This research has focussed specifically on the role of gene novelty in animal evolution at major nodes in the tree of life. I hereby focus on a node in deep time: Bilateria, a node at Order-level: Lepidoptera, a node at clade-level: Ditrysia and a node at family-level: Nymphalidae.

Although previous studies have explored gene novelty at multiple nodes of interest in the tree of life, none have utilised the breath of data made available through high-quality sequencing consortia such as the Darwin Tree of Life Project (DToL) (Blaxter 2022), typically either focussing on construction of deep level phylogenies using a large density of species but few loci or the study of specific gene families in depth, but not both (Heger et al., 2020; Kawahara et al., 2019; MacIas-Muñoz et al., 2019; Mantica et al., 2024; Mulhair, Crowley, Boyes, Harper, et al., 2023; Mulhair, Crowley, Boyes, Lewis, et al., 2023). As discussed in Chapter 2, a universally agreed approach on the identification, classification and functions in the context of gene novelty has not yet been established. Therefore, Chapter 2 was dedicated to establishing a suitable method for the above, understanding the associated caveats and understanding how methods may be adapted when working on nodes across different evolutionary timescales.

While each results chapter provides a focussed discussion of individual findings, this discussion chapter aims to synthesise results across different depths of nodes in evolutionary time, exploring the comprehensive themes and resulting implications. Therefore, the purpose of this discussion is not to revisit the details of chapter-specific findings, but instead to integrate findings in a way that addresses the thesis in its entirety.

The overarching aim of this thesis was to explore the mechanisms by which gene novelty may arise at these nodes, identify modes of origin of these genes of interest, investigate putative functions, importance and understand how this may relate to phenotype. This involved developing a pipeline to identify 'new' genes, followed by developing a consistent method to understand modes of origin, followed by methods to understand function. 'Novelty' in the context of evolution has multiple definitions. In this context, 'novelty' included genes arising via gene duplication and extensive sequence divergence, horizontal gene transfer (HGT) and arising 'de novo', meaning they are taxonomically restricted.

This chapter proceeds by synthesising key findings, discussing the limitations of this research, exploring opportunities for further research and finally, offering concluding remarks on findings.

6.1 Key Findings

6.1.1 Detecting gene mode of origin requires different approaches for different taxonomic levels

Chapter 2 discusses the requirement to establish a methodology for the identification, mode of origin and function of novel genes due to the lack of universally agreed approach for this purpose. Although overall a similar pipeline is applied to nodes of interest at different taxonomic levels, determination of the mode of origin or the novel genes identified required different approaches depending on the depth of evolutionary time a given node encompassed. Nodes which have diverged more recently in the tree of life e.g. the Lepidoptera, Ditrysia or Nymphalidae utilised the pipeline identified in Figure 2.4 of Chapter 2 and incorporated outgroup genomes annotated both by Ensembl Genebuild and BRAKER methods. In contrast, ancestral nodes of interest which are located deeper within the tree of life require an alternative approach. This was achieved by extracting hierarchical orthogroups (HOGs) from the OrthoFinder orthogroup outputs to account for the large span of evolutionary time over which these gene families may have evolved. HOGs are “sets of genes that are inferred to have descended from a single ancestral gene within a specific clade of species” and allows for better understanding of gene families by considering both orthologs and paralogs across species (Train et al. 2018). Secondly, GenEra was utilised for genes within HOGs of interest using a breadth of species within the Bilateria and relevant outgroups. These results, along with any information regarding the identity of gene within these HOGs were used to suggest a putative mode of origin.

6.1.2 Gene duplication is the most common form of gene novelty

Gene duplication is the most common mechanism by which gene novelty arises across all four taxonomic levels studied. Genes arising via duplication followed by sequence divergence comprises 82% of gene novelty in Lepidoptera, 74% in Ditrysia, 61% in Nymphalids and 69% of novel genes of interest in Bilateria. This is unsurprising as it is well documented that duplication and divergence is a major mechanism of gene novelty across Metazoa and within Bilateria and Lepidoptera (Fernández & Gabaldón, 2020; Heger et al., 2020b; Mulhair, Crowley, Boyes, Lewis, et al., 2023). In a similar manner, horizontal gene transfer (HGT) is the least common mechanism by which genomic novelty may arise, comprising only <1% in Lepidoptera, 2% in Ditrysia, 5% in Nymphalids and 0% of genes of interest in Bilateria. This is again supported by the current literature which considers HGT to be rare within the Metazoa as a whole in comparison to bacteria due to the requirement of a HGT to enter and be incorporated within the germline (Dunning Hotopp 2011; Ramulu et al. 2012; Nakabachi 2015).

6.1.3 Gene novelty does not always directly link to phenotype

This thesis focussed on the identification of novel genes emerging at nodes giving rise to major transitions within the tree of life, and aimed to identify functions of these novel genes through Pfam annotation, BLASTp searches, exploring expression patterns across tissues and investigating structural homology with genes of known function. Despite this, it is not always possible to ascertain a specific function of a novel gene and as a result it cannot always be linked to a specific phenotype. This is emphasised across chapters 3, 4 and 5, for example

the *propellin* gene families identified as a putative HGT event which appears to have occurred between *Spiroplasma* bacteria and Lepidoptera, resulting in multiple novel lepidopteran orthogroups (Chapter 3, Figure 5 and supplementary). In this instance, function remains unknown however it is clear that these genes are functional across a wide range of lepidopteran species which is suggestive of a gene which may be important in lepidopteran biology. Interestingly, even when analyses are directed towards novel genes which display tissue-specific enrichment, a similar result can be observed. This is observed in two novel bilaterian orthogroups in Chapter 5, one of which has previously been identified in humans, however neither have a known function and therefore cannot be linked to a specific phenotype.

Although this thesis primarily focusses on gene novelty emerging on nodes of interest as a mode of generating phenotypic novelty, Chapter 3 demonstrates that transcriptomic differences between Nymphalid legs do not arise as a result of gene novelty. Instead, differences between the reduced T1 legs and walking (T2 and T3) legs arise from differences in expression of preexisting genes which likely result from the co-option of new regulatory networks. As acknowledged in the introduction of this thesis, genetic novelty is not the sole driver of phenotypic novelty and therefore the emergence of novel genes alone cannot always be directly linked to a novel phenotype. This is something I believe should be acknowledged in further analyses.

6.2 Limitations and Further Directions

Although I do believe that the research covered within this thesis was conducted utilising the most appropriate data available, it is important to acknowledge the caveats resulting from data availability and the arising implication on results.

Firstly, and perhaps most importantly, it must be noted that all analyses conducted are limited by missing samples at the base of each node of interest which either result from lack of sample availability due to unsuccessful collection, or due to species extinction. This means that all work presented in this thesis demonstrates transitions along a stem and not necessarily a specific node. This challenge is presented in Chapter 5 where *Symsagittifera roscoffensis* is the only available Xenacoelomorph species available and therefore genes arising at the base of the Bilateria will be somewhat biased by the presence or absence of a gene in this species alone, to an extent. However, at the time of analysis further data was unavailable.

For this reason, I believe one of the greatest limitations on results arises from lack of species data or a lack of high-quality sequencing data. However, the emergence of data arising from the Darwin Tree of Life Project (DToL) (Blaxter 2022) has allowed for analyses to be more comprehensive than previously possible, resulting in a step forward in the understanding of gene novelty in nodes representing major transitions in the tree of life. Moore's Law observes that the number of components on a chip doubles roughly biennially, resulting in faster, smaller and cheaper computing power. The cost of DNA and RNA sequencing has evolved faster than this law, giving rise to reducing costs and results of increasingly higher

quality. Results are only as good as the genome annotation available. It is therefore hoped that with passing time, genomes will continue to be of higher and higher quality and reduced cost will allow for an even greater number of species to be sequenced. All of this would continue to strengthen and build upon the results of this thesis by filling the gaps of currently unavailable species and therefore presenting more comprehensive results which would give a clearer indication of novelty at a given node, rather than a transition along a stem, as mentioned above. This caveat was also observed particularly during analysis on the lepidopteran node in Chapter 3, where discrepancy in annotation method resulted in the identification of 'novel' genes which were actually present in outgroups. To alleviate this, the pipeline developed in Chapter 2, Figure 2.4 was utilised to ensure all initially identified genes were true positives.

A second consideration arising from these analyses is that altering parameters for the various analyses conducted will alter the number of observed novel genes arising at each node. To minimise this effect, analyses were conducted with more stringent parameters, while considering relevant biology. For example, a threshold value of 25% plus a presence in Amphioxus and Octopus was decided for the Bilaterian analysis in Chapter 5, resulting from the known high volume of gene loss occurring within the Bilateria (Guijarro-Clarke et al. 2020).

A third aspect is linking functional novelty to gene novelty. As explored above, a direct phenotype cannot always be deduced from these novel genes and it is also possible that multiple genes, or regulatory influences may contribute to a given novel trait as observed in the Nymphalidae in Chapter 4. For this reason further analyses may include differential gene expression analyses along with investigating gene novelty to build a better understanding of

how each of these components may interact and understand whether this may be the case in nodes of deeper evolutionary time, such as the Bilateria. Additionally, generating further tissue samples for species of interest may assist in understanding of expression or tissue enrichment, for example Chapter 5 lacked RNAseq data to represent cephalisation, an important bilaterian phenotypic characteristic.

There are multiple avenues for further research to build upon the findings of this field. In the era of advancing artificial intelligence (AI), utilising the tools available might allow for automation of a pipeline to detect genetic novelty, assess mode of origin and decipher functions. Chapters 3 and 5 have already applied AI to a limited extent through the use of AlphaFold 2 in 3D structure prediction of proteins (Jumper et al. 2021). This might be developed further to be able to identify gene functions and possibly link to phenotype by making use of all annotation and functional protein data available to AI. Integration of multi-omics data such as transcriptomics, proteomics and epigenomics may assist with understanding expression, translation and function of candidate novel genes, in addition to identifying protein domains and putative pathways.

Deep learning models such as protein language models may be used as an alternative to traditional methods of assessing sequence similarity such as BLAST, and may be more effective at detecting remote homology through understanding structural and functional constraints which may assist in identifying genes arising from duplication or through horizontal gene transfer (Hamamsy et al. 2023; Holst et al. 2023). Training AI models on coding and non-coding sequences may assist in the identification of genes arising de novo from non-coding DNA. Integration of these features into a single pipeline would improve

prediction and may provide greater understanding into the links between gene novelty, function and phenotype.

Projects of this nature can also be used in the context of scientific outreach and engagement. Generation of the .pdb 3D structure of the proteins arising from these novel genes can be used to create a 3D printed model which can enhance interest and assist with explaining and understanding what complex concepts to the general public. This was successfully achieved using the lepidopteran Propellin protein sequence from *Manduca sexta* (Figure 6.1). I believe this aspect is incredibly important to this ever-advancing field as it an help explain why large-scale initiatives such as the Darwin Tree of Life Project (DToL) are so important to future research. This kind of outreach can also assist non-specialists such as policymakers or researchers across other fields to understand how different approaches may be useful in their own fields. Outreach initiatives can also demonstrate the real-world impact of these studies and understand how they may be applied to fields such as conservation, medicine or biotechnology. This might also hope to inspire the next generation of scientists to continue to develop this field.

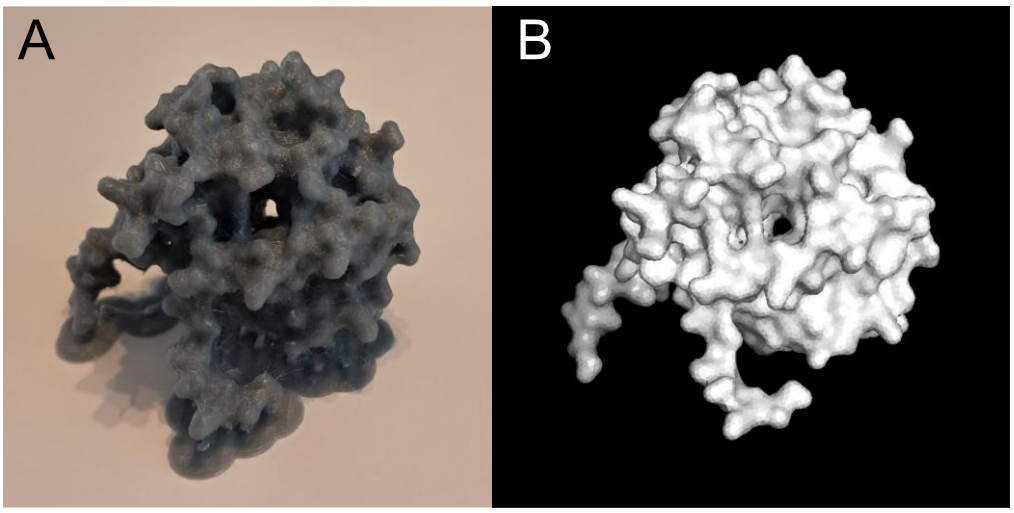


*Figure 6.1 - 3D structures of proteins of interest can be used to generate 3D-printed models. **A** printed model of Lepidopteran Propellin protein generated from **B** the original .pdb file which was converted to .stl format for printing.*

6.3 Concluding remarks

Overall, this thesis has explored gene novelty as one mechanism for phenotypic novelty across multiple taxonomic levels in the metazoan tree of life. This has been achieved through developing a suitable pipeline which can be adapted depending on the depth of the node of interest in evolutionary time. Although gene novelty remains the main focus of this study, it acknowledges that this is not the sole driver of phenotypic innovation and often acts in combination with other mechanisms such as regulatory, developmental or environmental factors.

Novel gene families emerging on the respective stem lineages leading to Bilateria, Lepidoptera, Ditrysia and Nymphalidae were successfully identified, along with putative functions and suggested modes of origin. The significance of these findings has been discussed in the context of the broader field, along with limitations of the study and suggestions for further research.

References

A Dictionary of Biology. 2019. doi: 10.1093/ACREF/9780198821489.001.0001.

Agić, H. et al. 2024. Life through an Ediacaran glaciation: Shale- and diamictite-hosted organic-walled microfossil assemblages from the late Neoproterozoic of the Tanafjorden area, northern Norway. *Palaeogeography, Palaeoclimatology, Palaeoecology* 635, p. 111956. Available at: https://www.sciencedirect.com/science/article/pii/S0031018223005746 [Accessed: 9 July 2025].

Agrawal, P., Heimbruch, K.E. and Rao, S. 2018. Genome-wide maps of transcription regulatory elements and transcription enhancers in development and disease. *Comprehensive Physiology* 9(1), p. 439. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC6566905/ [Accessed: 16 September 2025].

Alhesain, M., Ronan, H., LeBeau, F.E.N. and Clowry, G.J. 2023. Expression of the schizophrenia associated gene FEZ1 in the early developing fetal human forebrain. *Frontiers in Neuroscience* 17, p. 1249973. Available at: http://www.hdbr.org [Accessed: 30 August 2025].

Allen, B.J. and Levinton, J.S. 2007. Costs of bearing a sexually selected ornamental weapon in a fiddler crab. *Functional Ecology* 21(1), pp. 154–161. Available at: /doi/pdf/10.1111/j.1365-2435.2006.01219.x [Accessed: 31 August 2025].

Altenhoff, A.M., Gil, M., Gonnet, G.H. and Dessimoz, C. 2013. Inferring hierarchical orthologous groups from orthologous gene pairs. *PloS one* 8(1). Available at: https://pubmed.ncbi.nlm.nih.gov/23342000/ [Accessed: 16 September 2025].

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215(3), pp. 403–410. Available at: https://www.sciencedirect.com/science/article/pii/S0022283605803602 [Accessed: 3 September 2025].

Álvarez-Presas, M., Ruiz-Trillo, I. and Paps, J. 2024. Novel genomic approaches support Xenacoelomorpha as sister to all Bilateria. *Biorxiv* Available at: https://www.researchsquare.com [Accessed: 6 October 2025].

Andersson, D.I., Jerlström-Hultqvist, J. and Näsvall, J. 2015. Evolution of New Functions De Novo and from Preexisting Genes. *Cold Spring Harbor Perspectives in Biology* 7(6), p. a017996. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC4448608/ [Accessed: 9 July 2025].

Angers, B., Perez, M., Menicucci, T. and Leung, C. 2020. Sources of epigenetic variation and their applications in natural populations. *Evolutionary Applications* 13(6), pp. 1262–1278. Available at: /doi/pdf/10.1111/eva.12946 [Accessed: 30 August 2025].

Aplin, A.C. and Kaufman, T.C. 1997. Homeotic transformation of legs to mouthparts by proboscipedia expression in Drosophila imaginal discs. *Mechanisms of Development* 62(1), pp. 51–60. Available at: https://pubmed.ncbi.nlm.nih.gov/9106166/ [Accessed: 3 September 2025].

Atsumi, K., Lagisz, M. and Nakagawa, S. 2021. Nonadditive genetic effects induce novel phenotypic distributions in male mating traits of F1 hybrids. *Evolution* 75(6), pp. 1304–1315. Available at: /doi/pdf/10.1111/evo.14224 [Accessed: 31 August 2025].

Au, H.M., Nong, W. and Hui, J.H.L. 2025. Whole Genome Duplication in the Genomics Era: The Hidden Gems in Invertebrates? *Genome Biology and Evolution* 17(5), p. evaf073. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC12056724/ [Accessed: 1 October 2025].

Axén, A., Carlsson, A., Engström, Å. and Bennich, H. 1997. Gloverin, an antibacterial protein from the immune hemolymph of Hyalophora pupae. *European Journal of Biochemistry* 247(2), pp. 614–619. Available at: https://pubmed.ncbi.nlm.nih.gov/9266704/ [Accessed: 3 September 2025].

Baral, C., Baral, H.S., Inskipp, C. and Maharjan, R. 2025. Lepidoptera Diversity, Richness, and Distribution in Semi-Urban Farmland and other Habitats around Lumbini, Rupandehi. *bioRxiv*, p. 2025.01.26.634957. Available at: https://www.biorxiv.org/content/10.1101/2025.01.26.634957v1 [Accessed: 31 August 2025].

Barraclough, T.G. 2015. How Do Species Interactions Affect Evolutionary Dynamics Across Whole Communities? *Annual Review of Ecology, Evolution, and Systematics* 46(Volume 46, 2015), pp. 25–48. Available at: https://www.annualreviews.org/content/journals/10.1146/annurev-ecolsys-112414-054030 [Accessed: 4 September 2025].

Barrera-Redondo, J., Lotharukpong, J.S., Drost, H.G. and Coelho, S.M. 2023. Uncovering gene-family founder events during major evolutionary transitions in animals, plants and fungi using GenEra. *Genome Biology* 24(1), pp. 1–21. Available at: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-023-02895-z [Accessed: 29 August 2025].

Baur, R., Haribal, M., Renwick, J.A.A. and Stadler, E. 1998. Contact chemoreception related to host selection and oviposition behaviour in the monarch butterfly, Danaus plexippus. *Physiological Entomology (United Kingdom)* 23(1), pp. 7-19.

Bazinet, A.L., Mitter, K.T., Davis, D.R., Van Nieukerken, E.J., Cummings, M.P. and Mitter, C. 2017. Phylotranscriptomics resolves ancient divergences in the Lepidoptera. *Systematic Entomology* 42(2), pp. 305–316. doi: 10.1111/SYEN.12217.

Bell, E.A., Boehnke, P., Harrison, T.M. and Mao, W.L. 2015. Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon. *Proceedings of the National Academy of Sciences of the United States of America* 112(47), pp. 14518–14521. Available at: /doi/pdf/10.1073/pnas.1517557112?download=true [Accessed: 9 July 2025].

Bento, G., Ogawa, A. and Sommer, R.J. 2010. Co-option of the hormone-signalling module dafachronic acid-DAF-12 in nematode evolution. *Nature* 466(7305), pp. 494–497. Available at: https://pubmed.ncbi.nlm.nih.gov/20592728/ [Accessed: 31 August 2025].

Berry, C. et al. 2008. Brush-Footed Butterflies (Lepidoptera: Nymphalidae). *Encyclopedia of Entomology*, pp. 583–589. Available at: https://link.springer.com/rwe/10.1007/978-1-4020-6359-6_496 [Accessed: 1 September 2025].

Birchler, J.A. and Yang, H. 2022. The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. *The Plant Cell* 34(7), p. 2466. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC9252495/ [Accessed: 2 September 2025].

Bittrich, S., Segura, J., Duarte, J.M., Burley, S.K. and Rose, Y. 2024. RCSB protein Data Bank: exploring protein 3D similarities via comprehensive structural alignments.

Bioinformatics 40(6). Available at: https://dx.doi.org/10.1093/bioinformatics/btae370 [Accessed: 3 September 2025].

Black, C.R. and Berendzen, P.B. 2020. Shared ecological traits influence shape of the skeleton in flatfishes (Pleuronectiformes). *PeerJ* 2020(4), p. e8919. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC7134016/ [Accessed: 31 August 2025].

Blaustein, R. 2016. The Great Oxidation Event: Evolving understandings of how oxygenic life on Earth began. *BioScience* 66(3), pp. 189–195. Available at: https://dx.doi.org/10.1093/biosci/biv193 [Accessed: 4 September 2025].

Blaxter, M.L. 2022. Sequence locally, think globally: The Darwin Tree of Life Project. *Proceedings of the National Academy of Sciences of the United States of America* 119(4), p. e2115642118. Available at: /doi/pdf/10.1073/pnas.2115642118?download=true [Accessed: 8 July 2025].

