PROCEEDINGS B

rspb.royalsocietypublishing.org

Research

Cite this article: Smith G, Kelly JE, Macias-Muñoz A, Butts CT, Martin RW, Briscoe AD. 2018 Evolutionary and structural analyses uncover a role for solvent interactions in the diversification of cocoonases in butterflies. *Proc. R. Soc. B* **285**: 20172037. http://dx.doi.org/10.1098/rspb.2017.2037

Received: 14 September 2017
Accepted: 1 December 2017

Subject Category:
Evolution

Subject Areas:
evolution, genomics, biochemistry

Keywords:
cocoonase, pollen feeding, *Heliconius*, gene duplication, serine protease

Author for correspondence:
G. Smith
e-mail: gilbert.smith@bangor.ac.uk

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3954088.

THE ROYAL SOCIETY
PUBLISHING

Evolutionary and structural analyses uncover a role for solvent interactions in the diversification of cocoonases in butterflies

G. Smith[1,7], J. E. Kelly[2], A. Macias-Muñoz[1], C. T. Butts[3,4,5], R. W. Martin[2,6] and A. D. Briscoe[1]

[1]Department of Ecology and Evolutionary Biology, [2]Department of Chemistry, [3]Department of Sociology, [4]Department of Statistics, [5]Department of Electrical Engineering and Computer Science, and [6]Department of Molecular Biology and Biochemistry, University of California, Irvine, CA 92697, USA
[7]School of Biological Sciences, Bangor University, Brambell Laboratories, Bangor, Gwynedd, UK

Multi-omic approaches promise to supply the power to detect genes underlying disease and fitness-related phenotypes. Optimal use of the resulting profusion of data requires detailed investigation of individual candidate genes, a challenging proposition. Here, we combine transcriptomic and genomic data with molecular modelling of candidate enzymes to characterize the evolutionary history and function of the serine protease cocoonase. *Heliconius* butterflies possess the unique ability to feed on pollen; recent work has identified *cocoonase* as a candidate gene in pollen digestion. *Cocoonase* was first described in moths, where it aids in eclosure from the cocoon and is present as a single copy gene. In heliconiine butterflies it is duplicated and highly expressed in the mouthparts of adults. At least six copies of *cocoonase* are present in *Heliconius melpomene* and copy number varies across *H. melpomene* sub-populations. Most *cocoonase* genes are under purifying selection, however branch-site analyses suggest *cocoonase 3* genes may have evolved under episodic diversifying selection. Molecular modelling of cocoonase proteins and examination of their predicted structures revealed that the active site region of each type has a similar structure to trypsin, with the same predicted substrate specificity across types. Variation among heliconiine cocoonases instead lies in the outward-facing residues involved in solvent interaction. Thus, the neofunctionalization of *cocoonase* duplicates appears to have resulted from the need for these serine proteases to operate in diverse biochemical environments. We suggest that *cocoonase* may have played a buffering role in feeding during the diversification of *Heliconius* across the neotropics by enabling these butterflies to digest protein from a range of biochemical milieux.

1. Introduction

Adaptive evolution may occur through gene duplication events followed by neofunctionalization of the derived copy [1]. The serine protease *cocoonase* was first described in silkmoths, where it plays a key role in adult eclosion by degrading the sericin proteins that hold together the cocoon's silk fibres [2]. *Cocoonase* occurs in moths as a single copy gene, but recent work has identified multiple *cocoonase* duplication events in the *Heliconius melpomene* genome, resulting in at least five duplicates of recent origin [3]. The retention of gene duplicates in butterflies that do not spin a silk cocoon, and thus do not require chemical degradation during eclosion, suggests a novel function for these protease duplications. Furthermore, these novel proteins could have strong proteolytic properties that may have commercial value, particularly for degumming in the silk production industry or other applications requiring protein degradation.

The cocoonase enzyme has been intensively studied in silkmoths, including in *Antheraea polyphemus* [4–7], *A. peryni* [2,5,8,9], *A. mylitta* [5,10] and *Bombyx mori*