Bolger, A.M., Lohse, M. and Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15), pp. 2114–2120. Available at: https://dx.doi.org/10.1093/bioinformatics/btu170 [Accessed: 3 September 2025].

Bornberg-Bauer, E. and Albà, M.M. 2013. Dynamics and adaptive benefits of modular protein evolution. *Current Opinion in Structural Biology* 23(3), pp. 459–466. Available at: https://pubmed.ncbi.nlm.nih.gov/23562500/ [Accessed: 3 September 2025].

Le Bourg, É., Valenti, P., Lucchetta, P. and Payre, F. 2001. Effects of mild heat shocks at young age on aging and longevity in Drosophila melangoster. *Biogerontology* 2(3), pp. 155–164. Available at: https://link.springer.com/article/10.1023/A:1011561107055 [Accessed: 30 August 2025].

Bourque, G. et al. 2018. Ten things you should know about transposable elements 06 Biological Sciences 0604 Genetics. *Genome Biology* 19(1). Available at: https://pubmed.ncbi.nlm.nih.gov/30454069/ [Accessed: 9 July 2025].

Brenner, S. 1988. The molecular evolution of genes and proteins: a tale of two serines. *Nature* 334(6182), pp. 528–530. Available at: https://www.nature.com/articles/334528a0 [Accessed: 3 September 2025].

Briscoe, A.D. et al. 2013. Female Behaviour Drives Expression and Evolution of Gustatory Receptors in Butterflies. *PLoS Genetics* 9(7) e1003620. Available at: https://pubmed.ncbi.nlm.nih.gov/23950722/ [Accessed: 4 September 2025].

Brusatte, S.L. et al. 2015. The extinction of the dinosaurs. *Biological Reviews* 90(2), pp. 628–642. Available at: /doi/pdf/10.1111/brv.12128 [Accessed: 9 July 2025].

Brusca, R.C. and Shuster, S.M. 2016. Introduction to the Bilateria and the Phylum Xenacoelomorpha Triploblasty and Bilateral. In: *Invertebrates*. Massachusetts U.S.A: Sinauer Associates, Inc, pp. 345–372.

Bush, S.J., Chen, L., Tovar-Corona, J.M. and Urrutia, A.O. 2017. Alternative splicing and the evolution of phenotypic novelty. *Philosophical Transactions of the Royal Society B: Biological Sciences* 372(1713). Available at: /doi/pdf/10.1098/rstb.2015.0474 [Accessed: 31 August 2025].

Calvert, W.H. and Hanson, F.E. 1983. The role of sensory structures and preoviposition behaviour in oviposition by the patch butterfly, *Chlosyne lacinia*. *Entomologia Experimentalis et Applicata* 33(2), pp. 179–187. Available at: /doi/pdf/10.1111/j.1570-7458.1983.tb03254.x [Accessed: 1 September 2025].

Capella-Gutiérrez, S., Silla-Martínez, J.M. and Gabaldón, T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*

25(15), p. 1972. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC2712344/ [Accessed: 29 August 2025].

Carbone, C. and Narbonne, G.M. 2014. When Life Got Smart: The Evolution of Behavioral Complexity Through the Ediacaran and Early Cambrian of NW Canada. *Journal of Paleontology* 88(2), pp. 309–330. Available at: https://www.cambridge.org/core/journals/journal-of-paleontology/article/abs/when-life-got-smart-the-evolution-of-behavioral-complexity-through-the-ediacaran-and-early-cambrian-of-nw-canada/1DB8C6C8338909B037F0718D1975F03E [Accessed: 13 August 2025].

Carroll, S.B., Gates, J., Keys, D.N., Paddock, S.W., Panganiban, G.E.F., Selegue, J.E. and Williams, J.A. 1994. Pattern formation and eyespot determination in butterfly wings. *Science* 265(5168), pp. 109–114. Available at: https://pubmed.ncbi.nlm.nih.gov/7912449/ [Accessed: 3 September 2025].

Catoni, M., Jonesman, T., Cerruti, E. and Paszkowski, J. 2019. Mobilization of Pack-CACTA transposons in Arabidopsis suggests the mechanism of gene shuffling. *Nucleic Acids Research* 47(3), pp. 1311–1320. Available at: https://pubmed.ncbi.nlm.nih.gov/30476196/ [Accessed: 8 July 2025].

Chan, Y.F. et al. 2009. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science (New York, N.Y.)* 327(5963), p. 302. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC3109066/ [Accessed: 13 September 2025].

Chen, B., Piel, W.H. and Monteiro, A. 2016. Distal-less homeobox genes of insects and spiders: Genomic organization, function, regulation and evolution. *Insect Science* 23(3), pp. 335–352. Available at: https://pubmed.ncbi.nlm.nih.gov/26898323/ [Accessed: 3 September 2025].

Chen, C.K.M., Chan, N.L. and Wang, A.H.J. 2011. The many blades of the β-propeller proteins: Conserved but versatile. *Trends in Biochemical Sciences* 36(10), pp. 553–561. Available at: https://pubmed.ncbi.nlm.nih.gov/21924917/ [Accessed: 3 September 2025].

Chen, H. 2022. VennDiagram: Generate High-Resolution Venn and Euler Plots.

Chen, J.-Y. 2009. The sudden appearance of diverse animal body plans during the Cambrian explosion. *International journal of developmental biology, ISSN 0214-6282, Vol. 53, No. 5-6, 2009, págs. 733-752* 53(5), pp. 733–752. Available at: https://dialnet.unirioja.es/servlet/articulo?codigo=3924082&info=resumen&idioma=ENG [Accessed: 13 August 2025].

Choudhuri, S. 2014. Genes, Genomes, Molecular Evolution, Databases and Analytical Tools. . *Bioinformatics for Beginners*, pp. 27–53.

Cicconardi, F. et al. 2023. Evolutionary dynamics of genome size and content during the adaptive radiation of Heliconiini butterflies. *Nature Communications 2023 14:1* 14(1), pp. 1–24. Available at: https://www.nature.com/articles/s41467-023-41412-5 [Accessed: 3 September 2025].

Cohen, K.M., Finney, S.C., Gibbard, P.L. and Fan, J.-X. 2013. ChronostratChart2022-02. *ICS International Chronostratigraphic Chart*, pp. 199–204. Available at: https://stratigraphy.org/ICSchart/ChronostratChart2022-02.pdf [Accessed: 9 June 2024].

Cohen, O., Gophna, U. and Pupko, T. 2011. The Complexity Hypothesis Revisited: Connectivity Rather Than Function Constitutes a Barrier to Horizontal Gene Transfer.

Molecular Biology and Evolution 28(4), pp. 1481–1489. Available at: https://dx.doi.org/10.1093/molbev/msq333 [Accessed: 3 September 2025].

Conrad, B. and Antonarakis, S.E. 2007. Gene duplication: A drive for phenotypic diversity and cause of human disease. *Annual Review of Genomics and Human Genetics* 8, pp. 17–35. Available at: https://pubmed.ncbi.nlm.nih.gov/17386002/ [Accessed: 8 July 2025].

Conway Morris, S. 2000. The Cambrian 'explosion': Slow-fuse or megatonnage? *Proceedings of the National Academy of Sciences of the United States of America* 97(9), pp. 4426–4429. Available at: /doi/pdf/10.1073/pnas.97.9.4426?download=true [Accessed: 9 July 2025].

Copley, S.D. 2020. Evolution of new enzymes by gene duplication and divergence. *FEBS Journal* 287(7), pp. 1262–1283. Available at: https://pubmed.ncbi.nlm.nih.gov/32250558/ [Accessed: 3 September 2025].

Cosby, R.L., Judd, J., Zhang, R., Zhong, A., Garry, N., Pritham, E.J. and Feschotte, C. 2021. Recurrent evolution of vertebrate transcription factors by transposase capture. *Science* 371(6531). Available at: https://pubmed.ncbi.nlm.nih.gov/33602827/ [Accessed: 3 September 2025].

Crowner, A., Khatri, S., Blichmann, D. and Voss, S.R. 2019. Rediscovering the axolotl as a model for thyroid hormone dependent development. *Frontiers in Endocrinology* 10(APR). Available at: https://pubmed.ncbi.nlm.nih.gov/31031711/ [Accessed: 31 August 2025].

Cvekl, A., Yang, Y., Chauhan, B.K. and Cveklova, K. 2004. Regulation of gene expression by Pax6 in ocular cells: a case of tissue-preferred expression of crystallins in lens. *The International Journal of Developmental Biology* 48(8–9), pp. 829–844. Available at: https://ijdb.ehu.eus/article/041866ac [Accessed: 30 August 2025].

Dave, H.D., Shook, M. and Varacallo, M.A. 2023. Anatomy, Skeletal Muscle. *StatPearls*. Available at: https://www.ncbi.nlm.nih.gov/books/NBK537236/ [Accessed: 29 August 2025].

Dehal, P. and Boore, J.L. 2005. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLOS Biology* 3(10), p. e314. Available at: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0030314 [Accessed: 8 July 2025].

Denecke, S.M. et al. 2020. The identification and evolutionary trends of the solute carrier superfamily in arthropods. *Genome Biology and Evolution* 12(8), pp. 1429–1439. Available at: https://pubmed.ncbi.nlm.nih.gov/32681801/ [Accessed: 3 September 2025].

Diiak, K.T.A., Valuyskiy, M.Yu., Melnitsky, S.I. and Ivanov, V.D. 2023. Sensory structures on mouthpart palps in Trichoptera: ground plan and basal evolution trends. *Contributions to Entomology 73(1): 121-130* 73(1), pp. 121–130. Available at: https://contributions-to-entomology.arphahub.com/article/108068/ [Accessed: 1 September 2025].

Dobin, A. et al. 2013. STAR: ultrafast universal RNAseq aligner. *Bioinformatics* 29(1), pp. 15–21. Available at: https://dx.doi.org/10.1093/bioinformatics/bts635 [Accessed: 29 August 2025].

Domazet-Lošo, T., Brajković, J. and Tautz, D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics* 23(11), pp. 533–539. Available at:

https://www.sciencedirect.com/science/article/abs/pii/S0168952507002995 [Accessed: 16 September 2025].

Dubose, J.G. and De Roode, J.C. 2024. The link between gene duplication and divergent patterns of gene expression across a complex life cycle. *Evolution Letters* 8(5), p. 726. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11424080/ [Accessed: 25 August 2025].

Duncan, E.J., Cunningham, C.B. and Dearden, P.K. 2022. Phenotypic Plasticity: What Has DNA Methylation Got to Do with It? *Insects* 13(2), p. 110. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC8878681/ [Accessed: 30 August 2025].

Dunn, C.W., Giribet, G., Edgecombe, G.D. and Hejnol, A. 2014. Animal phylogeny and its evolutionary implications. *Annual Review of Ecology, Evolution, and Systematics* 45(Volume 45, 2014), pp. 371–395. Available at: https://www.annualreviews.org/content/journals/10.1146/annurev-ecolsys-120213-091627 [Accessed: 13 August 2025].

Dunning Hotopp, J.C. 2011. Horizontal gene transfer between bacteria and animals. *Trends in Genetics* 27(4), pp. 157–163. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0168952511000163 [Accessed: 24 September 2025].

Duplouy, A. and Hornett, E.A. 2018. Uncovering the hidden players in Lepidoptera biology: The heritable microbial endosymbionts. *PeerJ* 2018(5), p. e4629. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC5947162/ [Accessed: 3 September 2025].

Dutrow, E. V., Serpell, J.A. and Ostrander, E.A. 2022. Domestic dog lineages reveal genetic drivers of behavioral diversification. *Cell* 185(25), pp. 4737-4755.e18. Available at: https://pubmed.ncbi.nlm.nih.gov/36493753/ [Accessed: 3 September 2025].

Eacock, A., Rowland, H.M., van't Hof, A.E., Yung, C.J., Edmonds, N. and Saccheri, I.J. 2019. Adaptive colour change and background choice behaviour in peppered moth caterpillars is mediated by extraocular photoreception. *Communications Biology* 2(1), pp. 1–8. Available at: https://www.nature.com/articles/s42003-019-0502-7 [Accessed: 3 September 2025].

Eberhard, S.H. and Krenn, H.W. 2005. Anatomy of the oral valve in nymphalid butterflies and a functional model for fluid uptake in Lepidoptera. *Zoologischer Anzeiger - A Journal of Comparative Zoology* 243(4), pp. 305–312. Available at: https://www.sciencedirect.com/science/article/pii/S0044523105000148 [Accessed: 1 September 2025].

Eickbush, T.H. and Jamburuthugoda, V.K. 2008. The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus research* 134(1–2), p. 221. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC2695964/ [Accessed: 9 July 2025].

Van Eldijk, T.J.B., Wappler, T., Strother, P.K., Van Der Weijst, C.M.H., Rajaei, H., Visscher, H. and Van De Schootbrugge, B. 2018. A Triassic-Jurassic window into the evolution of Lepidoptera. *Science Advances* 4(1), p. e1701568. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC5770165/ [Accessed: 31 August 2025].

Emms, D.M. and Kelly, S. 2019. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology* 20(1), pp. 1–14. Available at: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1832-y [Accessed: 2 September 2025].

Erenler, H.E. and Gillman, M.P. 2010. Synchronisation of adult activity of the archaic moth, Micropterix calthella L. (Lepidoptera, Micropterigidae), with anthesis of sedges

(Carex spp., Cyperaceae) in an ancient wood. *Arthropod-Plant Interactions* 4(2), pp. 117–128. Available at: https://link.springer.com/article/10.1007/s11829-010-9090-7 [Accessed: 31 August 2025].

Erwin, D.H., Laflamme, M., Tweedt, S.M., Sperling, E.A., Pisani, D. and Peterson, K.J. 2011. The Cambrian conundrum: Early divergence and later ecological success in the early history of animals. *Science* 334(6059), pp. 1091–1097. doi: 10.1126/SCIENCE.1206375.

Esfahani, M., Movahedian, A., Baranchi, M. and Goodarzi, M.T. 2015. Adiponectin: an adipokine with protective features against metabolic syndrome. *Iranian Journal of Basic Medical Sciences* 18(5), p. 430. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC4475650/ [Accessed: 30 August 2025].

Estrada, C. and Jiggins, C.D. 2002. Patterns of pollen feeding and habitat preference among Heliconius species. *Ecological Entomology* 27(4), pp. 448–456. Available at: /doi/pdf/10.1046/j.1365-2311.2002.00434.x [Accessed: 4 September 2025].

Evans, K.M., Larouche, O., Watson, S.J., Farina, S., Habegger, M.L. and Friedman, M. 2021. Integration drives rapid phenotypic evolution in flatfishes. *Proceedings of the National Academy of Sciences of the United States of America* 118(18), p. e2101330118. Available at: /doi/pdf/10.1073/pnas.2101330118?download=true [Accessed: 31 August 2025].

Evans, S.D., Hughes, I. V., Gehling, J.G. and Droser, M.L. 2020. Discovery of the oldest bilaterian from the Ediacaran of South Australia. *Proceedings of the National Academy of Sciences of the United States of America* 117(14), pp. 7845–7850. Available at: https://pubmed.ncbi.nlm.nih.gov/32205432/ [Accessed: 8 July 2025].

Eyck, H.J.F., Buchanan, K.L., Crino, O.L. and Jessop, T.S. 2019. Effects of developmental stress on animal phenotype and performance: a quantitative review. *Biological Reviews* 94(3), pp. 1143–1160. Available at: /doi/pdf/10.1111/brv.12496 [Accessed: 30 August 2025].

Eyres, I., Boschetti, C., Crisp, A., Smith, T.P., Fontaneto, D., Tunnacliffe, A. and Barraclough, T.G. 2015. Horizontal gene transfer in bdelloid rotifers is ancient, ongoing and more frequent in species from desiccating habitats. *BMC Biology* 13(1), pp. 1–17. Available at: https://link.springer.com/articles/10.1186/s12915-015-0202-9 [Accessed: 4 September 2025].

Fernández, R. and Gabaldón, T. 2020. Gene gain and loss across the Metazoa Tree of Life. *Nature ecology & evolution* 4(4), p. 524. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC7124887/ [Accessed: 18 September 2025].

Feuda, R., Menon, A.K. and Göpfert, M.C. 2022. Rethinking Opsins. *Molecular Biology and Evolution* 39(3). Available at: https://dx.doi.org/10.1093/molbev/msac033 [Accessed: 3 September 2025].

Fiegler, H., Bassias, J., Jankovic, I. and Brückner, R. 1999. Identification of a gene in Staphylococcus xylosus encoding a novel glucose uptake protein. *Journal of Bacteriology* 181(16), pp. 4929–4936. Available at: https://pubmed.ncbi.nlm.nih.gov/10438764/ [Accessed: 3 September 2025].

Finn, R.D. et al. 2014. Pfam: The protein families database. *Nucleic Acids Research* 42(D1). Available at: https://pubmed.ncbi.nlm.nih.gov/24288371/ [Accessed: 3 September 2025].

Foster, W.J. et al. 2023. How predictable are mass extinction events? *Royal Society Open Science* 10(3). Available at: /doi/pdf/10.1098/rsos.221507 [Accessed: 4 September 2025].

Fox, R.J., Fromhage, L. and Jennions, M.D. 2019. Sexual selection, phenotypic plasticity and female reproductive output. *Philosophical Transactions of the Royal Society B: Biological Sciences* 374(1768), p. 20180184. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC6365872/ [Accessed: 31 August 2025].

Fox, R.M. 1966. Forelegs of Butterflies I. Introduction Chemoreception. *Journal of Research on the Lepidoptera* 5(1), pp. 1–12.

Gabriel, L., Brůna, T., Hoff, K.J., Ebel, M., Lomsadze, A., Borodovsky, M. and Stanke, M. 2024. BRAKER3: Fully automated genome annotation using RNAseq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. *bioRxiv : the preprint server for biology*. Available at: https://pubmed.ncbi.nlm.nih.gov/37398387/ [Accessed: 3 September 2025].

Gao, C.H., Yu, G. and Cai, P. 2021. ggVennDiagram: An Intuitive, Easy-to-Use, and Highly Customizable R Package to Generate Venn Diagram. *Frontiers in Genetics* 12, p. 706907. Available at: https://cran.r- [Accessed: 3 September 2025].

Ghasemizadeh, A. et al. 2021. Macf1 controls skeletal muscle function through the microtubule-dependent localization of extra-synaptic myonuclei and mitochondria biogenesis. *eLife* 10. doi: 10.7554/ELIFE.70490.

Gilbert, S.F. 2000. Modularity: The Prerequisite for Evolution through Development. Available at: https://www.ncbi.nlm.nih.gov/books/NBK10022/ [Accessed: 31 August 2025].

Gonzalez, D.R., Aramendia, A.C. and Davison, A. 2019. Recombination within the Cepaea nemoralis supergene is confounded by incomplete penetrance and epistasis. *Heredity* 123(2), pp. 153–161. Available at: https://www.nature.com/articles/s41437-019-0190-6 [Accessed: 31 August 2025].

Grant, B.R. and Grant, P.R. 2008. Fission and fusion of Darwin's finches populations. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363(1505), p. 2821. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC2606742/ [Accessed: 31 August 2025].

Grant, P.R., Rosemary Grant, B., Huey, R.B., Johnson, M.T.J., Knoll, A.H. and Schmitt, J. 2017. Evolution caused by extreme events. *Philosophical Transactions of the Royal Society B: Biological Sciences* 372(1723), p. 20160146. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC5434096/ [Accessed: 2 September 2025].

Grimson, A. et al. 2008. The early origins of microRNAs and Piwi-interacting RNAs in animals. *Nature* 455(7217), p. 10.1038/nature07415. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC3837422/ [Accessed: 16 September 2025].

Gu, Z. 2022. Complex heatmap visualization. *iMeta* 1(3). Available at: https://pubmed.ncbi.nlm.nih.gov/38868715/ [Accessed: 3 September 2025].

Gubala, A.M., Schmitz, J.F., Kearns, M.J., Vinh, T.T., Bornberg-Bauer, E., Wolfner, M.F. and Findlay, G.D. 2017. The Goddard and Saturn Genes Are Essential for Drosophila Male Fertility and May Have Arisen de Novo. *Molecular Biology and Evolution* 34(5), pp. 1066–1082. Available at: https://pubmed.ncbi.nlm.nih.gov/28104747/ [Accessed: 8 July 2025].

Guijarro-Clarke, C., Holland, P.W.H. and Paps, J. 2020. Publisher Correction: Widespread patterns of gene loss in the evolution of the animal kingdom. *Nature Ecology & Evolution 2020 4:4* 4(4), pp. 661–661. Available at: https://www.nature.com/articles/s41559-020-1159-9 [Accessed: 25 August 2025].

Haggerty, L.S. et al. 2014. A pluralistic account of homology: Adapting the models to the data. *Molecular Biology and Evolution* 31(3), pp. 501–516. Available at: https://pubmed.ncbi.nlm.nih.gov/24273322/ [Accessed: 3 September 2025].

Halanych, K.M. and Anderson, P.A. V. 2015. The ctenophore lineage is older than sponges? That cannot be right! Or can it? *Journal of Experimental Biology* 218(4), pp. 592–597. Available at: https://dx.doi.org/10.1242/jeb.111872 [Accessed: 13 August 2025].

Hall, B.K. and Hanken, J. 2023. Modularity, homology, heterochrony: Gavin de Beer's legacy to the mammalian skull. *Philosophical Transactions of the Royal Society B: Biological Sciences* 378(1880), p. 20220078. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC10184244/ [Accessed: 31 August 2025].

Hamamsy, T. et al. 2023. Protein remote homology detection and structural alignment using deep learning. *Nature Biotechnology* 42(6), pp. 975–985. doi: 10.1038/S41587-023-01917-2.

Hamm, C.A. and Fordyce, J.A. 2015. Patterns of host plant utilization and diversification in the brush-footed butterflies. *Evolution* 69(3), pp. 589–601. Available at: https://dx.doi.org/10.1111/evo.12593 [Accessed: 1 September 2025].

Hanly, J.J., Wallbank, R.W.R., McMillan, W.O. and Jiggins, C.D. 2019. Conservation and flexibility in the gene regulatory landscape of heliconiine butterfly wings. *EvoDevo* 10(1), pp. 1–14. Available at: https://evodevojournal.biomedcentral.com/articles/10.1186/s13227-019-0127-4 [Accessed: 13 September 2025].

Hansen, T.J., Fong, S.L., Day, J.K., Capra, J.A. and Hodges, E. 2024. Human gene regulatory evolution is driven by the divergence of regulatory element function in both cis and trans. *Cell Genomics* 4(4), p. 100536. Available at: https://www.sciencedirect.com/science/article/pii/S2666979X24000922 [Accessed: 13 September 2025].

Hanson, H.E. and Liebl, A.L. 2022. The Mutagenic Consequences of DNA Methylation within and across Generations. *Epigenomes 2022, Vol. 6, Page 33* 6(4), p. 33. Available at: https://www.mdpi.com/2075-4655/6/4/33/htm [Accessed: 30 August 2025].

Hao, Z., Lv, D., Ge, Y., Shi, J., Weijers, D., Yu, G. and Chen, J. 2020. RIdeogram: Drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Computer Science* 6, pp. 1–11. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC7924719/ [Accessed: 3 September 2025].

Harden, N., Yap, S.F., Chiam, M.A. and Lim, L. 1993. A Drosophila gene encoding a protein with similarity to diacylglycerol kinase is expressed in specific neurons. *Biochemical Journal* 289(Pt 2), p. 439. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC1132187/ [Accessed: 3 September 2025].

Harrison, P.W. et al. 2024. Ensembl 2024. *Nucleic Acids Research* 52(D1), pp. D891–D899. doi: 10.1093/NAR/GKAD1049.

Hastings Philip and Rosenberg Susan. 1998. Gene Conversion. In: Delves Peter and Roitt Ivan Maurice eds. *Encyclopedia of Immunology.* . 2nd ed. San Diego: Academic Press, pp. 969–973.

Heger, P., Zheng, W., Rottmann, A., Pan Lio, K.A. and Wiehe, T. 2020. The genetic factors of bilaterian evolution. *eLife* 9, pp. 1–45. doi: 10.7554/ELIFE.45530.

Hejnol, A. and Pang, K. 2016. Xenacoelomorpha's significance for understanding bilaterian evolution. *Current Opinion in Genetics and Development* 39, pp. 48–54. Available at: https://pubmed.ncbi.nlm.nih.gov/27322587/ [Accessed: 8 July 2025].

Herrera-Alsina, L. et al. 2021. Ancient geological dynamics impact neutral biodiversity accumulation and are detectable in phylogenetic reconstructions. *Global Ecology and Biogeography* 30(8), pp. 1633–1642. Available at: /doi/pdf/10.1111/geb.13326 [Accessed: 4 September 2025].

Hoekstra, H.E. and Coyne, J.A. 2007. The locus of evolution: Evo devo and the genetics of adaptation. *Evolution* 61(5), pp. 995–1016. Available at: https://pubmed.ncbi.nlm.nih.gov/17492956/ [Accessed: 3 September 2025].

Hoile, A.E., Holland, P.W.H. and Mulhair, P.O. 2025. Gene novelty and gene family expansion in the early evolution of Lepidoptera. *BMC Genomics* 26(1), pp. 1–14. Available at: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-025-11338-x [Accessed: 25 August 2025].

Holland, P.W.H. 1998. Major Transitions in Animal Evolution: A Developmental Genetic Perspective. *Integrative and Comparative Biology* 38(6), pp. 829–842. Available at: https://dx.doi.org/10.1093/icb/38.6.829 [Accessed: 31 August 2025].

Holland, P.W.H. 2015. Did homeobox gene duplications contribute to the Cambrian explosion? *Zoological Letters 2015 1:1* 1(1), pp. 1–8. Available at: https://zoologicalletters.biomedcentral.com/articles/10.1186/s40851-014-0004-x [Accessed: 8 July 2025].

Holland, P.W.H., Marlétaz, F., Maeso, I., Dunwell, T.L. and Paps, J. 2017. New genes from old: asymmetric divergence of gene duplicates and the evolution of development. *Philosophical Transactions of the Royal Society B: Biological Sciences* 372(1713), p. 20150480. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC5182412/ [Accessed: 8 July 2025].

Holst, F. et al. 2023. Helixer–de novo Prediction of Primary Eukaryotic Gene Models Combining Deep Learning and a Hidden Markov Model. *bioRxiv*, p. 2023.02.06.527280. Available at: https://www.biorxiv.org/content/10.1101/2023.02.06.527280v2 [Accessed: 19 September 2025].

Hombría, J.C.G., García-Ferrés, M. and Sánchez-Higueras, C. 2021. Anterior Hox Genes and the Process of Cephalization. *Frontiers in Cell and Developmental Biology* 9, p. 718175. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC8374599/ [Accessed: 25 August 2025].

Huang, J. 2013. Horizontal gene transfer in eukaryotes: The weak-link model. *BioEssays* 35(10), pp. 868–875. Available at: https://pubmed.ncbi.nlm.nih.gov/24037739/ [Accessed: 8 July 2025].

Huang, S., Yoshitake, K., Kinoshita, S. and Asakawa, S. 2024. Transcriptional landscape of small non-coding RNAs reveals diversity of categories and functions in molluscs. *RNA Biology* 21(1), p. 1. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11067994/ [Accessed: 16 September 2025].

Hughes, C.L. and Kaufman, T.C. 2002. Hox genes and the evolution of the arthropod body plan. *Evolution and Development* 4(6), pp. 459–499. Available at: https://pubmed.ncbi.nlm.nih.gov/12492146/ [Accessed: 1 September 2025].

Husby, A. 2020. On the Use of Blood Samples for Measuring DNA Methylation in Ecological Epigenetic Studies. *Integrative and Comparative Biology* 60(6), p. 1558. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC7742428/ [Accessed: 30 August 2025].

Husnik, F. and McCutcheon, J.P. 2018. Functional horizontal gene transfer from bacteria to eukaryotes. *Nature Reviews Microbiology* 16(2), pp. 67–79. Available at: https://pubmed.ncbi.nlm.nih.gov/29176581/ [Accessed: 8 July 2025].

Hwang, J. and Kim, Y. 2011. RNA interference of an antimicrobial peptide, gloverin, of the beet armyworm, Spodoptera exigua, enhances susceptibility to Bacillus thuringiensis. *Journal of Invertebrate Pathology* 108(3), pp. 194–200. Available at: https://pubmed.ncbi.nlm.nih.gov/21925182/ [Accessed: 3 September 2025].

Ioannidis, P. et al. 2022. A spatiotemporal atlas of the lepidopteran pest Helicoverpa armigera midgut provides insights into nutrient processing and pH regulation. *BMC Genomics* 23(1), pp. 1–12. Available at: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-021-08274-x [Accessed: 3 September 2025].

Jablonski, D. 2001. Lessons from the past: Evolutionary impacts of mass extinctions. *Proceedings of the National Academy of Sciences of the United States of America* 98(10), p. 5393. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC33224/ [Accessed: 2 September 2025].

Jékely, G., Paps, J. and Nielsen, C. 2015. The phylogenetic position of ctenophores and the origin(s) of nervous systems. *EvoDevo* 6(1), pp. 1–9. Available at: https://evodevojournal.biomedcentral.com/articles/10.1186/2041-9139-6-1 [Accessed: 9 July 2025].

Jiggins, F.M., Hurst, G.D.D., Jiggins, C.D., Schulenburg, J.H.G.V.D. and Majerus, M.E.N. 2000. The butterfly Danaus chrysippus is infected by a male-killing Spiroplasma bacterium. *Parasitology* 120(5), pp. 439–446. Available at: https://pubmed.ncbi.nlm.nih.gov/10840973/ [Accessed: 3 September 2025].

Johansen, M., Saenko, S., Schilthuizen, M., Blaxter, M. and Davison, A. 2023. Fine mapping of the Cepaea nemoralis shell colour and mid-banded loci using a high-density linkage map. *Heredity 2023 131:5* 131(5), pp. 327–337. Available at: https://www.nature.com/articles/s41437-023-00648-z [Accessed: 31 August 2025].

Johns, M.E., Tai, P.C. and Derby, C.D. 2004. Serine proteases in the spiny lobster olfactory organ: Their functional expression along a developmental axis, and the contribution of a CUB-serine protease. *Journal of Neurobiology* 61(3), pp. 377–391. Available at: https://pubmed.ncbi.nlm.nih.gov/15389692/ [Accessed: 3 September 2025].

Jumper, J. et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873* 596(7873), pp. 583–589. Available at: https://www.nature.com/articles/s41586-021-03819-2 [Accessed: 3 September 2025].

Juravel, K., Porras, L., Höhna, S., Pisani, D. and Wörheide, G. 2023. Exploring genome gene content and morphological analysis to test recalcitrant nodes in the animal phylogeny. *PLOS ONE* 18(3), p. e0282444. Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0282444 [Accessed: 8 July 2025].

Kaessmann, H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Research* 20(10), p. 1313. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC2945180/ [Accessed: 8 July 2025].

Kapli, P. et al. 2021. Lack of support for Deuterostomia prompts reinterpretation of the first Bilateria. *Science Advances* 7(12). Available at: /doi/pdf/10.1126/sciadv.abe2741?download=true [Accessed: 9 July 2025].

Katoh, K. and Standley, D.M. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* 30(4), pp. 772–780. Available at: https://dx.doi.org/10.1093/molbev/mst010 [Accessed: 29 August 2025].

Kawahara, A.Y. et al. 2019. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proceedings of the National Academy of Sciences of the United States of America* 116(45), pp. 22657–22663. Available at: /doi/pdf/10.1073/pnas.1907847116?download=true [Accessed: 31 August 2025].

Kawahara, A.Y. et al. 2023. A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts and biogeographic origins. *Nature Ecology and Evolution* 7(6), pp. 903–913. Available at: https://www.nature.com/articles/s41559-023-02041-9 [Accessed: 3 September 2025].

Keeling, P.J. 2024. Horizontal gene transfer in eukaryotes: aligning theory with data. *Nature Reviews Genetics* 25(6), pp. 416–430. Available at: https://pubmed.ncbi.nlm.nih.gov/38263430/ [Accessed: 25 August 2025].

Kim, T.K. and Shiekhattar, R. 2015. Architectural and Functional Commonalities between Enhancers and Promoters. *Cell* 162(5), pp. 948–959. Available at: https://www.sciencedirect.com/science/article/pii/S0092867415010211 [Accessed: 16 September 2025].

Kingston, A.C.N. and Cronin, T.W. 2016. Diverse Distributions of Extraocular Opsins in Crustaceans, Cephalopods, and Fish. *Integrative and Comparative Biology* 56(5), pp. 820–833. Available at: https://pubmed.ncbi.nlm.nih.gov/27252200/ [Accessed: 3 September 2025].

Kleinjan, D.A. et al. 2008. Subfunctionalization of Duplicated Zebrafish pax6 Genes by cis-Regulatory Divergence. *PLOS Genetics* 4(2), p. e29. Available at: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0040029 [Accessed: 8 July 2025].

Knopp, M.C.N. and Krenn, H.W. 2003. Efficiency of fruit juice feeding in Morpho peleides (Nymphalidae, Lepidoptera). *Journal of Insect Behavior* 16(1), pp. 67–77. Available at: https://link.springer.com/article/10.1023/A:1022849312195 [Accessed: 1 September 2025].

Kolde Ravio. 2025. Pheatmap.

Kondo, N., Nikoh, N., Ijichi, N., Shimada, M. and Fukatsu, T. 2002. Genome fragment of Wolbachia endosymbiont transferred to X chromosome of host insect. *Proceedings of the National Academy of Sciences of the United States of America* 99(22), pp. 14280–14285. Available at: /doi/pdf/10.1073/pnas.222228199?download=true [Accessed: 8 July 2025].

Krenn, H.W. 2010. Feeding Mechanisms of Adult Lepidoptera: Structure, Function, and Evolution of the Mouthparts. *Annual review of entomology* 55, p. 307. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC4040413/ [Accessed: 3 September 2025].

Krenn, H.W. and Penz, C.M. 1998. Mouthparts of heliconius butterflies (LEPIDOPTERA : NYMPHALIDAE) : a search for anatomical adaptations to pollen-feeding behavior. *International Journal of Insect Morphology and Embryology* 27(4), pp. 301–309. Available at: https://www.sciencedirect.com/science/article/pii/S0020732298000221 [Accessed: 4 September 2025].

Krenn, H.W., Zulka, K.P. and Gatschnegg, T. 2001. Proboscis morphology and food preferences in nymphalid butterflies (Lepidoptera: Nymphalidae). *Journal of Zoology* 254(1), pp. 17–26. doi: 10.1017/S0952836901000528.

Kristensen, N.P. and Skalski, A.W. 1998. Palaeontology and phylogeny. Lepidoptera: Moths and butterflies 1. *Handbuch der Zoologie/Handbook of Zoology IV* (35), pp. 7–25. Available at: https://researchprofiles.ku.dk/en/publications/palaeontology-and-phylogeny-lepidoptera-moths-and-butterflies-1 [Accessed: 3 September 2025].

Kubick, N. et al. 2021. Interleukins and Interleukin Receptors Evolutionary History and Origin in Relation to CD4+ T Cell Evolution. *Genes* 12(6). Available at: https://pubmed.ncbi.nlm.nih.gov/34073576/ [Accessed: 16 September 2025].

Lala, K. 2024. Understanding niche construction and phenotypic plasticity as causes of natural selection. *Palaeontology* 67(4), p. e12719. Available at: /doi/pdf/10.1111/pala.12719 [Accessed: 31 August 2025].

Laland, K.N. and O'Brien, M.J. 2011. Cultural Niche Construction: An Introduction. *Biological Theory* 6(3), pp. 191–202. Available at: https://link.springer.com/article/10.1007/s13752-012-0026-6 [Accessed: 31 August 2025].

Landing, E., Antcliffe, J.B., Geyer, G., Kouchinsky, A., Bowser, S.S. and Andreas, A. 2018. Early evolution of colonial animals (Ediacaran Evolutionary Radiation–Cambrian Evolutionary Radiation–Great Ordovician Biodiversification Interval). *Earth-Science Reviews* 178, pp. 105–135. Available at: https://www.sciencedirect.com/science/article/pii/S0012825217305238 [Accessed: 8 July 2025].

Laumer, C.E., Gruber-Vodicka, H., Hadfield, M.G., Pearse, V.B., Riesgo, A., Marioni, J.C. and Giribet, G. 2018. Support for a clade of placozoa and cnidaria in genes with minimal compositional bias. *eLife* 7. doi: 10.7554/ELIFE.36278.

Lavarone, E., Barbieri, C.M. and Pasini, D. 2019. Dissecting the role of H3K27 acetylation and methylation in PRC2 mediated control of cellular identity. *Nature Communications 2019 10:1* 10(1), pp. 1–16. Available at: https://www.nature.com/articles/s41467-019-09624-w [Accessed: 30 August 2025].

Lee, J.A. and Lupski, J.R. 2006. Genomic Rearrangements and Gene Copy-Number Alterations as a Cause of Nervous System Disorders. *Neuron* 52(1), pp. 103–121. Available at: https://pubmed.ncbi.nlm.nih.gov/17015230/ [Accessed: 8 July 2025].

Lee, M.S.Y., Soubrier, J. and Edgecombe, G.D. 2013a. Rates of Phenotypic and Genomic Evolution during the Cambrian Explosion. *Current Biology* 23(19), pp. 1889–1895. Available at: https://www.sciencedirect.com/science/article/pii/S0960982213009160 [Accessed: 8 July 2025].

Lee, S.Y. et al. 2013b. IL-17-mediated Bcl-2 expression regulates survival of fibroblast-like synoviocytes in rheumatoid arthritis through STAT3 activation. *Arthritis Research and Therapy* 15(1), pp. 1–10. Available at: https://arthritis-research.biomedcentral.com/articles/10.1186/ar4179 [Accessed: 30 August 2025].

Leonard, G. and Richards, T.A. 2012. Genome-scale comparative analysis of gene fusions, genefissions, and the fungal tree of life. *Proceedings of the National Academy of Sciences of the United States of America* 109(52), pp. 21402–21407. Available at: https://www.pnas.org/doi/pdf/10.1073/pnas.1210909110 [Accessed: 3 September 2025].

Levine, M.T., Jones, C.D., Kern, A.D., Lindfors, H.A. and Begun, D.J. 2006. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proceedings of the National Academy of Sciences of the United States of America* 103(26), pp. 9935–9939. Available at: https://pubmed.ncbi.nlm.nih.gov/16777968/ [Accessed: 25 August 2025].

Lewin, H.A. et al. 2018. Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences of the United States of America* 115(17), pp. 4325–4333. Available at: /doi/pdf/10.1073/pnas.1720115115?download=true [Accessed: 13 August 2025].

Li, M., Liu, J., Chen, S., Yao, J., Shi, L., Chen, H. and Chen, X. 2024. VOC Characterization of Byasa hedistus (Lepidoptera: Papilionidae) and Its Visual and Olfactory Responses during Foraging and Courtship. *Insects* 15(7), p. 548. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11276791/ [Accessed: 4 September 2025].

Li, X. et al. 2015. Outbred genome sequencing and CRISPR/Cas9 gene editing in butterflies. *Nature Communications 2015 6:1* 6(1), pp. 1–10. Available at: https://www.nature.com/articles/ncomms9212 [Accessed: 3 September 2025].

Li, Y. et al. 2022. HGT is widespread in insects and contributes to male courtship in lepidopterans. *Cell* 185(16), pp. 2975-2987.e10. Available at: https://pubmed.ncbi.nlm.nih.gov/35853453/ [Accessed: 3 September 2025].

Li, Y., Shen, X.X., Evans, B., Dunn, C.W. and Rokas, A. 2021. Rooting the Animal Tree of Life. *Molecular Biology and Evolution* 38(10), pp. 4322–4333. Available at: https://dx.doi.org/10.1093/molbev/msab170 [Accessed: 9 July 2025].

Li, Z., Tiley, G.P., Galuska, S.R., Reardon, C.R., Kidder, T.I., Rundell, R.J. and Barker, M.S. 2018. Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proceedings of the National Academy of Sciences* 115(18), pp. 4713–4718. Available at: /doi/pdf/10.1073/pnas.1710791115?download=true [Accessed: 1 October 2025].

Liang, K. 2024. Unveiling the Patterns and Impact of New Gene Recruitment in Development and Evolution. *Computational Molecular Biology* 14(0), pp. 202–210. Available at: https://bioscipublisher.com/index.php/cmb/article/view/3985 [Accessed: 30 August 2025].

Livraghi, L. et al. 2024. A long noncoding RNA at the cortex locus controls adaptive coloration in butterflies. *Proceedings of the National Academy of Sciences of the United States of America* 121(36), p. e2403326121. Available at: /doi/pdf/10.1073/pnas.2403326121?download=true [Accessed: 3 September 2025].

Locher, K.P. 2008. Structure and mechanism of ATP-binding cassette transporters. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364(1514), pp. 239–245. Available at: /doi/pdf/10.1098/rstb.2008.0125 [Accessed: 16 September 2025].

Lohse, K. and Weir, J. 2021. The genome sequence of the meadow brown, Maniola jurtina (Linnaeus, 1758). *Wellcome Open Research* 6. Available at: https://pubmed.ncbi.nlm.nih.gov/36866280/ [Accessed: 8 July 2025].

Long, M., Betrán, E., Thornton, K. and Wang, W. 2003. The origin of new genes: Glimpses from the young and old. *Nature Reviews Genetics* 4(11), pp. 865–875. Available at: https://pubmed.ncbi.nlm.nih.gov/14634634/ [Accessed: 3 September 2025].

Long, M. and Langley, C.H. 1993. Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. *Science* 260(5104), pp. 91–95. Available at: https://pubmed.ncbi.nlm.nih.gov/7682012/ [Accessed: 3 September 2025].

Love, A.C. and Wagner, G.P. 2022. Co-option of stress mechanisms in the origin of evolutionary novelties. *Evolution* 76(3), pp. 394–413. Available at: https://academic.oup.com/evolut/article/76/3/394/6728495 [Accessed: 30 August 2025].

Love, M.I., Huber, W. and Anders, S. 2014. Moderated estimation of fold change and dispersion for RNAseq data with DESeq2. *Genome Biology* 15(12), pp. 1–21. Available at: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8 [Accessed: 3 September 2025].