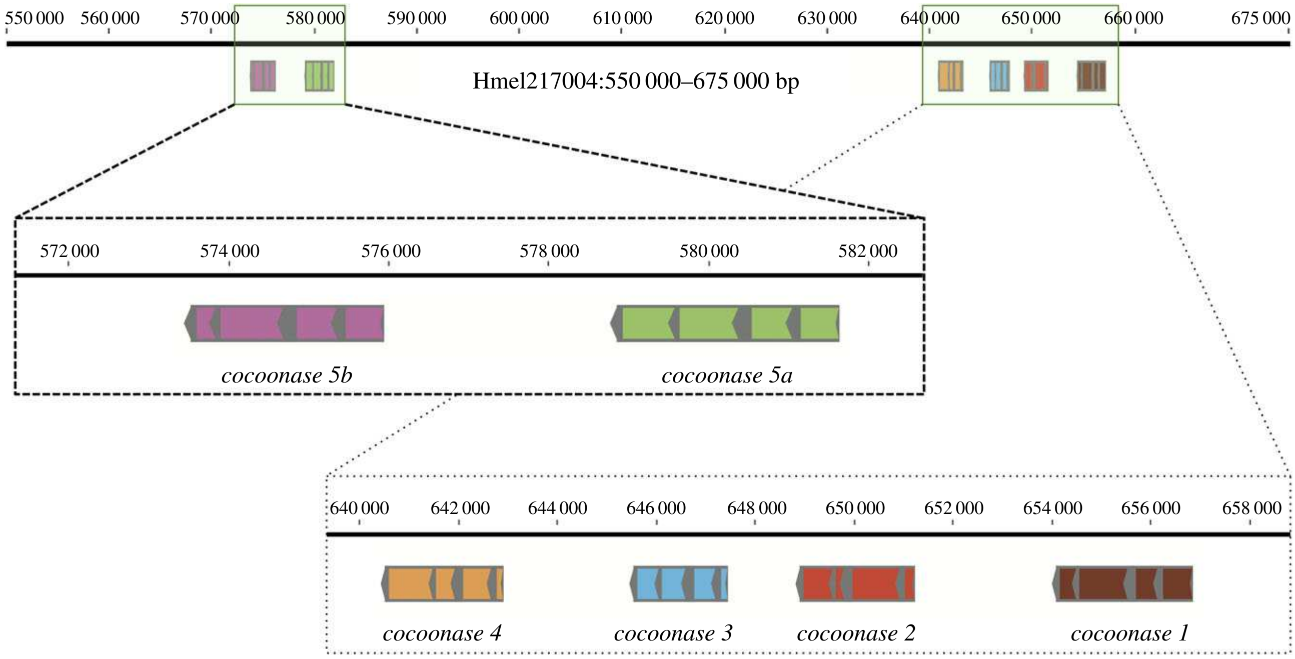

Figure 1. Genomic annotations of the *cocoonase* genes in *Heliconius melpomene*. Annotations indicate the positions of six duplicate copies of *cocoonase* located within the genomic coordinates 550 000–675 000 on scaffold Hmel217004, genome assembly version Hmel2 [21]. Each gene is coloured separately and exons are indicated in grey in the 5′ to 3′ direction (arrows).

[11,12]. In silkmoths, cocoonase is secreted directly by modified maxillary galeae (the proboscis), rather than, or perhaps in conjunction with [12], the digestive tract [2]. The cocoonase enzyme is deposited in a dry crystalline form on the surface of the proboscis by large specialized polyploid cells within the galeae (zymogen cells). The zymogen proenzyme is exuded through narrow cuticular ducts that connect to the polyploid cells to form a duct and valve apparatus linked to a large extracellular storage vacuole [13]. It is then dissolved on the surface of the proboscis by an alkaline solvent principally composed of aqueous potassium bicarbonate. Cocoonase has a strong and somewhat general proteolytic activity and is capable of digesting a wide range of proteins [2,8]. In pupating moths the enzyme is applied to the inside of the cocoon, degrading sericin and allowing the adult to escape.

Our characterization here of the heliconiine cocoonases as trypsin-like proteases of MEROPS family S1 [14] is consistent with previous studies of cocoonase from *Antheraea pernyi* [15]. Serine proteases act by cleaving peptide bonds in their targets, and play a range of physiological roles including in digestion, blood coagulation, signal transduction, reproduction and the immune response. Several categories of serine protease have been identified based on structure and substrate specificity (e.g. trypsin-like, chymotrypsin-like, elastase-like), with specificity driven by the identity of key residues in the specificity pocket [16,17]. All serine proteases hydrolyse peptide bonds using a Ser-His-Asp catalytic triad at the centre of the active site cleft. Binding sites adjacent to this triad comprise the specificity pocket; this region and adjacent loops help determine the enzymes' substrate preferences [17]. Cocoonase is expressed intracellularly as a preproenzyme 260 amino acids in length [18]. The enzyme is activated upon secretion, which entails removal of the N-terminal signal peptide [5,19] and the pro-sequence that prevents the enzyme from becoming prematurely activated. This is probably accomplished through autocatalysis [8]; in cocoonases a K or R residue at the end of the pro-sequence is followed by a conserved IVGG motif at the N-terminal end of the mature enzyme, consistent with trypsin-type cleavage. Amino acid residues forming the catalytic triad of serine proteases (Ser-His-Asp) have been seen in the cocoonase active site in moths [20]. Studies investigating cocoonases are limited to a few silkmoth species, and up until recently there have been little data from other lepidopteran species.