Ma, H., Wang, M., Zhang, Y.E. and Tan, S. 2023. The power of "controllers": Transposon-mediated duplicated genes evolve towards neofunctionalization. *Journal of Genetics and Genomics* 50(7), pp. 462–472. Available at: https://www.sciencedirect.com/science/article/pii/S1673852723000905 [Accessed: 8 July 2025].

MacIas-Muñoz, A., Rangel Olguin, A.G., Briscoe, A.D. and Li, W.H. 2019. Evolution of Phototransduction Genes in Lepidoptera. *Genome Biology and Evolution* 11(8), pp. 2107–2124. Available at: https://dx.doi.org/10.1093/gbe/evz150 [Accessed: 3 September 2025].

Mahram, A. and Herbordt, M.C. 2015. NCBI BLASTP on high-performance reconfigurable computing systems. *ACM Transactions on Reconfigurable Technology and Systems* 7(4). Available at: https://dl.acm.org/doi/pdf/10.1145/2629691 [Accessed: 29 August 2025].

Malakhov, V. V. and Gantsevich, M.M. 2022. The Origin and Main Trends in the Evolution of Bilaterally Symmetrical Animals. *Paleontological Journal* 56(8), pp. 887–937. doi: 10.1134/S0031030122080044.

Mantica, F. et al. 2024. Evolution of tissue-specific expression of ancestral genes across vertebrates and insects. *Nature Ecology and Evolution* 8(6), pp. 1140–1153. Available at: https://www.nature.com/articles/s41559-024-02398-5 [Accessed: 25 August 2025].

Mark Welch, D.B., Mark Welch, J.L. and Meselson, M. 2008. Evidence for degenerate tetraploidy in bdelloid rotifers. *Proceedings of the National Academy of Sciences* 105(13), pp. 5145–5149. Available at: /doi/pdf/10.1073/pnas.0800972105?download=true [Accessed: 1 October 2025].

Marshall, C.R. 2006. Explaining the Cambrian 'explosion' of animals. *Annual Review of Earth and Planetary Sciences* 34(Volume 34, 2006), pp. 355–384. Available at: https://www.annualreviews.org/content/journals/10.1146/annurev.earth.33.031504.103001 [Accessed: 13 August 2025].

Martay, B., Monteith, D.T., Brewer, M.J., Brereton, T., Shortall, C.R. and Pearce-Higgins, J.W. 2016. An indicator highlights seasonal variation in the response of Lepidoptera communities to warming. *Ecological Indicators* 68, pp. 126–133. Available at: https://www.sciencedirect.com/science/article/abs/pii/S1470160X16300115 [Accessed: 31 August 2025].

Masai, I., Hosoya, T., Kojima, S.I. and Hotta, Y. 1992. Molecular cloning of a Drosophila diacylglycerol kinase gene that is expressed in the nervous system and muscle. *Proceedings of the National Academy of Sciences of the United States of America* 89(13), pp. 6030–6034. Available at: /doi/pdf/10.1073/pnas.89.13.6030?download=true [Accessed: 3 September 2025].

McCintock, B. 1950. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America* 36(6), pp. 344–355. Available at: /doi/pdf/10.1073/pnas.36.6.344?download=true [Accessed: 8 July 2025].

McHale, F., Mulhair, P.O. and Holland, P.W.H. 2025. Evolution of Duplicated Hox Gene Clusters in Land Snails and Slugs. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 344(6), pp. 363–368. Available at: /doi/pdf/10.1002/jez.b.23322 [Accessed: 1 October 2025].

McLysaght, A. and Hurst, L.D. 2016. Open questions in the study of de novo genes: What, how and why. *Nature Reviews Genetics* 17(9), pp. 567–578. Available at: https://pubmed.ncbi.nlm.nih.gov/27452112/ [Accessed: 25 August 2025].

McNamara, K.J. 2012. Heterochrony: The Evolution of Development. *Evolution: Education and Outreach* 5(2), pp. 203–218. Available at: https://evolution-outreach.biomedcentral.com/articles/10.1007/s12052-012-0420-3 [Accessed: 31 August 2025].

Medina-Rivera, A., Santiago-Algarra, D., Puthier, D. and Spicuglia, S. 2018. Widespread Enhancer Activity from Core Promoters. *Trends in Biochemical Sciences* 43(6), pp. 452–468. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0968000418300574 [Accessed: 16 September 2025].

Meyer, A. and Van De Peer, Y. 2005. From 2R to 3R: Evidence for a fish-specific genome duplication (FSGD). *BioEssays* 27(9), pp. 937–945. Available at: https://pubmed.ncbi.nlm.nih.gov/16108068/ [Accessed: 9 July 2025].

Meyer, A.J., Ellefson, J.W. and Ellington, A.D. 2013. Library generation by gene shuffling. *Current Protocols in Molecular Biology* 105(SUPPL.105). Available at: https://pubmed.ncbi.nlm.nih.gov/24510437/ [Accessed: 8 July 2025].

Minelli, Alessandro. 2009. Perspectives in animal phylogeny and evolution. p. 345. Available at: https://global.oup.com/academic/product/perspectives-in-animal-phylogeny-and-evolution-9780198566212 [Accessed: 9 July 2025].

Minh, B.Q. et al. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* 37(5), pp. 1530–1534. Available at: https://pubmed.ncbi.nlm.nih.gov/32011700/ [Accessed: 3 September 2025].

Mistry, J. et al. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Research* 49(D1), pp. D412–D419. Available at: https://dx.doi.org/10.1093/nar/gkaa913 [Accessed: 29 August 2025].

Mitter, C., Davis, D.R. and Cummings, M.P. 2017. Phylogeny and Evolution of Lepidoptera. *Annual Review of Entomology* 62, pp. 265–283. Available at: https://pubmed.ncbi.nlm.nih.gov/27860521/ [Accessed: 31 August 2025].

Moen, R.A., Pastor, J. and Cohen, Y. 1999. Antler growth and extinction of Irish elk. *Evolutionary Ecology Research* 1(2), pp. 235–249. Available at:

https://experts.umn.edu/en/publications/antler-growth-and-extinction-of-irish-elk [Accessed: 31 August 2025].

Monteiro, A. 2021. Distinguishing serial homologs from novel traits: Experimental limitations and ideas for improvements. *BioEssays* 43(1). Available at: https://pubmed.ncbi.nlm.nih.gov/33118632/ [Accessed: 3 September 2025].

Monteiro, A., Murugesan, S.N., Prakash, A. and Papa, R. 2025. The Developmental Origin of Novel Complex Morphological Traits in Lepidoptera. *Annual Review of Entomology* 70(1), pp. 421–439. Available at: https://www.annualreviews.org/content/journals/10.1146/annurev-ento-021324-020504 [Accessed: 3 September 2025].

Mooi, R. 2009. Evolution, Second Edition. Douglas J. Futuyma. *Integrative and Comparative Biology* 49(6), pp. 722–723. doi: 10.1093/ICB/ICP095.

Moore, A.D. and Bornberg-Bauer, E. 2012. The Dynamics and Evolutionary Potential of Domain Loss and Emergence. *Molecular Biology and Evolution* 29(2), pp. 787–796. Available at: https://dx.doi.org/10.1093/molbev/msr250 [Accessed: 25 August 2025].

Moriyama, Y. and Koshiba-Takeuchi, K. 2018. Significance of whole-genome duplications on the emergence of evolutionary novelties. *Briefings in Functional Genomics* 17(5), pp. 329–338. Available at: https://dx.doi.org/10.1093/bfgp/ely007 [Accessed: 9 July 2025].

Mudge, J.M. et al. 2025. GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Research* 53(D1), pp. D966–D975. Available at: https://dx.doi.org/10.1093/nar/gkae1078 [Accessed: 16 September 2025].

Mulhair, P.O. et al. 2023a. Bursts of novel composite gene families at major nodes in animal evolution. *bioRxiv*, p. 2023.07.10.548381. Available at: https://www.biorxiv.org/content/10.1101/2023.07.10.548381v1 [Accessed: 3 September 2025].

Mulhair, P.O. et al. 2023b. Diversity, duplication, and genomic organization of homeobox genes in Lepidoptera. *Genome Research* 33(1), pp. 32–44. Available at: https://pubmed.ncbi.nlm.nih.gov/36617663/ [Accessed: 3 September 2025].

Mulhair, P.O., Crowley, L., Boyes, D.H., Lewis, O.T. and Holland, P.W.H. 2023. Opsin Gene Duplication in Lepidoptera: Retrotransposition, Sex Linkage, and Gene Expression. *Molecular Biology and Evolution* 40(11). Available at: https://pubmed.ncbi.nlm.nih.gov/37935057/ [Accessed: 3 September 2025].

Mulhair, P.O., McCarthy, C.G.P., Siu-Ting, K., Creevey, C.J. and O'Connell, M.J. 2022. Filtering artifactual signal increases support for Xenacoelomorpha and Ambulacraria sister relationship in the animal tree of life. *Current Biology* 32(23), pp. 5180-5188.e3. Available at: https://www.sciencedirect.com/science/article/pii/S0960982222016840 [Accessed: 9 July 2025].

Mullen, G.R. 2002. Moths and butterflies (Lepidoptera). *Medical and Veterinary Entomology*, pp. 363–381. doi: 10.1016/B978-012510451-7/50020-7.

Myers, J. 1969. Distribution of foodplant chemoreceptors on the female florida Queen butterfly, Danaus gilippus berenice (Nymphalidae). *J Lep Soc* 23, pp. 196–198. Available at: https://cir.nii.ac.jp/crid/1573668925009646464.bib?lang=en [Accessed: 1 September 2025].

Nakabachi, A. 2015. Horizontal gene transfers in insects. *Current Opinion in Insect Science* 7, pp. 24–29. Available at:

https://www.sciencedirect.com/science/article/abs/pii/S2214574515000371 [Accessed: 18 September 2025].

Nakatani, Y. and McLysaght, A. 2019. Macrosynteny analysis shows the absence of ancient whole-genome duplication in lepidopteran insects. *Proceedings of the National Academy of Sciences* 116(6), pp. 1816–1818. Available at: /doi/pdf/10.1073/pnas.1817937116?download=true [Accessed: 1 October 2025].

Namigai, E.K.O., Kenny, N.J. and Shimeld, S.M. 2014. Right across the tree of life: The evolution of left-right asymmetry in the Bilateria. *Genesis* 52(6), pp. 458–470. Available at: https://pubmed.ncbi.nlm.nih.gov/24510729/ [Accessed: 9 July 2025].

National Research Council (US) Committee on Research Opportunities in Biology. 1989. The Nervous System and Behavior. Available at: https://www.ncbi.nlm.nih.gov/books/NBK217810/ [Accessed: 29 August 2025].

Navarro, L., Harvey, A.É. and Morin, H. 2017. Lepidoptera wing scales: a new paleoecological indicator for reconstructing spruce budworm abundance1. *https://doi.org/10.1139/cjfr-2017-0009* 48(3), pp. 302–308. Available at: /doi/pdf/10.1139/cjfr-2017-0009?download=true [Accessed: 31 August 2025].

Nguyen, L.T., Schmidt, H.A., Von Haeseler, A. and Minh, B.Q. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* 32(1), pp. 268–274. Available at: https://dx.doi.org/10.1093/molbev/msu300 [Accessed: 29 August 2025].

Nijhout, H.F. 2025. Genetic assimilation, robustness and plasticity are key processes in the development and evolution of novel traits. *Developmental Biology* 523, pp. 132–138. Available at: https://www.sciencedirect.com/science/article/pii/S001216062500106X [Accessed: 31 August 2025].

Noffke, N., Christian, D., Wacey, D. and Hazen, R.M. 2013. Microbially Induced Sedimentary Structures Recording an Ancient Ecosystem in the ca. 3.48 Billion-Year-Old Dresser Formation, Pilbara, Western Australia. *Astrobiology* 13(12), p. 1103. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC3870916/ [Accessed: 9 July 2025].

Nowell, R.W., Rodriguez, F., Hecox-Lea, B.J., Mark Welch, D.B., Arkhipova, I.R., Barraclough, T.G. and Wilson, C.G. 2024. Bdelloid rotifers deploy horizontally acquired biosynthetic genes against a fungal pathogen. *Nature Communications 2024 15:1* 15(1), pp. 1–17. Available at: https://www.nature.com/articles/s41467-024-49919-1 [Accessed: 4 September 2025].

Nowińska, A., Franielczyk-Pietyra, B. and Polhemus, D.A. 2023. The Leg Sensilla of Insects from Different Habitats—Comparison of Strictly Aquatic and Riparian Bugs (Corixidae, Ochteridae, Gelastocoridae: Nepomorpha: Insecta: Heteroptera). *Insects* 14(5). doi: 10.3390/INSECTS14050441.

Nylin, S., Slove, J. and Janz, N. 2014. HOST PLANT UTILIZATION, HOST RANGE OSCILLATIONS AND DIVERSIFICATION IN NYMPHALID BUTTERFLIES: A PHYLOGENETIC INVESTIGATION. *Evolution* 68(1), pp. 105–124. Available at: https://dx.doi.org/10.1111/evo.12227 [Accessed: 3 September 2025].

O'Brien, J., Hayder, H., Zayed, Y. and Peng, C. 2018. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Frontiers in Endocrinology* 9(AUG), p. 402. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC6085463/ [Accessed: 16 September 2025].

Ohno, S. 1970a. Evolution by Gene Duplication. *Evolution by Gene Duplication*, pp. 59–87. doi: 10.1007/978-3-642-86659-3.

Ohno, S. 1970b. Evolution by Gene Duplication. *Evolution by Gene Duplication*. doi: 10.1007/978-3-642-86659-3.

Van Oss, S.B. and Carvunis, A.R. 2019. De novo gene birth. *PLoS Genetics* 15(5). Available at: https://pubmed.ncbi.nlm.nih.gov/31120894/ [Accessed: 3 September 2025].

Ota, K.G., Kuraku, S. and Kuratani, S. 2007. Hagfish embryology with reference to the evolution of the neural crest. *Nature* 446(7136), pp. 672–675. Available at: https://pubmed.ncbi.nlm.nih.gov/17377535/ [Accessed: 3 September 2025].

Otak, J.M. 2020. Morphological and Spatial Diversity of the Discal Spot on the Hindwings of Nymphalid Butterflies: Revision of the Nymphalid Groundplan. *Insects 2020, Vol. 11, Page 654* 11(10), p. 654. Available at: https://www.mdpi.com/2075-4450/11/10/654/htm [Accessed: 1 September 2025].

Panganiban, G. et al. 1997. The origin and evolution of animal appendages. *Proceedings of the National Academy of Sciences of the United States of America* 94(10), pp. 5162–5166. Available at: /doi/pdf/10.1073/pnas.94.10.5162?download=true [Accessed: 1 September 2025].

Paps, J. and Holland, P.W.H. 2018. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. *Nature Communications* 9(1), pp. 1–8. Available at: https://www.nature.com/articles/s41467-018-04136-5 [Accessed: 3 September 2025].

Payne, J.L., Bachan, A., Heim, N.A., Hull, P.M. and Knope, M.L. 2020. The evolution of complex life and the stabilization of the Earth system. *Interface Focus* 10(4), p. 20190106. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC7333899/ [Accessed: 2 September 2025].

Van De Peer, Y., Mizrachi, E. and Marchal, K. 2017. The evolutionary significance of polyploidy. *Nature Reviews Genetics* 18(7), pp. 411–424. Available at: https://pubmed.ncbi.nlm.nih.gov/28502977/ [Accessed: 9 July 2025].

Pennisi, E. 2021. The simplest of slumbers. *Science* 374(6567), pp. 526–529. Available at: https://pubmed.ncbi.nlm.nih.gov/34709907/ [Accessed: 9 July 2025].

Penz, C.M. and Krenn, H.W. 2000. Behavioral adaptations to pollen-feeding in Heliconius butterflies (nymphalidae, heliconiinae): An experiment using Latana flowers. *Journal of Insect Behavior* 13(6), pp. 865–880. doi: 10.1023/A:1007814618149.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T. and Salzberg, S.L. 2015. StringTie enables improved reconstruction of a transcriptome from RNAseq reads. *Nature Biotechnology* 33(3), pp. 290–295. Available at: https://www.nature.com/articles/nbt.3122 [Accessed: 29 August 2025].

Peter Jurtshuk, Jr. 1996. Bacterial Metabolism. *Medical Microbiology*. Available at: https://www.ncbi.nlm.nih.gov/books/NBK7919/ [Accessed: 4 September 2025].

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E. 2004. UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* 25(13), pp. 1605–1612. Available at: https://pubmed.ncbi.nlm.nih.gov/15264254/ [Accessed: 3 September 2025].

Popham, E.J. 1961. The function of the paleal pegs of corixidae (Hemiptera heteroptera). *Nature* 190(4777), pp. 742–743. Available at: https://www.nature.com/articles/190742a0 [Accessed: 1 September 2025].

Posit team. 2025. RStudio: Integrated Development Environment for R.

Poulton, E.B. and Poulton, E.B. 1890. *The colours of animals : their meaning and use, especially considered in the case of insects*. New York: D. Appleton. Available at:

https://www.biodiversitylibrary.org/bibliography/69899 [Accessed: 3 September 2025].

Powell, J.A. 2009. Lepidoptera: Moths, Butterflies. *Encyclopedia of Insects*, pp. 559–587. Available at: https://www.sciencedirect.com/science/article/abs/pii/B9780123741448001600 [Accessed: 31 August 2025].

Qiao, X., Yin, H., Li, L., Wang, R., Wu, J., Wu, J. and Zhang, S. 2018. Different modes of gene duplication show divergent evolutionary patterns and contribute differently to the expansion of gene families involved in important fruit traits in pear (Pyrus bretschneideri). *Frontiers in Plant Science* 9, p. 333415. Available at: www.frontiersin.org [Accessed: 9 July 2025].

Ramulu, H.G., Raoult, D. and Pontarotti, P. 2012. The rhizome of life: what about metazoa? *Frontiers in cellular and infection microbiology* 2, p. 50. Available at: www.frontiersin.org [Accessed: 18 September 2025].

Ranz, J.M. et al. 2021. A de novo transcriptional atlas in Danaus plexippus reveals variability in dosage compensation across tissues. *Communications Biology 2021 4:1* 4(1), pp. 1–13. Available at: https://www.nature.com/articles/s42003-021-02335-3 [Accessed: 3 September 2025].

Rao, J.N. and Wang, J.-Y. 2010. Intestinal Architecture and Development. Available at: https://www.ncbi.nlm.nih.gov/books/NBK54098/ [Accessed: 29 August 2025].

Rastogi, S. and Liberles, D.A. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evolutionary Biology* 5(1), pp. 1–7. Available at: https://bmcecolevol.biomedcentral.com/articles/10.1186/1471-2148-5-28 [Accessed: 9 July 2025].

Rawlings, N.D. and Barrett, A.J. 1994. Families of serine peptidases. *Methods in Enzymology* 244(C), pp. 19–61. Available at: https://pubmed.ncbi.nlm.nih.gov/7845208/ [Accessed: 3 September 2025].

Rebeiz, M., Jikomes, N., Kassner, V.A. and Carroll, S.B. 2011. Evolutionary origin of a novel gene expression pattern through co-option of the latent activities of existing regulatory sequences. *Proceedings of the National Academy of Sciences of the United States of America* 108(25), pp. 10036–10043. Available at: /doi/pdf/10.1073/pnas.1105937108?download=true [Accessed: 4 September 2025].

Redmond, A.K. 2024. Acoelomorph flatworm monophyly is a long-branch attraction artefact obscuring a clade of Acoela and Xenoturbellida. *Proceedings of the Royal Society B* 291(2031). Available at: /doi/pdf/10.1098/rspb.2024.0329 [Accessed: 6 October 2025].

Reinwald, C., Bauder, J.A.S., Karolyi, F., Neulinger, M., Jaros, S., Metscher, B. and Krenn, H.W. 2022. Evolutionary functional morphology of the proboscis and feeding apparatus of hawk moths (Sphingidae: Lepidoptera). *Journal of Morphology* 283(11), p. 1390. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC9825987/ [Accessed: 31 August 2025].

Renwick, J.A.A. and Chew, F.S. 1994. Oviposition behavior in lepidoptera. *Annual Review of Entomology* 39(1), pp. 377–400. doi: 10.1146/ANNUREV.EN.39.010194.002113.

Retallack, G.J. 2013. Ediacaran life on land. *Nature* 493(7430), pp. 89–92. Available at: https://pubmed.ncbi.nlm.nih.gov/23235827/ [Accessed: 9 July 2025].

Richter, D.J., Fozouni, P., Eisen, M.B. and King, N. 2018. Gene family innovation, conservation and loss on the animal stem lineage. *eLife* 7. doi: 10.7554/ELIFE.34226.

Riley, C.L., Oostra, V. and Plaistow, S.J. 2023. Does the definition of a novel environment affect the ability to detect cryptic genetic variation? *Journal of Evolutionary Biology* 36(11), pp. 1618–1629. Available at: https://dx.doi.org/10.1111/jeb.14238 [Accessed: 31 August 2025].

Rödelsperger, C., Prabh, N. and Sommer, R.J. 2019. New Gene Origin and Deep Taxon Phylogenomics: Opportunities and Challenges. *Trends in Genetics* 35(12), pp. 914–922. Available at: https://pubmed.ncbi.nlm.nih.gov/31610892/ [Accessed: 3 September 2025].

Roelofs, D. et al. 2020. Multi-faceted analysis provides little evidence for recurrent whole-genome duplications during hexapod evolution. *BMC Biology* 18(1), pp. 1–13. Available at: https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-020-00789-1 [Accessed: 1 October 2025].

Rota, J., Twort, V., Chiocchio, A., Peña, C., Wheat, C.W., Kaila, L. and Wahlberg, N. 2022. The unresolved phylogenomic tree of butterflies and moths (Lepidoptera): Assessing the potential causes and consequences. *Systematic Entomology* 47(4), pp. 531–550. Available at: https://resjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/syen.12545 [Accessed: 3 September 2025].

Royal Ontario Museum. 2025. *The Tree of Life - The Burgess Shale*. Available at: https://burgess-shale.rom.on.ca/science/origin-of-animals-and-the-cambrian-explosion/the-tree-of-life/ [Accessed: 31 August 2025].

Ruiz-Narváez, E.A. 2013. Use of alternative promoters may hide genetic effects on phenotypic traits. *Journal of Human Genetics* 58(1), pp. 47–50. Available at: https://www.nature.com/articles/jhg2012115 [Accessed: 16 September 2025].

Saco, A., Rey-Campos, M., Rosani, U., Novoa, B. and Figueras, A. 2021. The Evolution and Diversity of Interleukin-17 Highlight an Expansion in Marine Invertebrates and Its Conserved Role in Mucosal Immunity. *Frontiers in Immunology* 12, p. 692997. Available at: www.frontiersin.org [Accessed: 16 September 2025].

Sankar, A., Mohammad, F., Sundaramurthy, A.K., Wang, H., Lerdrup, M., Tatar, T. and Helin, K. 2022. Histone editing elucidates the functional roles of H3K27 methylation and acetylation in mammals. *Nature Genetics* 54(6), pp. 754–760. Available at: https://pubmed.ncbi.nlm.nih.gov/35668298/ [Accessed: 30 August 2025].

Santos, M.E., Le Bouquin, A., Crumiere, A.J.J. and Khila, A. 2017. Taxon-restricted genes at the origin of a novel trait allowing access to a new environment. *Science* 358(6361), pp. 386–390. Available at: https://pubmed.ncbi.nlm.nih.gov/29051384/ [Accessed: 3 September 2025].

Scannell, D.R., Butler, G. and Wolfe, K.H. 2007. Yeast genome evolution - The origin of the species. *Yeast* 24(11), pp. 929–942. Available at: https://pubmed.ncbi.nlm.nih.gov/17621376/ [Accessed: 9 July 2025].

Schierwater, B. and DeSalle, R. 2021. Invertebrate Zoology: A Tree of Life Approach. *Invertebrate Zoology: A Tree of Life Approach*, pp. 1–627. doi: 10.1201/9780429159053/INVERTEBRATE-ZOOLOGY-BERND-SCHIERWATER-ROB-DESALLE/RIGHTS-AND-PERMISSIONS.

van Schooten, B., Meléndez-Rosa, J., van Belleghem, S.M., Jiggins, C.D., Tan, J.D., Owen McMillan, W. and Papa, R. 2020. Divergence of chemosensing during the early stages of speciation. *Proceedings of the National Academy of Sciences of the United States of America* 117(28), pp. 16438–16447. Available at: /doi/pdf/10.1073/pnas.1921318117?download=true [Accessed: 3 September 2025].

Schreiber, A.M. 2006. Asymmetric craniofacial remodeling and liberalized behavior in larval flatfish. *Journal of Experimental Biology* 209(4), pp. 610–621. Available at: https://pubmed.ncbi.nlm.nih.gov/16449556/ [Accessed: 31 August 2025].

Scoble, M.J.. 1995. The lepidoptera : form, function and diversity. p. 404. Available at: https://search.worldcat.org/title/25282932 [Accessed: 3 September 2025].

Sémon, M. and Wolfe, K.H. 2008a. Preferential subfunctionalization of slow-evolving genes after allopolyploidization in Xenopus laevis. *Proceedings of the National Academy of Sciences of the United States of America* 105(24), pp. 8333–8338. Available at: /doi/pdf/10.1073/pnas.0708705105?download=true [Accessed: 9 July 2025].

Sémon, M. and Wolfe, K.H. 2008b. Preferential subfunctionalization of slow-evolving genes after allopolyploidization in Xenopus laevis. *Proceedings of the National Academy of Sciences of the United States of America* 105(24), pp. 8333–8338. Available at: https://pubmed.ncbi.nlm.nih.gov/18541921/ [Accessed: 25 August 2025].

Serif. 2024. Affinity Designer2.

Shirai, L.T. et al. 2012. Evolutionary history of the recruitment of conserved developmental genes in association to the formation and diversification of a novel trait. *BMC Evolutionary Biology* 12(1), pp. 1–11. Available at: https://bmcecolevol.biomedcentral.com/articles/10.1186/1471-2148-12-21 [Accessed: 30 August 2025].

Shirvaliloo, M. 2022. The landscape of histone modifications in epigenomics since 2020. *Epigenomics* 14(23), pp. 1465–1477. Available at: https://pubmed.ncbi.nlm.nih.gov/36710634/ [Accessed: 30 August 2025].

Signor, S.A. and Nuzhdin, S. V. 2018. The evolution of gene expression in cis and trans. *Trends in genetics : TIG* 34(7), p. 532. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC6094946/ [Accessed: 13 September 2025].

Slingsby, C., Wistow, G.J. and Clark, A.R. 2013. Evolution of crystallins for a role in the vertebrate eye lens. *Protein Science : A Publication of the Protein Society* 22(4), p. 367. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC3610043/ [Accessed: 30 August 2025].

Soltis, D.E. et al. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* 96(1), pp. 336–348. Available at: https://pubmed.ncbi.nlm.nih.gov/21628192/ [Accessed: 9 July 2025].

Sommer, R.J. 2020. Phenotypic Plasticity: From Theory and Genetics to Current and Future Challenges. *Genetics* 215(1), p. 1. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC7198268/ [Accessed: 31 August 2025].

Sommer, R.J., Dardiry, M., Lenuzzi, M., Namdeo, S., Renahan, T., Sieriebriennikov, B. and Werner, M.S. 2017. The genetics of phenotypic plasticity in nematode feeding structures. *Open Biology* 7(3), p. 160332. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC5376706/ [Accessed: 31 August 2025].

Sparks, M.E., Blackburn, M.B., Kuhar, D. and Gundersen-Rindal, D.E. 2013. Transcriptome of the Lymantria dispar (Gypsy Moth) Larval Midgut in Response to Infection by Bacillus thuringiensis. *PLOS ONE* 8(5), p. e61190. Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061190 [Accessed: 3 September 2025].

Steenwyk, J.L., Buida, T.J., Labella, A.L., Li, Y., Shen, X.X. and Rokas, A. 2021. PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data.

Bioinformatics 37(16), pp. 2325–2331. Available at: https://dx.doi.org/10.1093/bioinformatics/btab096 [Accessed: 29 August 2025].

Stelkens, R.B., Schmid, C., Selz, O. and Seehausen, O. 2009. Phenotypic novelty in experimental hybrids is predicted by the genetic distance between species of cichlid fish. *BMC Evolutionary Biology* 9(1), pp. 1–13. Available at: https://bmcecolevol.biomedcentral.com/articles/10.1186/1471-2148-9-283 [Accessed: 31 August 2025].

Stern, R.J. 2016. Is plate tectonics needed to evolve technological species on exoplanets? *Geoscience Frontiers* 7(4), pp. 573–580. Available at: https://www.sciencedirect.com/science/article/pii/S1674987115300062 [Accessed: 4 September 2025].

Stern, R.J. and Gerya, T. V. 2023. Co-Evolution of Life and Plate Tectonics: The Biogeodynamic Perspective on the Mesoproterozoic-Neoproterozoic Transitions. *Dynamics of Plate Tectonics and Mantle Convection*, pp. 295–319. doi: 10.1016/B978-0-323-85733-8.00013-5.

Suzuki David T, Griffiths Anthony J.F. and Miller Jeffrey H. 2000. *An Introduction to Genetic Analysis Page 1 of 1*. 7th ed. W. H. Freeman Publishers.

Swanson, B.O., George, M.N., Anderson, S.P. and Christy, J.H. 2013. Evolutionary variation in the mechanics of fiddler crab claws. *BMC Evolutionary Biology* 13(1), pp. 1–11. Available at: https://bmcecolevol.biomedcentral.com/articles/10.1186/1471-2148-13-137 [Accessed: 31 August 2025].

Takahashi, A. and Ohnishi, T. 2004. The significance of the study about the biological effects of solar ultraviolet radiation using the Exposed Facility on the International Space Station. *Biological sciences in space = Uchū seibutsu kagaku* 18(4), pp. 255–260. Available at: https://pubmed.ncbi.nlm.nih.gov/15858393/ [Accessed: 4 September 2025].

Thiele, S.C., Rodrigues, D. and Moreira, G.R.P. 2016. Oviposition in Heliconius erato (Lepidoptera, Nymphalidae): how Essential Is Drumming Behavior for Host-Plant Selection? *Journal of Insect Behavior* 29(3), pp. 283–300. Available at: https://link.springer.com/article/10.1007/s10905-016-9559-z [Accessed: 1 September 2025].

Thomas, G.W.C. et al. 2020. Gene content evolution in the arthropods. *Genome Biology* 21(1), pp. 1–14. Available at: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1925-7 [Accessed: 3 September 2025].

Tian, S. et al. 2024. A micro-RNA is the effector gene of a classic evolutionary hotspot locus. *bioRxiv : the preprint server for biology*. Available at: https://pubmed.ncbi.nlm.nih.gov/38659873/ [Accessed: 3 September 2025].

Train, C.M., Pignatelli, M., Altenhoff, A. and Dessimoz, C. 2018. iHam and pyHam: visualizing and processing hierarchical orthologous groups. *Bioinformatics* 35(14), p. 2504. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC6612847/ [Accessed: 25 August 2025].

Tsai, I.J. et al. 2013. The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* 496(7443), pp. 57–63. Available at: https://pubmed.ncbi.nlm.nih.gov/23485966/ [Accessed: 25 August 2025].

Upham, N.S., Esselstyn, J.A. and Jetz, W. 2019. Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLOS Biology*

17(12), p. e3000494. Available at: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000494 [Accessed: 9 July 2025].

Urban, J.M. et al. 2021. High contiguity de novo genome assembly and DNA modification analyses for the fungus fly, Sciara coprophila, using single-molecule sequencing. *BMC Genomics 2021 22:1* 22(1), pp. 1–23. Available at: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-021-07926-2 [Accessed: 3 September 2025].

Usinger, R.L. 1956. *Aquatic insects of California, with keys to North American genera and California species*. Berkeley: University of California Press. Available at: https://www.biodiversitylibrary.org/bibliography/61952 [Accessed: 1 September 2025].

Velicky, P. et al. 2018. Genome amplification and cellular senescence are hallmarks of human placenta development. *PLoS Genetics* 14(10). Available at: https://pubmed.ncbi.nlm.nih.gov/30312291/ [Accessed: 9 July 2025].

Vianello, S.D. et al. 2025. Deconstructing the common anteroposterior organisation of adult bilaterian guts. *bioRxiv*, p. 2025.07.02.662275. Available at: https://www.biorxiv.org/content/10.1101/2025.07.02.662275v1 [Accessed: 29 September 2025].

Wahlberg, N., Leneveu, J., Kodandaramaiah, U., Pena, C., Nylin, S., Freitas, A.V.L. and Brower, A.V.Z. 2009. Nymphalid butterflies diversify following near demise at the Cretaceous/Tertiary boundary. *Proceedings of the Royal Society B: Biological Sciences* 276(1677), pp. 4295–4302. Available at: /doi/pdf/10.1098/rspb.2009.1303 [Accessed: 1 September 2025].

Wald, G. 1934. Carotenoids and the vitamin A cycle in vision [8]. *Nature* 134(3376), p. 65. Available at: https://www.nature.com/articles/134065a0 [Accessed: 3 September 2025].

Wang, J., Zhang, W., Engel, M.S., Sheng, X., Shih, C. and Ren, D. 2022. Early evolution of wing scales prior to the rise of moths and butterflies. *Current Biology* 32(17), pp. 3808-3814.e2. Available at: https://www.sciencedirect.com/science/article/pii/S0960982222010910 [Accessed: 31 August 2025].

Waterhouse, R.M., Tegenfeldt, F., Li, J., Zdobnov, E.M. and Kriventseva, E. V. 2012. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Research* 41(Database issue), p. D358. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC3531149/ [Accessed: 16 September 2025].

Webster, M. and Zelditch, M.L. 2005. Evolutionary modifications of ontogeny: heterochrony and beyond. *Paleobiology* 31(3), pp. 354–372. Available at: https://www.cambridge.org/core/journals/paleobiology/article/abs/evolutionary-modifications-of-ontogeny-heterochrony-and-beyond/8B7B393EC72A2F2243955717153F8E2E [Accessed: 31 August 2025].

Weisman, C.M., Murray, A.W. and Eddy, S.R. 2020. Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLOS Biology* 18(11), p. e3000862. Available at: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000862 [Accessed: 29 August 2025].

Weisman, C.M., Murray, A.W. and Eddy, S.R. 2022. Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. *Current Biology* 32(12), pp. 2632-2639.e2. Available at: https://pubmed.ncbi.nlm.nih.gov/35588743/ [Accessed: 3 September 2025].

West-Eberhard, M.J. 2003. Heterotopy. *Developmental Plasticity and Evolution*. Available at: https://academic.oup.com/book/40908/chapter/349015232 [Accessed: 31 August 2025].

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag.

Willen, E. 2003. A new species of Stenhelia (Copepoda, Harpacticoida) from a hydrothermal, active, submarine volcano in the New Ireland Fore-Arc system (Papua New Guinea) with notes on deep sea colonization within the Stenheliinae. *Journal of Natural History* 37(14), pp. 1691–1711. Available at: https://www.tandfonline.com/doi/pdf/10.1080/00222930110114437 [Accessed: 4 September 2025].

Wolfe, J.M., Oliver, J.C. and Monteiro, A. 2011. Evolutionary Reduction of the First Thoracic Limb in Butterflies. *Journal of Insect Science* 11, p. 66. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC3281478/ [Accessed: 1 September 2025].

Wörheide, G. et al. 2012. Deep Phylogeny and Evolution of Sponges (Phylum Porifera). *Advances in Marine Biology* 61, pp. 1–78. Available at: https://pubmed.ncbi.nlm.nih.gov/22560777/ [Accessed: 9 July 2025].

Worsaae, K., Vinther, J. and Sørensen, M.V. 2023. Evolution of Bilateria from a Meiofauna Perspective—Miniaturization in the Focus. *New Horizons in Meiobenthos Research: Profiles, Patterns and Potentials*, pp. 1–31. Available at: https://link.springer.com/chapter/10.1007/978-3-031-21622-0_1 [Accessed: 9 July 2025].

Wright, C.J., Stevens, L., Mackintosh, A., Lawniczak, M. and Blaxter, M. 2024. Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera. *Nature Ecology & Evolution 2024 8:4* 8(4), pp. 777–790. Available at: https://www.nature.com/articles/s41559-024-02329-4 [Accessed: 31 August 2025].

Wright, J.P., Jones, C.G. and Flecker, A.S. 2002. An ecosystem engineer, the beaver, increases species richness at the landscape scale. *Oecologia* 132(1), pp. 96–101. Available at: https://pubmed.ncbi.nlm.nih.gov/28547281/ [Accessed: 31 August 2025].

Wu, D.D. and Zhang, Y.P. 2013. Evolution and function of de novo originated genes. *Molecular Phylogenetics and Evolution* 67(2), pp. 541–545. Available at: https://pubmed.ncbi.nlm.nih.gov/23454495/ [Accessed: 9 July 2025].

Wu, L. and Lambert, J.D. 2023. Clade-specific genes and the evolutionary origin of novelty; new tools in the toolkit. *Seminars in Cell & Developmental Biology* 145, pp. 52–59. Available at: https://www.sciencedirect.com/science/article/pii/S1084952122001835?casa_token=HDFOL2Gmt18AAAAA:uDd0AmIUfrmeYATvZ0rJ7CXdW9J-duoh1Sx4nLdfMhrIo8V0-hroG5ZoVI_rVQ3r59gIGlb_g1w [Accessed: 13 August 2025].

Wu, N. et al. 2022. Widespread Gene Expression Divergence in Butterfly Sensory Tissues Plays a Fundamental Role During Reproductive Isolation and Speciation. *Molecular Biology and Evolution* 39(11). Available at: https://dx.doi.org/10.1093/molbev/msac225 [Accessed: 3 September 2025].

Wu, P.H. and Zamore, P.D. 2021. Defining the functions of PIWI-interacting RNAs. *Nature Reviews Molecular Cell Biology* 22(4), pp. 239–240. Available at:

https://www.nature.com/articles/s41580-021-00336-y [Accessed: 16 September 2025].

Wucherpfennig, J.I. et al. 2022. Evolution of stickleback spines through independent cis-regulatory changes at HOXDB. *Nature Ecology and Evolution* 6(10), pp. 1537–1552. Available at: https://pubmed.ncbi.nlm.nih.gov/36050398/ [Accessed: 3 September 2025].

Wybouw, N., Pauchet, Y., Heckel, D.G. and Leeuwen, T. Van. 2016. Horizontal Gene Transfer Contributes to the Evolution of Arthropod Herbivory. *Genome Biology and Evolution* 8(6), pp. 1785–1801. Available at: https://dx.doi.org/10.1093/gbe/evw119 [Accessed: 2 September 2025].

Xia, S. et al. 2021. Genomic analyses of new genes and their phenotypic effects reveal rapid evolution of essential functions in Drosophila development. *PLOS Genetics* 17(7). Available at: https://knowledge.uchicago.edu/record/5948 [Accessed: 30 August 2025].

Xu, G.F., Gong, C.C., Lyu, H., Deng, H.M. and Zheng, S.C. 2022. Dynamic transcriptome analysis of Bombyx mori embryonic development. *Insect Science* 29(2), pp. 344–362. Available at: https://pubmed.ncbi.nlm.nih.gov/34388292/ [Accessed: 3 September 2025].

Yanai, I., Peshkin, L., Jorgensen, P. and Kirschner, M.W. 2011. Mapping Gene Expression in Two Xenopus Species: Evolutionary Constraints and Developmental Flexibility. *Developmental Cell* 20(4), pp. 483–496. Available at: https://www.sciencedirect.com/science/article/pii/S1534580711001213 [Accessed: 31 August 2025].

Yokoi, K. et al. 2021. Reference Transcriptome Data in Silkworm Bombyx mori. *Insects 2021, Vol. 12, Page 519* 12(6), p. 519. Available at: https://www.mdpi.com/2075-4450/12/6/519/htm [Accessed: 3 September 2025].

Young, F.J. and Montgomery, S.H. 2020. Pollen feeding in Heliconius butterflies: The singular evolution of an adaptive suite: Pollen feeding in Heliconius butterflies. *Proceedings of the Royal Society B: Biological Sciences* 287(1938). doi: 10.1098/RSPB.2020.1304.

Yu, G., Smith, D.K., Zhu, H., Guan, Y. and Lam, T.T.Y. 2017. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 8(1), pp. 28–36. Available at: /doi/pdf/10.1111/2041-210X.12628 [Accessed: 25 August 2025].

Yung, P.Y.K., Stuetzer, A., Fischle, W., Martinez, A.M. and Cavalli, G. 2015. Histone H3 Serine 28 Is Essential for Efficient Polycomb-Mediated Gene Repression in Drosophila. *Cell Reports* 11(9), pp. 1437–1445. doi: 10.1016/J.CELREP.2015.04.055.

Zahnle, K., Schaefer, L. and Fegley, B. 2010. Earth's earliest atmospheres. *Cold Spring Harbor perspectives in biology* 2(10). Available at: https://pubmed.ncbi.nlm.nih.gov/20573713/ [Accessed: 9 July 2025].

Zhang, J. 2003. Evolution by gene duplication: an update. *Trends in Ecology & Evolution* 18(6), pp. 292–298. Available at: https://www.sciencedirect.com/science/article/pii/S0169534703000338 [Accessed: 9 July 2025].

Zhang, R., Guo, C., Shan, H. and Kong, H. 2014. Developmental repatterning and biodiversity. *Biodiversity Science* 22(1), p. 66. Available at: https://www.biodiversity-science.net/EN/10.3724/SP.J.1003.2014.13248 [Accessed: 31 August 2025].

Zhang, X. and Shu, D. 2021. Current understanding on the Cambrian Explosion: questions and answers. *PalZ* 95(4), pp. 641–660. doi: 10.1007/S12542-021-00568-5.

Zhao, L., Svetec, N. and Begun, D.J. 2024. De Novo Genes. *Annual Review of Genetics* 58(1), pp. 211–232. Available at: https://pubmed.ncbi.nlm.nih.gov/39088850/ [Accessed: 25 August 2025].

Zhou, S., Chen, Y., Guo, C. and Qi, J. 2020. PhyloMCL: Accurate clustering of hierarchical orthogroups guided by phylogenetic relationship and inference of polyploidy events. *Methods in Ecology and Evolution* 11(8), pp. 943–954. Available at: /doi/pdf/10.1111/2041-210X.13401 [Accessed: 16 September 2025].