The *cocoonase* gene is a single-copy gene in several butterfly and moth genomes (the silkmoth *Bombyx mori*, diamond backed moth *Plutella xylostella* and monarch butterfly *Danaus plexippus*, and the Glanville fritillary *Melitaea cinxia* [3]). However, heliconiine butterflies harbour at least five duplicate copies (figure 1), including recent duplication events unique to *Heliconius* butterflies. Smith *et al*. [3] discovered multiple paralogues of *cocoonase* mRNA are upregulated in the proboscis of heliconiines when compared to two other tissues (antennae and legs). Further, they saw that these *cocoonases* are not expressed in the salivary glands of *Heliconius melpomene*, suggesting that, like moths, butterflies directly secrete this digestive enzyme from the proboscis. This is further supported by the presence of cocoonase in the saliva of *H. melpomene* adults [22]. Unlike moths, adult butterflies express multiple, sequence-divergent versions of *cocoonase* in their mouthparts. This, and the fact that butterflies do not pupate within a silk cocoon but escape with relative ease from a chrysalis, suggests that cocoonase may have a different function in butterfly adults post-emergence. One potential function is the pre-digestion of pollen granules during feeding. *Heliconius* butterflies directly feed on pollen by collecting and digesting pollen on their proboscides, a behaviour that is not seen in other butterflies [23,24]. Nectar also contains small amounts of amino acids from dissolved pollen, and thus heliconiine butterflies may have coopted cocoonase for the digestion of peptides found in their natural diet, and *Heliconius* butterflies may use these specifically for feeding on pollen.

New cocoonase enzymes have potential commercial value across numerous applications. In the silk industry, degumming

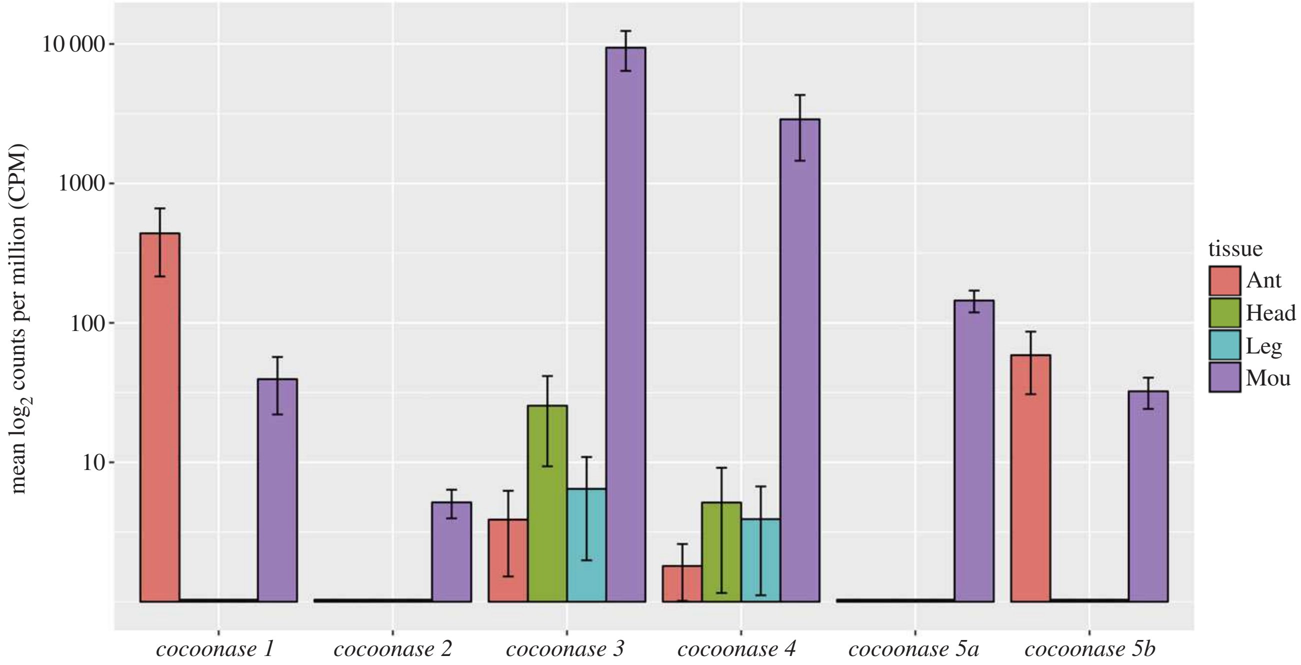

Figure 2. Comparison of the mRNA expression levels of *cocoonases* across multiple *H. melpomene* tissues. Bars are mean plus standard error of $\log_2$ CPM across four tissue types (14 individuals): antennae (Ant), head (Head), legs (Leg) and mouthparts (Mou).