Zhuravlev, A.Y. and Wood, R.A. 2018. The two phases of the Cambrian Explosion. *Scientific Reports* 8(1), pp. 1–10. Available at: https://www.nature.com/articles/s41598-018-34962-y [Accessed: 13 August 2025].

Appendices

Appendix A: List of Abbreviations

This appendix contains non-S.I. abbreviations used in this thesis. Gene names, protein names and species names are excluded.

AI - Artificial Intelligence
BLAST - Basic Local Alignment Search Tool
CSP - Chemosensory proteins
DE - Differentially expressed
DGE - Differential gene expression
DNA - Deoxyribonucleic acid
DToL - Darwin Tree of Life project
FC - Fold change
FPKM - Fragments Per Kilobase per Million mapped reads
GO - Gene ontology
GR - Gustatory receptors
HGT - Horizontal gene transfer
HOG - Hierarchical orthogroup
IR - Ionotropic receptors
LCA - Last common ancestor
MRCA - Most recent common ancestor
mRNA - Messenger RNA
NCBI - National Centre for Biotechnology Information
nr - NCBI non-redundant database
OBP - Olfactory binding proteins
OR - Olfactory receptors
PCA - Principal component analysis
PCR - Polymerase chain reaction
piRNA - Piwi-interacting RNA
QC - Quality control
RNA - Ribonucleic acid
RNAseq - RNA sequencing
sncRNAs - Small noncoding RNAs
SNMP - Sensory neuron membrane proteins
T1 - Prothoracic legs
T2 - Midlegs / mesothoracic legs
T3 - Metathoracic legs
TE - Transposable elements
WGD - Whole genome duplication
WT - Wild type

Appendix B: Resulting publications

Hoile et al. BMC Genomics (2025) 26:161
https://doi.org/10.1186/s12864-025-11338-x

RESEARCH Open Access

Gene novelty and gene family expansion in the early evolution of Lepidoptera

Asia E. Hoile[1], Peter W. H. Holland[1*] and Peter O. Mulhair[1*]

Abstract

Background Almost 10% of all known animal species belong to Lepidoptera: moths and butterflies. To understand how this incredible diversity evolved we assess the role of gene gain in driving early lepidopteran evolution. Here, we compared the complete genomes of 115 insect species, including 99 Lepidoptera, to search for novel genes coincident with the emergence of Lepidoptera.

Results We find 217 orthogroups or gene families which emerged on the branch leading to Lepidoptera; of these 177 likely arose by gene duplication followed by extensive sequence divergence, 2 are candidates for origin by horizontal gene transfer, and 38 have no known homology outside of Lepidoptera and possibly arose via de novo gene genesis. We focus on two new gene families that are conserved across all lepidopteran species and underwent extensive duplication, suggesting important roles in lepidopteran biology. One encodes a family of sugar and ion transporter molecules, potentially involved in the evolution of diverse feeding behaviours in early Lepidoptera. The second encodes a family of unusual propeller-shaped proteins that likely originated by horizontal gene transfer from Spiroplasma bacteria; we name these the Lepidoptera propellin genes.

Conclusion We provide the first insights into the role of genetic novelty in the early evolution of Lepidoptera. This gives new insight into the rate of gene gain during the evolution of the order as well as providing context on the likely mechanisms of origin. We describe examples of new genes which were retained and duplicated further in all lepidopteran species, suggesting their importance in Lepidoptera evolution.

Keywords Insect evolution, HGT, Gene duplication, Genome evolution

Background

Diversification and adaptation depend on genetic change but associating genomic drivers underpinning phenotypic change is challenging. Many studies have approached this problem by starting with phenotypic polymorphisms within a species or differences between closely related species and then using genomic and experimental approaches to identify underlying causative mutations. Several of these studies have uncovered sequence changes in non-coding DNA affecting the expression of conserved genes [1–4]. Other studies have identified coding sequence changes causing amino acid substitutions, or loss of function, as causative mutations that were subsequently fixed under selection [5–7]. It is clear, however, that changes in existing genes, whether they affect gene expression or protein sequence, cannot explain all adaptive evolution. Perhaps the best evidence lies in comparative genomics: when genome sequences are compared ample evidence is uncovered for the role of gene number variation, gene duplications, and gene novelty in driving evolution and adaptation [8–13].

Gene novelty is a multi-faceted concept [14–17]. We define novel genes as protein-coding loci that are

*Correspondence:
Peter W. H. Holland
peter.holland@biology.ox.ac.uk
Peter O. Mulhair
peter.mulhair@biology.ox.ac.uk
[1] Department of Biology, University of Oxford, Mansfield Road, Oxford OX1 3SZ, UK

lineage-specific (i.e. taxonomically restricted genes), without close homologues in other taxa [18]. This is a pragmatic definition rather than a mechanistic one since we cannot always determine the mechanism by which a novel gene arose. The mode of origin of taxonomically restricted genes might be gene duplication followed by extensive sequence divergence [19–24], fusion of distinct loci or a transposable element into a pre-existing locus [25–30], horizontal gene transfer [31–33] or de novo origin from non-coding DNA [17, 34–36]. Whatever the mode of origin, novel genes likely reflect novel biology as they will encode proteins with potentially distinct activity or function not present in the outgroup taxa. Examples in arthropods include horizontally acquired genes from bacteria underpinning adaptations to phytophagy [37] or male courtship behaviour in moths and butterflies [32], and divergent gene duplicates recruited for limb patterning in water striders [38].

Here we investigate the origin of novel genes in the early evolution of the insect order Lepidoptera. Lepidoptera are a holometabolous order of insects consisting of the moths and butterflies and comprise nearly 160,000 described species or 8–10% of known animal species on the planet [39]. The oldest members of the Lepidoptera crown group are estimated to have appeared in the Late Carboniferous (~300 mya) and were likely pollen feeders, with the evolution of a tube-like proboscis and nectar feeding occurring later in the Middle Triassic (~240 Ma). Today the Lepidoptera inhabit almost all terrestrial ecosystems, displaying a large variety of ecological adaptations relating to feeding, defence, and survival [39–41]. Larvae of the earliest lineages were likely endophagous, feeding internally in the tissue of nonvascular land plants, with adults possessing mandibulate chewing mouthparts (as seen in extant members of the family Micropterigidae) suitable for pollen feeding [42, 43]. A period of diversification early in the evolution of Lepidoptera coincided with the development of the tube-like proboscis, used by adults to feed on nectar, and the expansion of angiosperms. The remarkable diversity present in Lepidoptera today can be attributed to continued co-evolution with diverse angiosperm lineages, major transitions in morphology and habitat, and the emergence of diverse feeding behaviours [41].

To assess whether novel genes arose in the early evolution of Lepidoptera, and whether any of these underwent further gene family expansion, we require complete genome sequences from a dense sampling of Lepidoptera and related insect orders. Previous studies have constructed deep-level phylogenies of Lepidoptera using a large density of species but relatively few loci [39], while other studies have studied specific gene families in depth [44–46]. Large genomic datasets have only recently become available through sequencing consortia such as the Darwin Tree of Life Project [47] affiliated to the Earth Biogenome Project [48]. Here, we avail of this data by analysing 115 high quality insect genomes and identify 217 novel genes that arose on the stem lineage of Lepidoptera and 541 novel genes that arose on the stem lineage of the Ditrysia, a major clade encompassing most of lepidopteran diversity [49]. We infer the likely modes of origin for these novel genes. We then focus attention on two gene families gained on the ancestral lepidopteran branch that were subsequently retained across all species, suggestive of recruitment to important roles in lepidopteran biology. One is a gene family encoding divergent sugar transporter proteins; the other is a likely horizontal gene transfer from bacteria.

Materials and methods

Gene family construction and discovery of novel genes

Proteome data from 99 species of Lepidoptera and 16 other arthropod species (Supplementary Table S1) were obtained from Ensembl Rapid Release (rapid.ensembl.org; accessed February 2023); taxon sampling was based on obtaining robust phylogenetic coverage across Lepidoptera while also preferentially selecting species with proteome predictions based on the Ensembl genebuild annotation pipeline (i.e. annotation which incorporated RNA sequence data). Primary transcripts were obtained from the predicted proteome data and Orthofinder v2.3.14 was run with default parameters to determine orthogroups within the dataset [50]. To relate these to a species tree, amino acid sequences from 25 single copy orthologues present in all species, as obtained from the Orthofinder output, were aligned using MAFFT v7.505 [51], trimmed using trimAl v1.4.rev15 build [52], and concatenated with PhyKIT [53]. This concatenated alignment was used to generate a species tree using IQ-TREE version 2.0-rc1 with 1000 bootstrap iterations, the given model LG+G4 and option -nt AUTO which automatically determines the best number of cores given the current data and computer capacity [54]. Orthogroups gained at nodes of interest (i.e. the branch leading to Lepidoptera and the branch leading to Ditrysia) were extracted using Orthoparser (github.com/PeterMulhair/ortho_parser). To test further whether orthogroups inferred by the analysis to be specific to Lepidoptera were actually present in outgroups but missing from predicted proteomes, Trichoptera genomes annotated by the alternative Augustus-Gaius pipeline (BRAKER) [55] were analysed. This was carried out using a BLASTp search of the orthogroups against the trichopteran BRAKER proteomes to find any potential missing homologues (using an e-value cutoff of 1e-5 and filtering hits above 25% sequence identity match along with query and subject

coverage of 60% to remove hits due to partial homology). Downstream of these steps, genes within orthogroups were analysed by exploring gene copy number, conducting synteny analyses, and generating expression matrices using publicly available RNAseq data. Figures including phylogenetic trees and heatmaps were generated in R using ggtree v3.6.2 [56], ggplot2 v3.4.4 [57], and Pheatmap v1.0.12. Protein models were predicted using AlphaFold (ColabFold v1.5.5: AlphaFold2) [58] and imported into Chimera v1.18 [59]. Molecular graphics and analyses of protein models were performed with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311. Chromosome plots showing gene positions were created using RIdeogram v0.2.2 [60].

Phylogenetic analysis of gene families

Phylogenetic trees of the two gene families of interest were built by aligning deduced protein sequences using MAFFT v7.505 followed by trimming using trimAl v1.4.rev15 build and tree building using maximum likelihood in IQ-TREE version 2.0-rc1. Trees were visualised using ggtree v3.6.2 in Rstudio. In the sugar transporter orthogroup analyses, PfamScan (command line tool pfam_scan.pl) was used to search each orthogroup against the Pfam-A.hmm database with cutoff –cut_ga and an e-value threshold of 1e-3 [61] to annotate functional domains in each gene. This was used to detect additional gene families labelled as belonging to sugar transporters (possessing Pfam domain Sugar_tr; PF00083), followed by phylogenetic analysis including Drosophila and other arthropod SLC sequences to infer the class of SLC each orthogroup belonged to [62]. In the propeller protein analyses, putative HGT was investigated using a BLASTp search (e-value threshold of 1e-3) [63] against the BLAST nr database with all lepidopteran sequences removed (Supplementary Table S3). The source of the HGT was then inferred by building a gene tree from the BLAST hits. Additional orthogroups in our datasets possessing the propellin gene were discovered by running a BLASTp search of the initial orthogroup (OG0000175) against all orthogroups in our dataset, retaining only those with percent identity equal to or above 25% and query and subject equal to or above 60%. This uncovered 8 additional homologous orthogroups, each of which contained only lepidopteran species. To further test the likely mode of origin of each of the 9 orthogroups, we carried out sequence similarity searches against the non-redundant protein sequence database (nr) and the core nucleotide database (core_nt) using a set of 10 representative species from each of the orthogroups (Supplementary Table S4). In one of the orthogroups (OG0008135), two of the species had hits against genes/proteins belonging to other insects. To test whether these BLAST hits represented true homologs, or the result of spurious homology, we aligned both insect and Spiroplasma proteins to a Manduca sexta propellin protein. This was carried out using the RCSB pairwise structure alignment tool [64].

Gene expression quantification

RNA-seq data for Bombyx mori were obtained from NCBI datasets PRJDB8614 and PRJNA675719 [65, 66], for Danaus plexippus from PRJNA663267 [67], and for Papilio machaon from PRJNA270386 [68]. RNA reads were trimmed using Trimmomatic v0.39 [69], and mapped to the reference genome using STAR 2.7.10b [70]. Stringtie v2.2.1 was used to quantify expression in each of the species datasets [71] and expression matrices were generated in RStudio using Pheatmap. Where multiple samples were available for a given tissue of lifestage, these were averaged to give one value.

Gene synteny analysis

Synteny analyses were used to test orthology of genes within and beyond Lepidoptera. For genes of interest, the gene ID, chromosome number, and location were determined from the genome annotation and gene track browser on Ensembl Rapid Release [72]. Two conserved 'marker genes' either side of the gene of interest were chosen and BLASTp searches (using Ensembl default parameters) conducted against the genomes of four Lepidoptera (Danaus plexippus, Papilio machaon, Tinea trinotella and Micropterix aruncella) and eight outgroups (Limnephilus lunatus, Limnephilus marmoratus, Limnephilus rhombicus, Glyphotaelius pellucidus, Bibio marci, Drosophila melanogaster, Adalia bipunctata and Vespula vulgaris). These data were used to compare chromosomal organisation and gene neighbourhoods surrounding the genes of interest, revealing if individual genes within lepidopteran orthology groups were 1:1 homologues between species and also whether highly divergent orthologues were present in outgroups.

Results

Novel genes emerging at the base of Lepidoptera

To build a framework for comparative analyses, a phylogenetic tree was built from 25 single copy genes from 115 species, comprising 99 Lepidoptera species representing 24 families, and 16 outgroup taxa (Fig. 1, Supplementary Table S1). The tree is broadly consistent with previously hypothesized evolutionary relationships, including placing the Micropterigidae family (Micropterix aruncella and Neomicropterix facetella in our dataset) sister to the rest of the lepidopteran lineages, the presence of

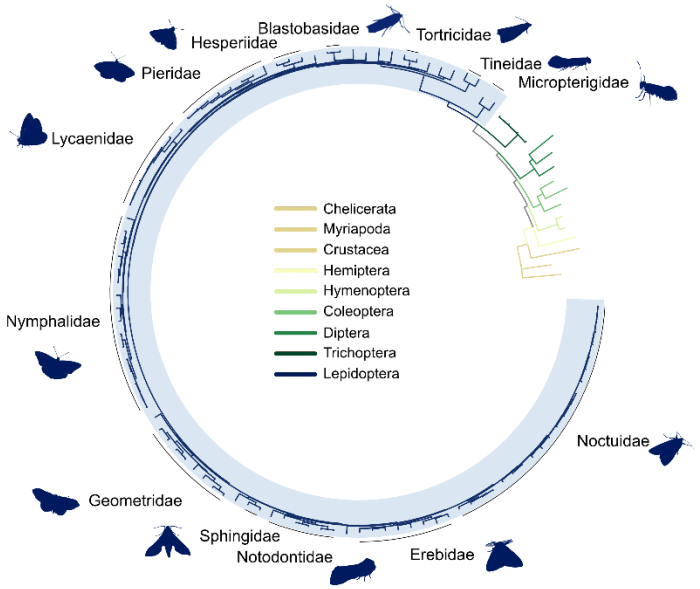


Fig. 1 Molecular phylogenetic tree of the 99 lepidopteran species from 24 families and 16 outgroup species inferred from 25 single-copy orthologues. Branches are coloured by insect order; species belonging to the named lepidopteran families are labelled with black lines on the outside of the tree

the large, established groups of Ditrysia, Apoditrysia, and Macroheterocera [39], and recovering monophyletic groups for all taxonomic families in the dataset [39, 49] (Fig. 1).

To identify novel genes or novel gene families that emerged early in lepidopteran evolution, we first constructed homologous gene groups ('orthogroups') using OrthoFinder [50]. Novel gene families here are defined as orthogroups present in a clade but missing from all outgroup taxa i.e. taxonomically restricted genes. We filtered the complete set of orthogroups to only retain those present in greater than two species. To place each of these orthogroups onto the species tree, we took the parsimonious assumption that the common ancestor of all species present in each orthogroup represented the node of origin (Fig. 2A). We identified 217 putative novel gene families originating on the branch leading to Lepidoptera (Fig. 2A).

To assess the mode of origin for each gene family we applied Pfam annotations to search for protein domains (indicative of duplication and divergence from pre-existing genes) as well as carrying out sequence similarity searches against metazoan (excluding Lepidoptera; further suggestive of duplication) and non-metazoan sequences (suggestive of HGT) from the nr protein database. We deduce that the majority of novel gene families (177 orthogroups) which originated along the lepidopteran branch likely arose via duplication followed by extensive sequence divergence (Fig. 2B). Putative HGTs accounted for only two orthogroups, as indicated by presence in Lepidoptera and non-metazoan proteomes but absent from animals other than Lepidoptera. We suggest that 38 orthogroups are potential orphan genes, candidates for origin by de novo gene genesis, although further analysis and additional data would be needed to test this hypothesis. We also detected 541 putative novel orthogroups on the branch leading to Ditrysia (representing all species outside of Micropterigidae in our dataset) (Fig. 2B). Of these, 398 likely arose from duplication, 13 via HGT, and 130 genes potentially originated de novo.

We hypothesized that novel genes of particular importance to lepidopteran biology would be present in most species of Lepidoptera analysed, with little or no gene loss after gene emergence. Furthermore, some genes of functional importance may have undergone duplication and divergence since their emergence [73]. We therefore calculated gene copy number for every orthogroup

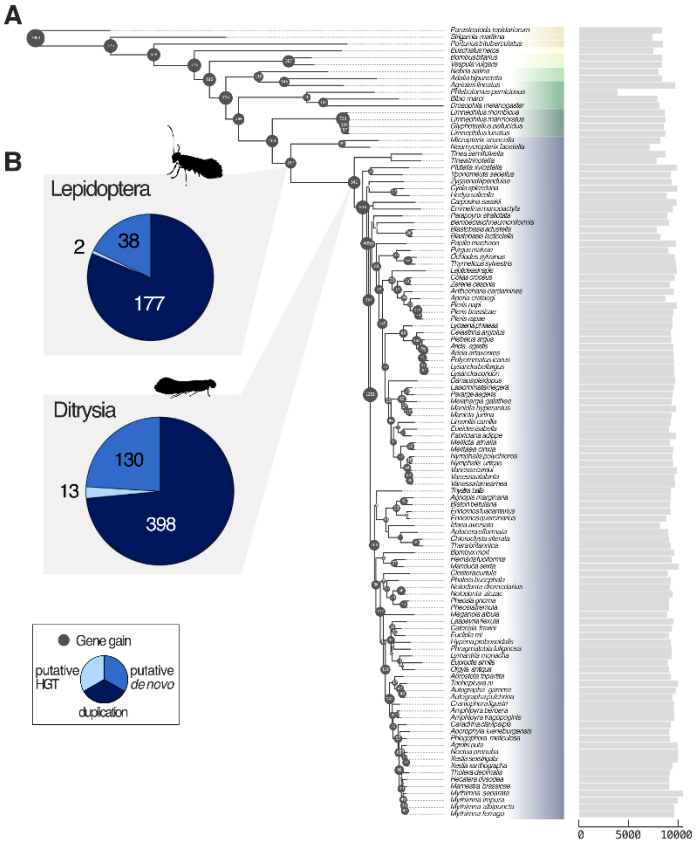


Fig. 2 A Speciestree showing numbers of orthogroups gained at each phylogenetic node. Insect orders are separated by colours. Barchart to the right of the tree displaysthe total number of orthogroups identified in each species. B Pie charts show the number of orthogroups originating at the Lepidoptera and Ditrysia nodesand proportionsof the putative modesof new gene origin

originating at the Lepidoptera node and plotted these as a heatmap against a phylogenetic tree (Fig. 3A). Approximately half of the 217 orthogroups showed a scattered phylogenetic distribution (present in a low number of species within Lepidoptera); these may represent genes that are frequently lost, or which underwent extensive sequence divergence within Lepidoptera complicating orthology assignment (right-hand columns in Fig. 3A). 87 orthogroups are present in 75% or more of the lepidopteran species in this dataset, with sporadic gene loss and occasional gene duplication on some internal branches (left-hand columns in Fig. 3A). To identify orthogroups with higher ratesof duplication patterns, we first determined that the data does not follow a normal distribution (positive, non-symmetric, right skew) and is non-parametric (Anderson–Darling test, $p < 0.05$), and that at least one orthogroup has a gene copy distribution across species which differs from the mean number of gene copies per orthogroup per species (Kruskal–Wallis rank test, $p < 0.05$; mean number of gene copies= 1.1645). We found that two orthogroups deviate significantly from the mean number of gene copies within a given orthogroup gained at the lepidopteran node: OG0000164 and OG0000175 (Dunn test, $p < 0.0001$; Fig. 3B). These two orthogroups have the highest variation in copy number, implying they have undergone extensive gene

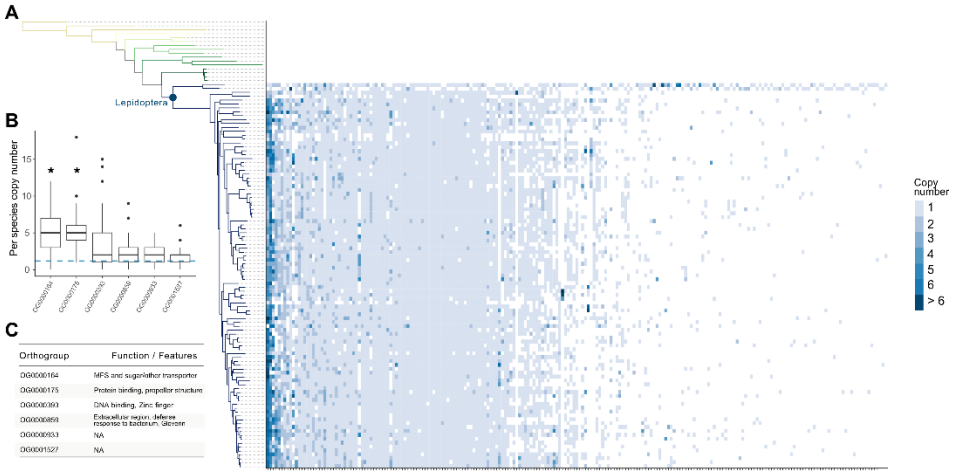


Fig. 3 Copy number of genes gained on the ancestral node of Lepidoptera. A Heatmap (right) showing gene copy number for each orthogroup originating at the Lepidoptera node mapped to the species tree (left). Lepidoptera node is labelled with a blue circle. Orthogroups on the right-hand side of the figure have genes present in few species and may include spurious homologies. B Boxplots showing copy number variation per species in the top 6 orthogroups present in all or most lepidopteran species. Blue broken line signifies the mean copy number per species for all orthogroups. Orthogroups OG0000164 and OG0000175 have a mean copy number significantly different from the mean copy number of lepidopteran orthogroups, as signified by an asterisk ($p < 0.05$). C Table showing functions and features from six orthogroups deviating above the average copy number per orthogroup

duplication within Lepidoptera, and they are also present in every lepidopteran species analysed. Sequence homology from BLASTp searches and domain annotation from Pfam revealed that these proteins have a putative sugar transporter domain (OG0000164; MFS and Sugar/other transporter, PF00083.27, GO:0016020|GO:0022857|GO:0055085) and a 6-bladed beta propeller 3D structure (OG0000175; GO:0005515) (Fig. 3C). To determine whether there were any functions enriched in the full set of 217 orthogroups gained on the lepidopteran node, we analysed the functional domains of each to determine whether there were any categories which were significantly overrepresented. Although no functional categories were found to be enriched within this dataset, approximately 9% of the orthogroups (19 out of 217) were found to contain a zinc finger domain (Supplementary Table S2). We also discover that the Gloverin gene family (OG0000859) emerged on the branch leading to Lepidoptera (Fig. 3C). The gloverin gene has previously been described as a lepidopteran novelty, and we confirm its emergence coincident with the evolution of Lepidoptera, where it has been retained in 86 of the 99 lepidopteran species in our dataset including Micropterix aruncella (Fig. 3A). Gloverin, first purified from Hyalophora gloveri [74], is a glycine rich protein with no detectable homology outside of Lepidoptera. It functions as an antimicrobial peptide against a range of bacteria, with greater specificity to Gram-negative bacteria, and appears to be commonly and widely expressed across a range of life stages and tissues, with significant increases in expression observed following exposure to bacteria [75, 76].