of silk by sericin removal is carried out in order to improve the quality of the fibroin silk fibres [25]. Current industrial methods for removing sericin involve chemical incubation (e.g. with alkaline solutions), which also degrades fibroin, reducing silk quality [8]. Because cocoonase hydrolyses sericin but leaves the fibroin fibres untouched the enzyme is of great interest to the silk producing industry. The discovery of duplicate copies of *cocoonase* with divergent amino acid sequences might lead to the discovery of a protease that is more effective at degumming. Proteases are also useful for medical applications; recent work on the Chinese silkworm cocoonase has shown that it cleaves fibrin and fibrinogen both *in vitro* and in an animal model of thrombosis [15], demonstrating its utility for studying thrombosis and potentially treating blood clots. As we show below, the heliconiine cocoonases exhibit substantial variation in surface properties (particularly hydrophobicity), potentially facilitating their diffusion into diverse chemical environments. Such a mixture of enzymes with differing surface characteristics could prove useful in degrading complex proteinaceous material in the presence of chemical detergents, a common requirement of enzymatic cleaning agents.

Here, we use multiple sources of -omic data in order to examine the evolutionary history and functional properties of the heliconiine *cocoonase* gene duplications. Specifically, using new transcriptomic and genomic data, we examine their expression across additional tissues, reconstruct their phylogenetic relationships, and examine rates of gene duplication and deletion. Because the heliconiine *cocoonase* duplications have been retained, and their diversification appears to be ongoing, we hypothesized that duplicates have undergone neofunctionalization. Our initial hypothesis involved their enzyme products acting on different substrates. We use comparative modelling and protein structure analysis to infer functional (and perhaps adaptive) differences when compared with the single-copy moth *cocoonase*. Our modelling data of 30 individual cocoonases indicate that, contrary to our hypothesis, all the cocoonase enzymes have trypsin-like specificity, while significant differences are found among the surface residues of different cocoonase types, suggesting enzyme adaptation to different chemical environments.

2. Results and discussion

(a) Heliconiine cocoonases exhibit copy-number variation

Smith *et al*. [3] suggested that the diversification of *cocoonase* genes through duplication might be linked to the ability of *Heliconius* butterflies to feed on pollen. The annotation of *cocoonase* genes from a new *H. melpomene* genome assembly (Hmel2) suggests that duplication events have given rise to six gene duplicates in this species (figure 1). Expression of *cocoonase* occurs primarily in the mouthpart tissues (proboscis and labial palps) of butterflies, where mRNA levels are orders of magnitude greater than that of other tissues. All six *cocoonase* genes are expressed in the mouthparts of *H. melpomene* to some degree, with *cocoonase 3* and *cocoonase 4* being highly expressed compared with the other 4 gene duplicates (figure 2). Such high expression levels could be indicative of an important function for *3* and *4* in the mouthparts tissues. *Cocoonase* expression in other *H. melpomene* tissues is limited to *cocoonase 1* and *cocoonase 5b* in the antennae, and low-level expression of *3* and *4* in antennae, head and legs.

Phylogenetic reconstruction of the coding nucleotide sequences of *cocoonase* genes in butterflies and moths reveals multiple duplication events, including in the non-pollen feeding species, *H. aoede* (electronic supplementary material, figure S1). Four major clusters were detected, with duplications of *cocoonase 3*/*4* and *5* clusters being unique to *Heliconius* [3] (electronic supplementary material, figure S1). Several duplication events predate heliconiine divergence and the evolution of pollen feeding, suggesting that *Eueides* may use *cocoonases* for feeding-related proteolysis if not for pollen

Table 1. Copy number variation in *cocoonase* genes across 18 individuals of four *Heliconius melpomene* (*H. m.*) subspecies.

	H. m. rosina		*H. m. melpomene*		*H. m. amaryllis*		*H. m. aglaope*	
	duplication	deletion	duplication	deletion	duplication	deletion	duplication	deletion
Cocoonase 1	0	0	0	0	0	0	1	0
Cocoonase 2	0	0	0	0	0	0	1	0
Cocoonase 3	0	0	0	0	0	0	2	2
Cocoonase 4	0	0	0	0	0	0	0	1
Cocoonase 5a	0	2	1	2	2	3	0	1
Cocoonase 5b	0	1	1	0	1	0	0	0
total samples	4		6		4		4	

feeding, and this may have facilitated the evolution of pollen feeding in *Heliconius* [3]. The rate of synonymous substitutions between pairs of sequences mirrored this clustering; pairwise comparisons between *cocoonase* sequences grouped by cluster were substantially lower within (mean of 0.3 for heliconiine *cocoonases*; electronic supplementary material, table S1) than between clusters (mean of 1.92; electronic supplementary material, table S2). Although not every species's transcriptome has a representative copy from each cluster, each species has at least three and as many as six paralogues. Particularly interesting is the *cocoonase 3*/*4* cluster containing copies *3* and *4* that are of more recent origin, and highly expressed. Copies *5a* and *b* have a very recent origin, with highly similar nucleotide and amino acid sequences, suggesting that copy number variation (CNV) could exist in populations.