Gene expansion of lepidopteran sugar and solute transporters

The orthogroup originating on the node leading to the Lepidoptera with the highest mean copy number is a sugar transporter gene family (OG0000164) (Fig. 3). Across the species analysed, the copy number for this lepidopteran-specific orthogroup ranged from one gene (Micropterix aruncella) to twelve genes (Manduca sexta). As the sugar transporter protein superfamily is large and diverse in animals [62], and to understand the significance of this Lepidoptera-specific orthogroup, we extended our analysis to include all orthogroups containing a sugar transporter domain. We found 99 orthogroups with genes possessing a sugar transporter domain present across all species in our dataset (Fig. 4A), nine of which are annotated as emerging on the lepidopteran or ditrysian node; gene copy number for all nine orthogoups

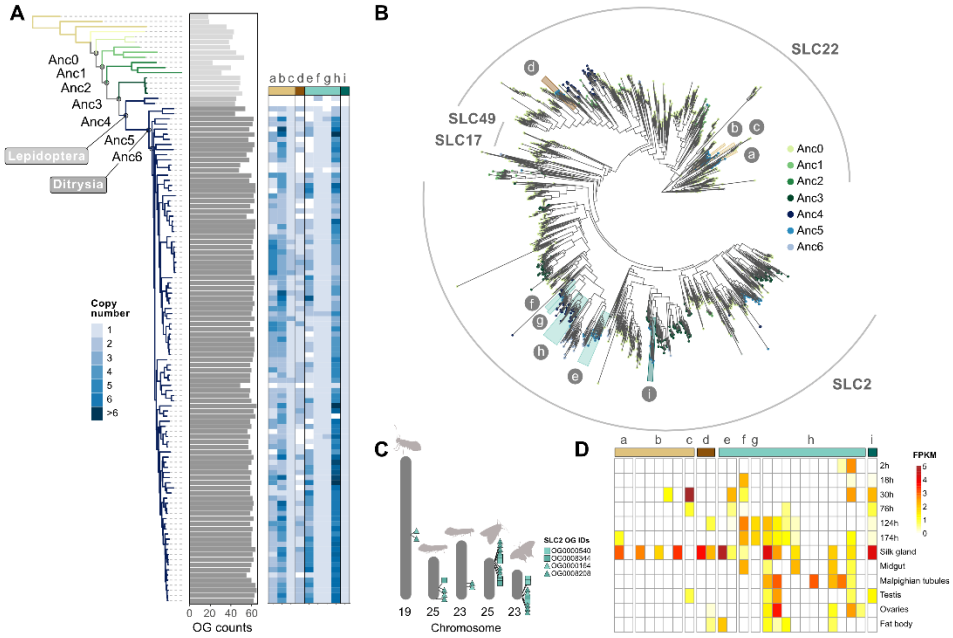


Fig. 4 Origins and evolution of lepidopteran and ditrysian-specific sugar transporter genes. Orthogroups are identified as follows: a—OG0000700, b—OG0000319, c—OG0007512, d—OG0001801, e—OG0000540, f—OG0008208, g—OG0008344, h—OG0000164, i—OG000840 (A) Species tree on the left is coloured by taxonomic group, with the Lepidoptera and Ditrysia nodes labelled. The numbered ancestral nodes correlate to the node of origin for the orthogroups shown in the gene tree (part B). The grey bar chart (middle) shows that Lepidoptera (darker grey bars) have a higher total number of sugar transporter orthogroups compared to outgroup species. Copy number of lepidopteran and ditrysian-specific orthogroups varies greatly (heatmap, right): SLC22 transporters are below the light and dark brown bars (labelled a-d); SLC2 transporters are below the light and dark blue/green bars (labelled with letters e-i). B Phylogenetic tree built using a representative sample of outgroup sugar transporters, combined with sugar transporters identified in the Lepidoptera. SLC2 transporters are highlighted in light and dark blue/green along with letters e-i, while SLC22 transporters are highlighted in light and dark brown with letters a-d. Tip colours represent the node of origin (as shown in the species tree in part A) for each orthogroup. C The four closely related SLC2 transporters were mapped to a selection of lepidopteran chromosomes (left to right: Micropterix aruncella, Tinea trinotella, Tinea semifulvella, Papilio machaon and Autographa gamma). Sugar transporters of lepidopteran origin are represented by a triangle, while those of ditrysian origin are represented by a square. All four transporter orthogroups group in close physical proximity, on the same chromosome. D Heatmap of expression patterns of nine lepidopteran/ditrysian originating sugar transporters in Bombyx mori tissues

in each species shows varying rates of copy number expansion and gene loss (Fig. 4A). Four of these orthogroups were single copy in all or most species, while five have undergone extensive gene duplication since their lepidopteran origins (Fig. 4A and B).

Next, we wanted to determine whether these nine lepidopteran sugar transporter-containing orthogroups had a single evolutionary origin or whether they had evolved independently from separate ancestral sugar transporter genes. To resolve this, a phylogenetic tree of transporter proteins from representative lepidopteran and outgroup species was constructed using all sugar transporter orthogroups (Fig. 4B). The tree topology suggests multiple origins of the lepidopteran- and ditrysian-specific sugar transporter genes, although not each of the nine orthogroups had independent origins. Notably, two lepidopteran-originating orthogroups (OG0000164 and OG0008208), and two ditrysian-originating orthogroups (OG0008344 and OG0000540) group close to each other in the phylogenetic tree (e–h). Another orthogroup (i; OG0008400) is located outside this clade, however each of these orthogroups are present in a larger clade consisting of members of the solute carrier 2 (SLC2) family (Fig. 4B). In some taxa SLC2 genes have been shown to

encode proteins that facilitate transport of small sugars across cell membranes [62].

The remaining four sugar transporter orthogroups (a-d) are all ditrysian-specific, three of which form a single monophyletic group. The fourth orthogroup is located in a more phylogenetically distinct group, however, all orthogroups belong to the SLC22 protein subfamily (Fig. 4B). The SLC22 proteins are membrane transporters known to regulate metabolic functions, transporting a broader range of small molecules than SLC2 [62]. In all instances, the most closely related orthogroups in the gene tree contain both outgroups and lepidopteran species (Fig. 4B). This implies that the lepidopteran- and ditrysian-specific transporter orthogroups originated from more ancient gene families that were present across all or most insects including Lepidoptera. These ancestral genes duplicated and underwent extensive amino acid substitutions specifically in the lineage leading to Lepidoptera or Ditrysia.

The four SLC2-like orthogroups [e–h] which group closely together in the gene tree (Fig. 4B) are also co-located in the genome, found consistently in close association with one another across diverse lepidopteran species (Fig. 4C). This suggests that these sugar transporter genes originated from a single ancestral duplication event at the base of the Lepidoptera and subsequently underwent tandem duplication in the ancestral lepidopteran and again in the branch leading to Ditrysia. In contrast, for the SLC22-like orthogroups gained on the ditrysian branch (a-d), we do not see the same close linkage and instead they are scattered on separate chromosomes (Supplementary Figure S1). If these originated from a single ancestral gene, as suggested by branching patterns in the gene tree, they dispersed around the genome after duplication.

To investigate possible functions of these lepidopteran-species transporter gene families, we assessed their patterns of expression across multiple time points and tissues from Bombyx mori [65, 66]. All genes from the nine lepidopteran and ditrysian-specific gene families are expressed in at least one tissue or at one developmental time point (Fig. 4D). The silk gland was the most frequent site of expression across the nine orthogroups, but some of the genes have wide and distributed expression (Fig. 4D).

Lepidoptera propellin genes arose through horizontal gene transfer

The second gene family which emerged at the base of Lepidoptera and is maintained in significantly high copy number across all butterfly and moth species analysed is orthogroup OG0000175 (Fig. 3). The genes in this family were previously undescribed in insects. Below we show they encode proteins with a beta-propeller structure; we therefore name this the propellin gene family. Intriguingly, this group of genes is conserved across Lepidoptera yet does not have detectable sequence identity to any orthogroups in the arthropod outgroups included in our initial analysis. Furthermore, iterative BLAST searching revealed that OG000175 is not the only set of propellin genes in Lepidoptera; the genes are split into eight distinct orthogroups, including the original group OG000175 with the highest copy number. All eight propellin orthogroups are specific to Lepidoptera (Supplementary Figure S2). Combining all propellin orthogroups together, we find Lepidoptera genomes have an average of 11 gene copies, ranging from 3 copies in Neofacetella micropterix (Micropterigidae) to 25 copies in Phragmatobia fuliginosa (Erebidae) (Supplementary Table S5).

To assess the likely origin of the propellin gene in Lepidoptera, we carried out a BLASTp search using all proteins in orthogroup OG0000175 against the NCBI nr protein database excluding Lepidoptera sequences. This revealed significant sequence similarity matches to proteins from bacterial species. The most frequent bacterial genus in the set of matches was Spiroplasma, with additional matches in Macrococcus and Escherichia (Supplementary Table S3). Using iterative rounds of BLAST searching, we found very few matches outside bacteria; we identified a potentially related unnamed gene in the genome of the plant Picea sitchensis (spruce; ABK22491.1) and a fungus gnat Bradysia coprophila (30% identity to a Spiroplasma homologue of Lepidoptera propellin genes over 16% query cover). To confirm the likely bacterial origins of the different propellin orthogroups, we also carried out BLASTp and tBLASTn searches against the nr and core nt databases, respectively, for each of the propellin orthogroups. We searched protein sequences from 10 representative species in each of the 8 orthogroups using both methods and found that bacterial, specifically Spiroplasma, sequences represented the majority of sequence similarity hits (Supplementary Table S4). While two genes from one orthogroup showed sequence similarity to other insect proteins (E3 ubiquitin ligases), we deduce that these hits are likely a result of spurious homology, with low query coverage (34–40%) and sequence similarity likely resulting from convergent amino acid residues in repetitive regions. In addition to this, all other hits from other species in the same orthogroup showed sequence similarity to Spiroplasma and other bacterial proteins. These Spiroplasma proteins were deduced to have similar tertiary structures to propellin (i.e. 6-bladed beta propeller; RMSD value of 3.73); in contrast, the spurious insect protein hits possessed multiple alpha helices and no structural similarity (RMSD value of 5.57).

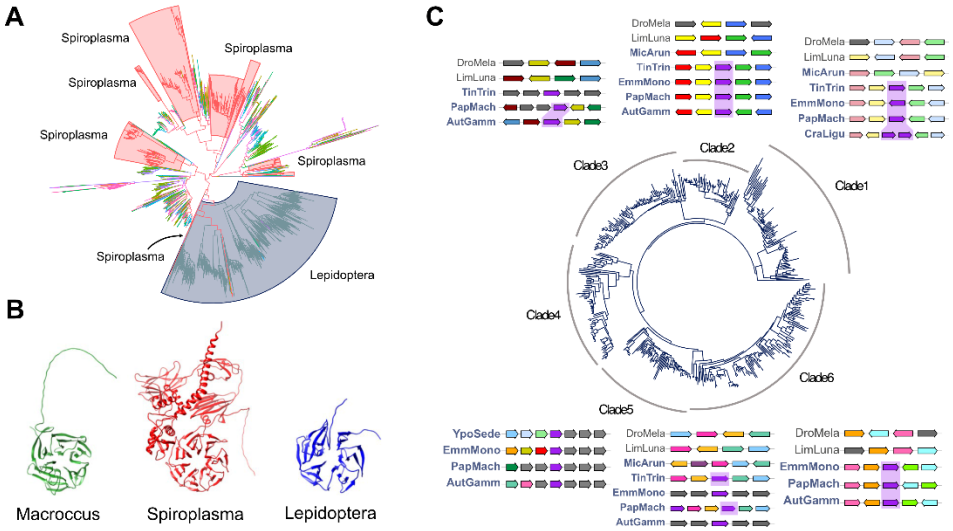


Fig. 5 Lepidoptera-specific genes encoding proteins with sequence identity and structural similarity to bacterial 6-bladed propeller proteins. A Gene tree of propellin and putative bacterial homologs. The Lepidoptera clade (blue) and Spiroplasma clades (red) are labelled with coloured boxes and text. All other branches represent a range of bacterial species which are shown in Supplementary Figure S3. Molecular phylogenetic analysis indicates the propellin genes of Lepidoptera are monophyletic, whose most closely branching lineages are Spiroplasma genes, and sister group to a clade dominated by Spiroplasma genes (highlighted in red). B AlphaFold predictions suggest lepidopteran propellin proteins form 6-bladed propeller structures similar to bacterial homologues; examples shown from Macrococcus (green), Spiroplasma (red) and the lepidopteran M. sexta (blue). Additional protein structure predictions in Supplementary Figure S4. C Molecular phylogenetic analysis indicates that the largest orthogroup of lepidopteran propellin genes divides into 6 clades, each gene (purple) located at a different chromosomal location, most of which show conserved synteny between lepidopteran species (synteny indicated by shaded purple regions). The Micropterigidae species M. aruncella only has a gene in clades 1. Marker genes are shown by various colours

Next, we constructed a phylogenetic tree of all propellin copies and their putatively homologous sequences. This shows that all lepidopteran propellin sequences are closely related in the gene tree (Fig. 5A, Supplementary Figure S3). The most closely related branches to the Lepidoptera clade are Spiroplasma sequences which, along with a larger sister clade dominated by Spiroplasma sequences, suggests there has been a putative horizontal gene transfer from bacteria to Lepidoptera (Fig. 5A). Based on the gene tree topology, we cannot exclude the possibility of multiple horizontal transfer events into Lepidoptera. Although there are clear sequence similarity matches between Lepidoptera and Spiroplasma, the level of primary sequence identity is low. The highest percentage identity found between a lepidopteran protein and a Spiroplasma protein had only 35% identity over a sequence alignment of 134aa. This represents 45% coverage of a lepidopteran propellin protein (ENSAGMG00005008917.1) and 18% of a Spiroplasma protein (WP_164028422.1). To further assess legitimacy of the homologous relationships between these genes with low sequence identity, we predicted 3D structures of the deduced proteins from Spiroplasma, Macrococcus, and Lepidoptera (using six genes from Manduca sexta as representative of Lepidoptera) with AlphaFold [58] (Fig. 5B; Supplementary Figure S4). We find clear similarity in predicted protein structure with all lepidopteran and bacterial sequences having a 6-bladed propeller structure (Fig. 5B). Further support for homology between lepidopteran propellin proteins and bacterial proteins was found when we aligned representative protein structures, which showed an RMSD value of 3.83 and TM-score of 0.69 (Supplementary Figure S5). Each structured propeller region within a propellin protein is approximately 221aa long consisting of blades of 30aa in length. There is variability outside of the beta-propeller domain including regions of varying length and structure, most notably in the additional domains in the Spiroplasma protein model (Fig. 5B). To reflect this protein structure, we name the Lepidoptera genes the propellin gene family.

To further investigate the evolution of propellin genes, we focussed attention on the high-copy number propellin orthogroup (OG0000175; Supplementary Figure S2). Phylogenetic analysis divides this orthogroup into six clades within Lepidoptera, which we refer to as gene subfamilies (Fig. 5C). The early diverging lineage of Lepidoptera, represented by the family Micropterigidae, has a propellin gene in clade 1 in the gene tree (Fig. 5C). Next, we examined the local gene synteny for these six subfamilies across representative lepidopteran species. Genes from each subfamily, excluding clade four, were found in a microsyntenic cluster of genes ('marker genes') which are homologous across most or all species (Fig. 5C). This confirms that each of these 6 propellin subfamilies are one-to-one orthologues across Lepidoptera. Importantly, many of the marker genes also exist in microsyntenic blocks in the arthropod outgroups, consistent with these being the genomic sites where the Lepidoptera-specific propellin gene was integrated (Fig. 5C). Since the six propellin subfamilies are at distinct chromosomal locations, yet form a monophyletic group in molecular phylogenetic analysis, we propose that this orthogroup emerged through a single HGT event from Spiroplasma or another bacterial source to Lepidoptera, followed by duplication and transposition around the genome. These duplications generated not only the six subfamilies analysed in detail, but also likely the additional propellin genes referred to above. We note that genes in subfamily 1 are intronless (or have one intron), while the remaining subfamilies and additional propellin orthogroups have between 0 and 8 introns, with the median count being 1 intron. This could reflect transposition via an RNA intermediate or could be a legacy of the gene's bacterial origin (Supplementary Table S5).

For a first insight into the possible functional role of the lepidopteran propellin genes, we analysed the expression of all copies of propellin using transcriptomic data sets from three species: Bombyx mori, Danaus plexippus, and Papilio machaon (Supplementary Figure S6). While we find evidence for expression of all gene copies in each species, the patterns are complex and variable within and between species. In Danaus plexippus for example, while most propellin copies show some expression in larval or pupal stages (8 of the 11 genes), levels of expression are highest in the adult life stage, with particularly high expression found in the thorax, compared to the head or abdomen (Supplementary Figure S6). In Papilio machaon most copies are restricted to one or two life stages, while others are strongly expressed throughout the life cycle of the butterfly. Expression in Bombyx mori shows wider coverage across life stages and tissue types, with most gene copies expressed in early developmental and adult life stages. While expression is common across most adult tissue types in Bombyx mori, there is little, or no expression found in the midgut or silk glands (Supplementary Figure S6). While there is little correlation in expression between homologous copies of propellin across all three species, we note that transcriptomic datasets available are not comprehensive. However, such pervasive expression across life stages and tissues in multiple species provides support to the fact that these genes are functional across a wide range of lepidopteran species.

We noted above that there were some sequence similarity matches outside bacteria and Lepidoptera. The putatively homologous gene from Diptera is an uncharacterised locus (LOC119081672, encoding putative protein XP_037046651) on an unplaced scaffold in the genome assembly of a fungus gnat Bradysia coprophila [77]. We find this gene is present in two species of Bradysia. It is unlikely that the fungus gnat scaffold is a contaminating sequence since it is present in two species, and because it is adjacent in the genome to recognisable insect genes (Supplementary Figure S6). Analysis of the unplaced scaffold reveals clearly dipteran genes immediately 3' (LOC119081668) and relatively close 5' (LOC119081673 and LOC119081675) to the gene of interest. Intriguingly, a locus immediately 5' (LOC119081585) has high similarity to springtail (Collembola) tyrosine kinases, and the next neighbouring gene (LOC119081673) is Bradysia-specific (Supplementary Figure S6). We therefore suggest the Diptera gene LOC119081672 arose by an independent HGT from Spiroplasma in the Bradysia fungus gnat genus, which has likely also acquired other genes by HGT. We have not deduced the origin of the loci with a sequence match in Picea sitchensis (spruce).