In order to examine population structural variation we analysed 18 *H. melpomene* re-sequenced genomes across four different subspecies. CNV was detected for all *cocoonase* copies (table 1), but was particularly evident for *cocoonase 5a* and *5b*. Duplication events were detected for *cocoonase 1-3* in the *agalope* subspecies, and *5a* and *5b* in *amaryllis* and *melpomene*. Deletions were seen for *cocoonase 3* and *4* in the *agalope* subspecies, *cocoonase 5a* in all four subspecies, and *5b* in *rosina*. Deletions in *cocoonases 1* and *2* were not detected in any subspecies. Smaller-scale differences (insertions/deletions) were seen frequently across all genes (electronic supplementary material, tables S3 and S4). Thus, *cocoonase* gene duplications vary in number across *Heliconius* populations, varying more when they are recent, and are probably maintained in populations, indicating their functional importance.

Clustering of the heliconiine cocoonase protein sequences resulted in a clear division into subfamilies comprising type 1, type 2, types 3 and 4, and type 5 (electronic supplementary material, figure S2). These enzymes cluster according to type rather than by species, consistent with the idea that paralogues have different functions. Indeed, the majority of branches on the nucleotide tree demonstrate signals of purifying selection, suggesting that selection has acted to maintain their function (electronic supplementary material, table S5). However, a branch-site REL test [26] revealed two branches with sites showing enhanced rates of non-synonymous substitutions that are likely to be due to episodic diversifying selection: the branch leading to the *cocoonase 3* clade (corrected p-value = 0.025) and *Eueides isabella cocoonase 2* (corrected p-value = 0.001; Holm–Bonferroni corrected p-value threshold of less than 0.05; electronic supplementary material, table S5).

(b) Heliconiine cocoonase clusters represent functional subfamilies

In order to compare the sequence diversity among and between the heliconiine cocoonases and more distant relatives, two sequence alignments are presented (electronic supplementary material, figures S3 and S4), annotated to indicate amino acid properties, catalytic residues and functional sequence regions. The aligned sequences of the outgroup enzymes with those of representative cocoonases from *Heliconius melpomene*, and details of the region including the specificity pocket are shown in electronic supplementary material, figures S3 and S5. Electronic supplementary material, figure S4 shows a similar alignment for every heliconiine cocoonase. These alignments show a general overview of sequence elements that are broadly conserved and hence likely to be functionally essential for this enzyme class, and those that are more specific to the heliconiine butterflies. Generally, the cocoonase catalytic triad residues are conserved in 28 of the 30 cocoonases examined; the exceptions are *E. isabella* cocoonase 2, in which the active His is mutated to Ile, and *H. sara* cocoonase 3, which is truncated at the C-terminal end (the full mRNA was not recovered). Conservation of the key catalytic residues indicates that almost all of these enzymes are functional serine proteases.

Our initial hypothesis about the function of cocoonases in *Heliconius* butterflies focused on the possibility of diversification for different substrate preferences following gene duplication. However, this hypothesis is not supported by examination of the sequence alignments; all the cocoonases investigated here share sequence elements that are common to enzymes with trypsin-like functionality, where the peptide backbone is cleaved following a positively charged Arg or Lys residue in the P1′ position. In particular, the Asp residue at the bottom of the specificity pocket (D189 in trypsin and D212 in *D. plexippus* cocoonase) is conserved in all cocoonases. The presence of a negatively charged residue in this location is essential for positioning the substrate in trypsin-like enzymes. Other serine proteases with different substrate cleavage preferences are characterized by different residues in this critical position, e.g. Ser for chymotrypsin.