Discussion

In this study we identified 217 'novel' genes arising on the evolutionary lineage leading Lepidoptera, after it had diverged from outgroups including the closest related order Trichoptera (caddisflies). We caution, however, against this as a quantitative measure of genomic novelty. First, we are using a pragmatic definition of novelty that includes de novo genes, horizontally transferred genes, and gene duplication followed by sequence divergence; altering parameters relating to sequence divergence could increase or decrease the gene count [78]. To improve inference of new genes in early lepidopteran evolution, we employed a phylogenetically informed approach to construct gene families, minimising the effects of bias resulting from rapid sequence divergence [50]. Second, novelty at the Lepidoptera node could be 'undercounted' if some genome annotations are incomplete, particularly those of early diverging lepidopteran taxa. Third, there are factors that could spuriously 'overcount' novelty. For example, in our study around half the

novel orthogroups were found sporadically in a small number of distantly related Lepidoptera species. This could indicate repeated gene loss following the origin of the novel gene but could also include 'noise' as a result of some proteins being grouped incorrectly due to spurious sequence identity. Secondary loss of genes from caddisfly genomes could theoretically cause overcounting of genes on the Lepidoptera node, but we have minimized this risk through use of four caddisfly genomes. We also noted a small degree of overcounting (<2%) emerging from alternative genome annotation methods [79], but we accounted for this (see Methods). Specifically, the initial input data consisted of proteomes predicted from the Ensembl Genebuild annotation which incorporates RNA sequence data and filters poorly supported potential coding transcript proteins. A second method of genome annotation, the BRAKER method, is potentially less stringent and found some genes that had been missed by Genebuild. The difference amounted to just two orthogroups. The same caveats apply to counts of novel genes at other similarly deep phylogenetic nodes. Despite this caveat, we find it interesting that even more apparent gene novelty (541 gene families) dates to the node leading to Ditrysia. These genes require further analysis, but the observation suggests that the evolution of new biological traits continued during the early evolutionary radiation of moths. More important than an absolute number of novel genes, the analysis gives us a first look into the relative importance of different modes of gene origin during the emergence of Lepidoptera. We find the majority of novel gene families gained on the ancestral lepidopteran branch arose via gene duplication and divergence (~ 82%) while around~18% genes had no sequence matches or any recognisable domains. These are putative candidates for genes arising de novo from non-coding genes. Just two genes (< 1%) are candidates for having arisen via HGT (including the propellin gene), with hits to bacterial or fungal species.

One of the genes that likely arose via HGT was highlighted in our analysis as a novel gene that underwent extensive gene duplication in Lepidoptera to generate a large gene family. This gene family, which we name the propellin genes, is potentially functional as evidenced by the extensive retention through evolution and conserved domain structure. Currently, however, its precise role in lepidopteran biology is unclear. Phylogenetic analyses suggest that the progenitor of the propellin gene family was transferred to an insect from Spiroplasma bacteria, some time on the Lepidoptera stem lineage. Spiroplasma is a well-known intracellular symbiont in arthropods. Furthermore, Spiroplasma is known to colonise reproductive tissues, which in turn impacts upon the host's reproduction, and indeed this genus is one of two bacterial symbionts in Lepidoptera for which maternal transmission has been demonstrated [80]. In some cases, transmission is enhanced by manipulation of host physiology, such as male-killing which increases the number of female offspring as observed in Danaus chrysippus [81]. Clearly, persistent association with reproductive tissues gives opportunity for horizontal gene transfer, as the symbiont DNA is in close physical proximity to the DNA of the host germline. This has been seen in the relationship between a mealybug and two endosymbiont species Tremblaya and Moranella [82]. Interestingly, we also found a putatively homologous gene in two species of Diptera (genus Bradysia), possibly reflecting an independent HGT event. This is consistent with previous findings that some types of gene are more prone to HGT than others, perhaps those encoding proteins with few interaction partners [83]. The evolutionary retention of the likely HGT-derived propellin gene, plus its extensive gene duplication in Lepidoptera, suggest this gene family likely evolved to perform functions that are important for the biology of moths and butterflies. We do not know the biological role, or roles, of propellin genes in Lepidoptera, but note that their bacterial homologues have diverse functions including ligand-binding proteins, signalling proteins, lysases, structural proteins, isomerases and hydrolases [84]. It is worth noting that the Spiroplasma genes which group closest to the propellin genes in the gene tree are annotated as hypothetical proteins without known function, suggesting more work is needed to understand the functional context of this gene.

The only other novel lepidopteran gene to show such widespread retention and extensive gene duplication encodes a family of SLC2-like sugar transporter proteins. In other animals, members of the SLC2 sugar transporter superfamily encode glucose-uptake proteins, ribose transport proteins, and several putative membrane proteins probably involved in sugar transport [62, 85, 86]. The functional link to sugars is particularly intriguing since the ecological association between Lepidoptera and sugar-feeding changed markedly in early lepidopteran evolution. Specifically, adult moths in the basal family Micropterigidae primitively lack a proboscis and are pollen feeders, whereas adult moths and butterflies in the Ditrysia use a proboscis to access sugar-rich nectar in flowers. Our wider comparative survey of sugar transporter gene families picks up potentially interesting co-evolution between this ecological shift and the sugar transporter genes. We find that although OG0000164 (and one other sugar transporter gene family) are present in pollen-feeding Micropterigidae, it is not until the evolution of the nectar-feeding Ditrysia that we see extensive gene duplication, widespread gene retention and the emergence of additional SLC-like sugar transporter

gene families [62]. We suggest, therefore, that novel sugar transporter gene families emerged at the base of Lepidoptera, but it was only later in lepidopteran evolution that massive gene duplication and functional divergence of sugar-transporter genes took place, in association with nectar feeding. The causal link between these genetic changes and the evolution of novel feeding behaviour in the early evolution of Lepidoptera warrants further study.

Conclusion

We have demonstrated the emergence of 217 novel gene families (orthogroups) on the node leading to Lepidoptera and 541 novel gene families emerging on the node leading to the Ditrysia. Two orthogroups have significantly higher gene copy per species across Lepidoptera indicative of extensive gene duplication following their origins. One likely originated by horizontal gene transfer from the endosymbiont bacterium Spiroplasma and then duplicated to generate a diverse group of 'propellin' genes encoding a 6-bladed propeller domain. The other encodes a large set of sugar transporter proteins and is part of a diverse set of sugar and solute transporter genes that duplicated and diverged extensively in early lepidopteran evolution.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11338-x.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

We thank the taxonomic experts who collected the insects used in the study and all members of the DToL project at the Sanger Institute for sequencing and assembling the genomes. For the purpose of Open Access, the author has applied a CCBY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Authors' contributions

P.W.H.H. and P.O.M. conceived the study and oversaw the research. A.E.H. and P.O.M. designed analyses and carried out the bioinformatic research presented. A.E.H., P.W.H.H. and P.O.M. interpreted all results. A.E.H. wrote the initial draft of the manuscript, and P.W.H.H. and P.O.M. edited versions. All authors read and approved the final manuscript.

Funding

AEH was supported by the Oxford Interdisciplinary DTP and funding from the Biotechnology and Biological Sciences Research Council (UKRI-BBSRC) [grant number BB/T008784/1]; POM and PWH acknowledge funding from Wellcome Trust Darwin Tree of Life Awards (grant agreements 218328, 226458).

Data availability

Genome data associated with this study is listed in Supplementary Table S1 along with accession numbers. Data and code generated in this study can be found on figshare; figshare.com/s/32d1b9055257dad1892f.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 6 December 2024 Accepted: 10 February 2025
Published online: 19 February 2025

References

1. Carroll SB, Gates J, Keys DN, Paddock SW, Panganiban GE, Selegue JE, et al. Pattern formation and eyespot determination in butterfly wings. Science. 1994;265:109–14.
2. Wucherpfennig JI, Howes TR, Au JN, Au EH, Roberts Kingman GA, Brady SD, et al. Evolution of stickleback spines through independent cis-regulatory changes at HOXDB. Nat Ecol Evol. 2022;6:1537–52.
3. Tian S, Asano Y, Banerjee TD, Wee JLQ, Lamb A, Wang Y, et al. A micro-RNA is the effector gene of a classic evolutionary hotspot locus. bioRxiv org. 2024;:2024.02.09.579741.
4. Livraghi L, Hanly JJ, Evans E, Wright CJ, Loh LS, Mazo-Vargas A, et al. A long noncoding RNA at the cortex locus controls adaptive coloration in butterflies. Proc Natl Acad Sci U SA. 2024;121: e2403326121.
5. Hoekstra HE, Coyne JA. The locus of evolution: evo devo and the genetics of adaptation: The locus of evolution. Evolution. 2007;61:995–1016.
6. Ota KG, Kuraku S, Kuratani S. Hagfish embryology with reference to the evolution of the neural crest. Nature. 2007;446:672–5.
7. Dutrow EV, Serpell JA, Ostrander EA. Domestic dog lineages reveal genetic drivers of behavioral diversification. Cell. 2022;185:4737–55.e18.
8. Richter DJ, Fozouni P, Eisen MB, King N. Gene family innovation, conservation and loss on the animal stem lineage. eLife. 2018;7:e34226.
9. Paps J, Holland PWH. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. Nat Commun. 2018;9:1730.
10. Thomas GWC, Dohmen E, Hughes DST, Murali SC, Poelchau M, Glastad K, et al. Gene content evolution in the arthropods. Genome Biol. 2020;21:15.
11. Fernández R, Gabaldón T. Gene gain and loss across the metazoan tree of life. Nat Ecol Evol. 2020;4:524–33.
12. Guijarro-Clarke C, Holland PWH, Paps J. Widespread patterns of gene loss in the evolution of the animal kingdom. Nat Ecol Evol. 2020;4:519–23.
13. Cicconardi F, Milanetti E, Pinheiro de Castro EC, Mazo-Vargas A, Van Belleghem SM, Ruggieri AA, et al. Evolutionary dynamics of genome size and content during the adaptive radiation of Heliconiini butterflies. Nat Commun. 2023;14:5620.
14. Long M, Betrán E, Thornton K, Wang W. The origin of new genes: glimpses from the young and old. Nat Rev Genet. 2003;4:865–75.
15. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res. 2010;20:1313–26.
16. Haggerty LS, Jachiet P-A, Hanage WP, Fitzpatrick DA, Lopez P, O'Connell MJ, et al. A pluralistic account of homology: adapting the models to the data. Mol Biol Evol. 2014;31:501–16.
17. Van Oss SB, Carvunis A-R. De novo gene birth. PLoS Genet. 2019;15: e1008160.
18. Rödelsperger C, Prabh N, Sommer RJ. New gene origin and deep taxon phylogenomics: Opportunities and challenges. Trends Genet. 2019;35:914–22.
19. Ohno S. Evolution by gene duplication. 1970th ed. Berlin, Germany: Springer; 2013.
20. Rastogi S, Liberles DA. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol Biol. 2005;5:28.
21. Conrad B, Antonarakis SE. Gene duplication: a drive for phenotypic diversity and cause of human disease. Annu Rev Genomics Hum Genet. 2007;8:17–35.

22. Sémon M, Wolfe KH. Preferential subfunctionalization of slow-evolving genes after allopolyploidization in Xenopus laevis. Proc Natl Acad Sci U S A. 2008;105:8333–8.
23. Holland PWH, Marlétaz F, Maeso I, Dunwell TL, Paps J. New genes from old: asymmetric divergence of gene duplicates and the evolution of development. Philos Trans R Soc Lond B Biol Sci. 2017;372:20150480.
24. DuBose JG, de Roode JC. The link between gene duplication and divergent patterns of gene expression across a complex life cycle. Evol Lett. 2024;8:726–34.
25. McClintock B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U SA. 1950;36:344–55.
26. Long M, Langley CH. Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science. 1993;260:91–5.
27. Leonard G, Richards TA. Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life. Proc Natl Acad Sci U SA. 2012;109:21402–7.
28. Bornberg-Bauer E, Albà MM. Dynamics and adaptive benefits of modular protein evolution. Curr Opin Struct Biol. 2013;23:459–66.
29. Cosby RL, Judd J, Zhang R, Zhong A, Garry N, Pritham EJ, et al. Recurrent evolution of vertebrate transcription factors by transposase capture. Science. 2021;371:eabc6405.
30. Mulhair PO, Moran RJ, Pathmanathan JS, Sussfeld D, Creevey CJ, Siu-Ting K, et al. Bursts of novel composite gene families at major nodes in animal evolution. bioRxiv. 2023;:2023.07.10.548381.
31. Husnik F, McCutcheon JP. Functional horizontal gene transfer from bacteria to eukaryotes. Nat Rev Microbiol. 2018;16:67–79.
32. Li Y, Liu Z, Liu C, Shi Z, Pang L, Chen C, et al. HGTis widespread in insects and contributes to male courtship in lepidopterans. Cell. 2022;185:2975–87.e10.
33. Keeling PJ. Horizontal gene transfer in eukaryotes: aligning theory with data. Nat Rev Genet. 2024;25:416–30.
34. Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. Proc Natl Acad Sci U SA. 2006;103:9935–9.
35. McLysaght A, Hurst LD. Open questions in the study of de novo genes: what, how and why. Nat Rev Genet. 2016;17:567–78.
36. Zhao L, Svetec N, Begun DJ. De Novo genes Annu Rev Genet. 2024. https://doi.org/10.1146/annurev-genet-111523-102413.
37. Wybouw N, Pauchet Y, Heckel DG, Van Leeuwen T. Horizontal gene transfer contributes to the evolution of arthropod herbivory. Genome Biol Evol. 2016;8:1785–801.
38. Santos ME, Le Bouquin A, Crumière AJJ, Khila A. Taxon-restricted genes at the origin of a novel trait allowing access to a new environment. Science. 2017;358:386–90.
39. Kawahara AY, Plotkin D, Espeland M, Meusemann K, Toussaint EFA, Donath A, et al. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. Proc Natl Acad Sci U SA. 2019;116:22657–63.
40. Mitter C, Davis DR, Cummings MP. Phylogeny and evolution of Lepidoptera. Annu Rev Entomol. 2017;62:265–83.
41. Kawahara AY, Storer C, Carvalho APS, Plotkin DM, Condamine FL, Braga MP, et al. A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts and biogeographic origins. Nat Ecol Evol. 2023;7:903–13.
42. Krenn HW. Feeding mechanisms of adult Lepidoptera: structure, function, and evolution of the mouthparts. Annu Rev Entomol. 2010;55:307–27.
43. Bazinet AL, Mitter KT, Davis DR, Van Nieukerken EJ, Cummings MP, Mitter C. Phylotranscriptomics resolves ancient divergences in the Lepidoptera: Ancient divergences in Lepidoptera. Syst Entomol. 2017;42:305–16.
44. Macias-Muñoz A, Rangel Olguin AG, Briscoe AD. Evolution of phototransduction genes in Lepidoptera. Genome Biol Evol. 2019;11:2107–24.
45. Mulhair PO, Crowley L, Boyes DH, Harper A, Lewis OT, Darwin Tree of Life Consortium, et al. Diversity, duplication, and genomic organization of homeobox genes in Lepidoptera. Genome Res. 2023;33:32–44.
46. Mulhair PO, Crowley L, Boyes DH, Lewis OT, Holland PWH. Opsin gene duplication in Lepidoptera: Retrotransposition, sex linkage, and gene expression. Mol Biol Evol. 2023;40:msad241.
47. Darwin Tree of Life Project Consortium. Sequence locally, think globally: The Darwin Tree of Life Project. Proc Natl Acad Sci U SA. 2022;119:e2115642118.
48. Lewin HA, Richards S, Lieberman Aiden E, Allende ML, Archibald JM, Bálint M, et al. The Earth BioGenome Project 2020: Starting the clock. Proc Natl Acad Sci U SA. 2022;119: e2115635118.
49. Rota J, Twort V, Chiocchio A, Peña C, Wheat CW, Kaila L, et al. The unresolved phylogenomic tree of butterflies and moths (Lepidoptera): Assessing the potential causes and consequences. Syst Entomol. 2022;47:531–50.
50. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:238.
51. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
52. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.
53. Steenwyk JL, Buida TJ, Labella AL, Li Y, Shen X-X, Rokas A. PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. Bioinformatics. 2021;37:2325–31.
54. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37:1530–4.
55. Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, et al. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Genome Res. 2024;34:769–77.
56. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. Ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8:28–36.
57. Wickham H. Ggplot2: Elegant graphics for data analysis. 2nd ed. Cham, Switzerland: Springer International Publishing; 2016.
58. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.
59. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem. 2004;25:1605–12.
60. Hao Z, Lv D, Ge Y, Shi J, Weijers D, Yu G, et al. RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. PeerJ Comput Sci. 2020;6: e251.
61. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42 Database issue:D222–30.
62. Denecke SM, Driva O, Luong HNB, Ioannidis P, Linka M, Nauen R, et al. The identification and evolutionary trends of the solute carrier superfamily in arthropods. Genome Biol Evol. 2020;12:1429–39.
63. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
64. Bittrich S, Segura J, Duarte JM, Burley SK, Rose Y. RCSB Protein Data Bank: exploring protein 3D similarities via comprehensive structural alignments. Bioinformatics. 2024;40:btae370.
65. Xu G-F, Gong C-C, Lyu H, Deng H-M, Zheng S-C. Dynamic transcriptome analysis of Bombyx mori embryonic development. Insect Sci. 2022;29:344–62.
66. Yokoi K, Tsubota T, Jouraku A, Sezutsu H, Bono H. Reference Transcriptome Data in Silkworm Bombyx mori. Insects. 2021;12:519.
67. Ranz JM, González PM, Clifton BD, Nazario-Yepiz NO, Hernández-Cervantes PL, Palma-Martínez MJ, et al. A de novo transcriptional atlas in Danaus plexippus reveals variability in dosage compensation across tissues. Commun Biol. 2021;4:791.
68. Li X, Fan D, Zhang W, Liu G, Zhang L, Zhao L, et al. Outbred genome sequencing and CRISPR/Cas9 gene editing in butterflies. Nat Commun. 2015;6:8212.
69. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
70. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
71. Shumate A, Wong B, Pertea G, Pertea M. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. PLoS Comput Biol. 2022;18:e1009730.
72. Harrison PW, Amode MR, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, et al. Ensembl 2024. Nucleic Acids Res. 2024;52:D891–9.

73. Copley SD. Evolution of new enzymes by gene duplication and divergence. FEBS J. 2020;287:1262–83.
74. Axén A, Carlsson A, Engström A, Bennich H. Gloverin, an antibacterial protein from the immune hemolymph of Hyalophora pupae. Eur J Biochem. 1997;247:614–9.
75. Hwang J, Kim Y. RNA interference of an antimicrobial peptide, gloverin, of the beet armyworm, Spodoptera exigua, enhances susceptibility to Bacillus thuringiensis. J Invertebr Pathol. 2011;108:194–200.
76. Sparks ME, Blackburn MB, Kuhar D, Gundersen-Rindal DE. Transcriptome of the Lymantria dispar (gypsy moth) larval midgut in response to infection by Bacillus thuringiensis. PLoS ONE. 2013;8:e61190.
77. Urban JM, Foulk MS, Bliss JE, Coleman CM, Lu N, Mazloom R, et al. High contiguity de novo genome assembly and DNA modification analyses for the fungus fly, Sciaracoprophila, using single-molecule sequencing. BMC Genomics. 2021;22:643.
78. Weisman CM, Murray AW, Eddy SR. Many, but not all, lineage-specific genes can be explained by homology detection failure. PLoS Biol. 2020;18: e3000862.
79. Weisman CM, Murray AW, Eddy SR. Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. Curr Biol. 2022;32:2632–9.e2.
80. Duplouy A, Hornett EA. Uncovering the hidden players in Lepidoptera biology: the heritable microbial endosymbionts. PeerJ. 2018;6:e4629.
81. Jiggins FM, Hurst GD, Jiggins CD, d Schulenburg JH v, Majerus ME. The butterfly Danaus chrysippus is infected by a male-killing Spiroplasma bacterium. Parasitology. 2000;120(Pt5):439–46.
82. Husnik F, Nikoh N, Koga R, Ross L, Duncan RP, Fujie M, et al. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. Cell. 2013;153:1567–78.
83. Cohen O, Gophna U, Pupko T. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol Biol Evol. 2011;28:1481–9.
84. Chen CK-M, Chan N-L, Wang AH-J. The many blades of the β-propeller proteins: conserved but versatile. Trends Biochem Sci. 2011;36:553–61.
85. Fiegler H, Bassias J, Jankovic I, Brückner R. Identification of a gene in Staphylococcus xylosus encoding a novel glucose uptake protein. J Bacteriol. 1999;181:4929–36.
86. Ioannidis P, Buer B, Ilias A, Kaforou S, Aivaliotis M, Orfanoudaki G, et al. A spatiotemporal atlas of the lepidopteran pest Helicoverpa armigera midgut provides insights into nutrient processing and pH regulation. BMC Genomics. 2022;23:75.