In addition to this critical residue, it has long been recognized that other features also affect serine protease activity and specificity. For example, preference for Arg versus Lys is modulated by whether the residue immediately following the Asp is Ala or Ser [27]; examples of both are found in

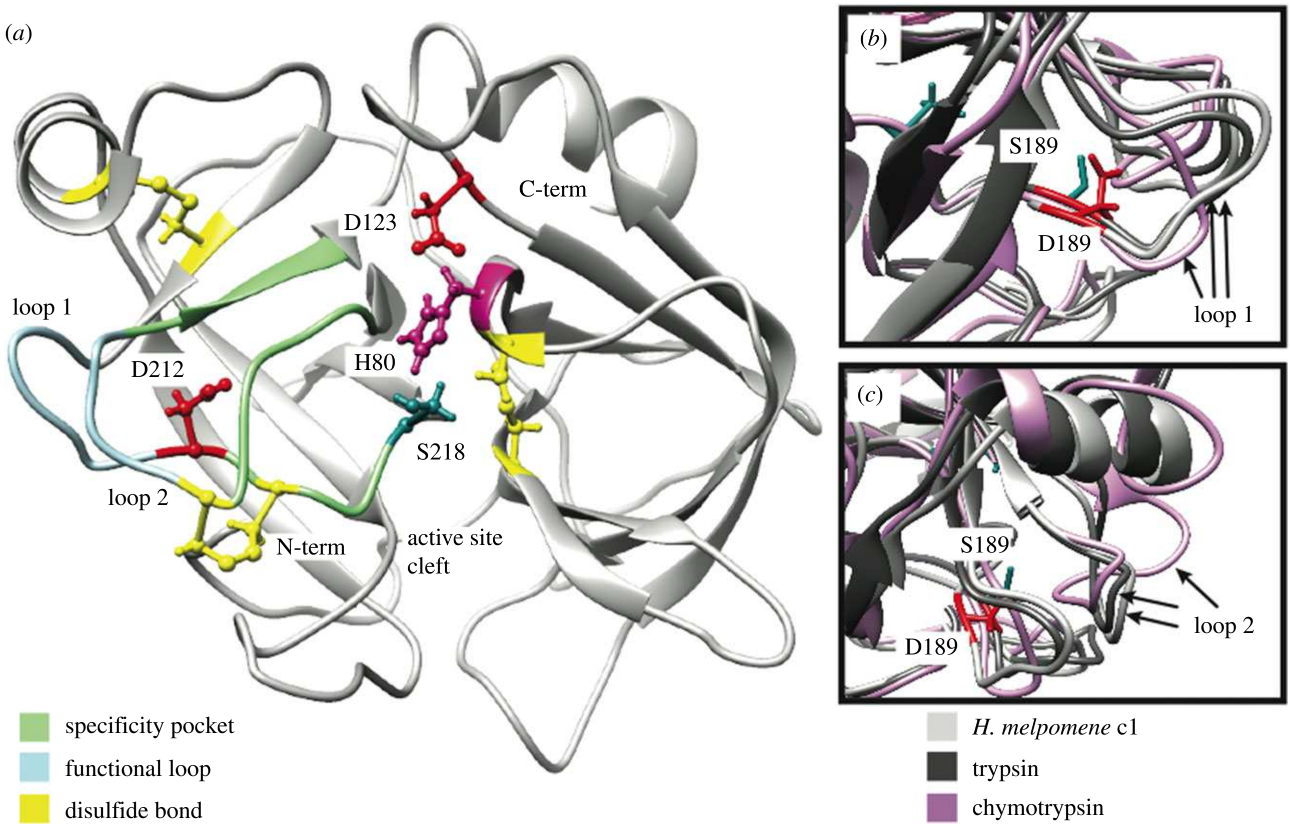

Figure 3. (*a*) Equilibrated molecular model of the active form of *D. plexippus cocoonase*, labelled to show the essential features of a trypsin-like serine protease. The catalytic triad residues H80, D123 and S218 (zymogen numbering) are shown as ball and stick models. The presence of a negatively charged residue (D212) in the bottom of the specificity pocket (green) indicates that this enzyme, like the other cocoonases, has a preference for a positively Arg or Lys in the position immediately prior to the cut site. The loops surrounding the pocket (blue) are also functionally important. The three conserved disulfide bonds are shown in yellow. (*b*,*c*) Comparison of *H. melpomene* cocoonase 1 (c1; light grey) to trypsin (dark grey) and chymotrypsin (lavender).

the set of sequences examined here. Unusually, *E. isabella* cocoonase 2 has Tyr in this position, however it is presumed to be inactive due to mutation of the catalytic His residue. In trypsin a Tyr residue (Y172) interacts with the specificity pocket and surface loops to further influence substrate specificity [28]. This Tyr is conserved in all of the heliconiine cocoonases, but is replaced with Leu in *D. plexippus* cocoonase and *M. sexta* SP54, and with Phe in *B. mori* cocoonase. Although some variation does exist among the lepidopteran cocoonases at this secondary site, overall the heliconiine cocoonases are extremely consistent in their sequence identity to trypsin at functional positions. Therefore, the differences in their functions and hence the reason for their diversification must lie elsewhere. In order to investigate the potential role of three-dimensional structural features, molecular models were calculated for the cocoonases. A similar approach has been used to perform structural comparisons of clip domains from immune system serine proteases from *M. sexta* [29].

(c) Molecular modelling predicts high structural homology to trypsin

The molecular models of cocoonase are characterized by a trypsin-like fold, as illustrated for a representative example (*D. plexippus* cocoonase; figure 3*a*). Functional features examined include the length, sequence and positioning of the surface loops [30,31], and the conformation and dynamics [32] of the backbone around a glycine residue near the entrance to the specificity pocket (Gly 216 in trypsin) [33]. Figure 3*b*,*c* show the surface loops for a representative cocoonase (*H. melpomene* cocoonase 1) in comparison with trypsin and chymotrypsin. Although there are minor differences in backbone position, for both loops the conformation adopted is clearly that of trypsin. This is true for all of the full-length cocoonase models examined here.

Strong conservation of sequence and structural properties in the cocoonase specificity pockets is not limited to the stabilizing Asp residue; examination of the sequence alignments (electronic supplementary material, figures S3 and S4, detail in figure S5) reveals significant variation in only three sites in the specificity pocket itself, one of which is the Ser/Ala in position 188 (trypsinogen numbering). The side chains of the other residues involved (W215 and D217 in trypsinogen) point outwards, away from the substrate-binding pocket. Furthermore, the conformation about G216, a critical structural residue involved in regulating substrate binding [33], is similar to that of trypsin in the cocoonases. Based on the sequence and structural evidence, the evolution of these enzymes does not seem to be the result of pressure to produce proteases with different substrate specificity. Therefore, understanding the functional origin of cocoonase gene duplication in pollen-feeding butterflies must focus on other structural and sequence features.

(d) Diversity of surface residues suggests specialization for different environments

Within each cocoonase type, sequences from different species demonstrate a high level of conservation (electronic supplementary material, figure S6). Figure 4 illustrates the

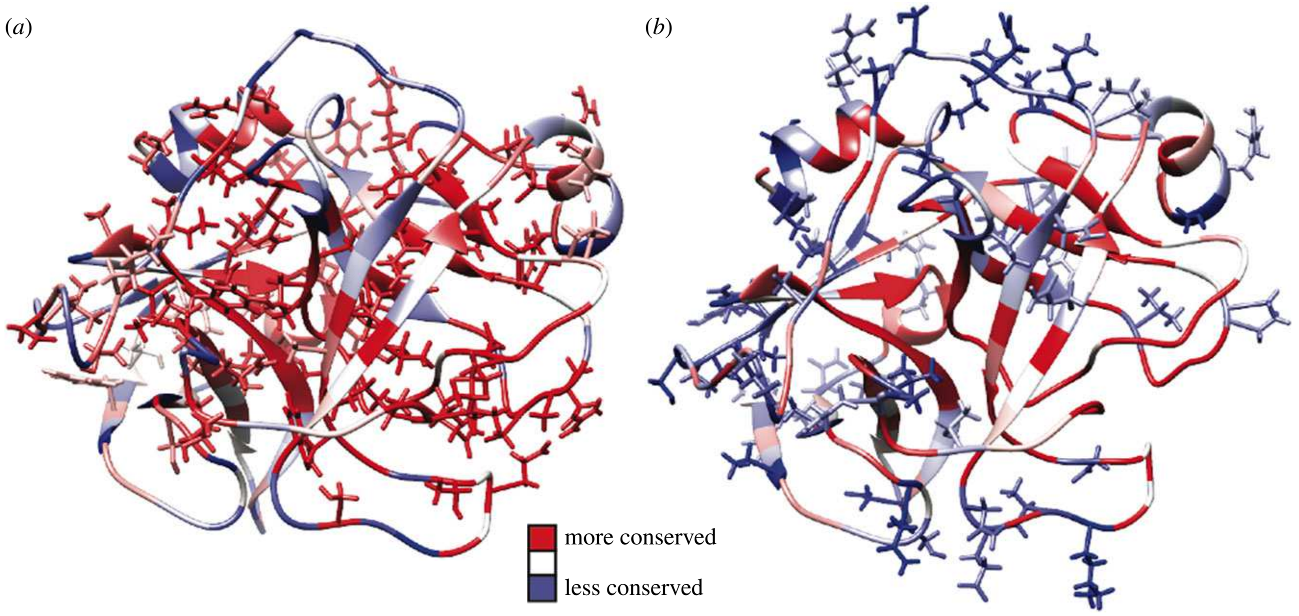

Figure 4. Conservation map of all cocoonase sequences plotted on the molecular model of *H. melpomene* cocoonase 1. The percentage conservation for each position ranges from red (most conserved) to blue (least conserved). (*a*) Side chains for the most conserved residues. (*b*) Side chains for the least conserved residues.

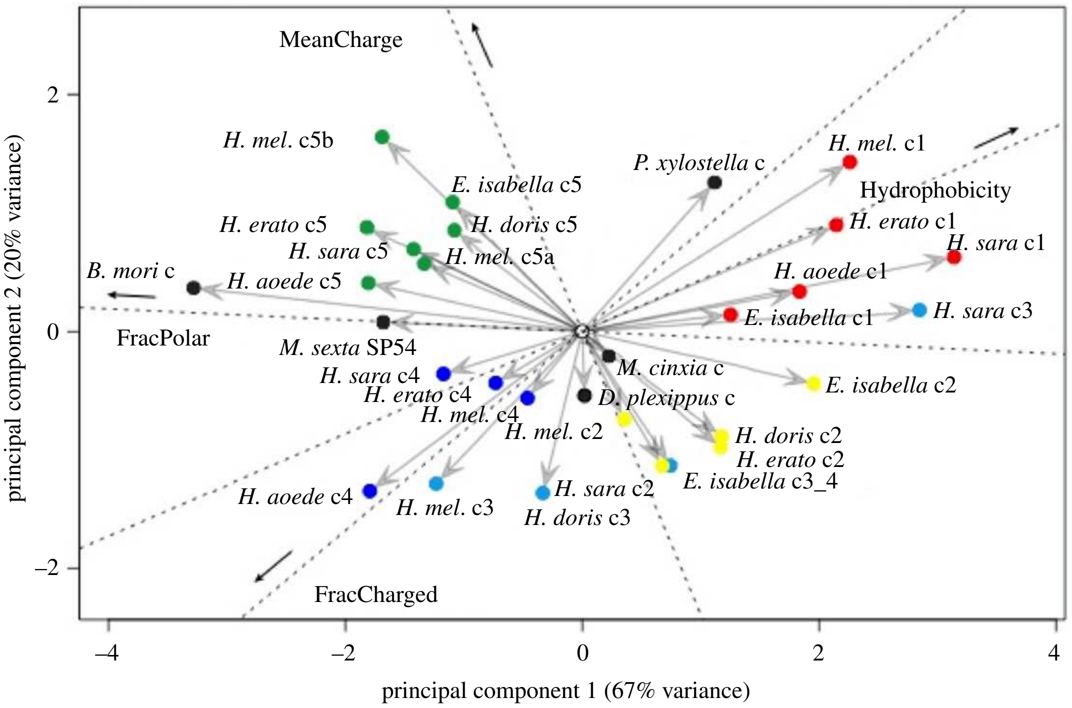

Figure 5. PCA of solvation-relevant surface residue properties. Projections of original variables into the PCA space are shown by dotted lines; arrows indicate positive direction. Original variables include overall mean residue charge (MeanCharge), fraction of residues that are charged (FracCharged), fraction of residues that are polar (FracPolar) and overall hydrophobicity (Hydrophobicity). Colours indicate protein sequence-based clustering (see electronic supplementary material, figure S2 for clusters).

sequence conservation for all cocoonases, plotted on the structure of *H. melpomene* cocoonase 1. Strikingly, the side chains of most of the highly conserved residues are directed toward the interior of the protein, while many of the variable residues project into the solvent. This pattern suggests that the evolution of different cocoonase types may have been driven by the need to diffuse into and remain solvated in different environments, rather than by differences in substrate specificity.

The observation that cocoonases differ primarily in their outward-facing residues suggests that functional differences may alter their surface properties; such an adaptation could arise, for example, in response to the need to efficiently diffuse through the chemically heterogeneous materials that comprise pollen grains. The properties most likely to vary in such a scenario are residue charge, overall hydrophobicity and whether or not the residue side chain is polar, all of which influence solvation and ability to diffuse within relatively polar versus non-polar environments. Principal component analysis (PCA) of solvation-relevant surface properties for the protein set (figure 5) shows that the five sequence-based clusters clearly follow a well-ordered, circumplex pattern from those whose surfaces are relatively non-polar and hydrophobic (cocoonase 1) to those with more moderate levels of hydrophobicity and relatively negative mean charge (cocoonases 2 and 3), those that are relatively polar, hydrophilic, and neutral (cocoonase 4), and those that are polar and hydrophilic with a relatively positive mean charge (cocoonase 5). With the exception of cluster 3, the five clusters are well separated in PCA space, and there is no indication of clustering by species (as might be expected if surface characteristics were primarily a response to species-specific feeding requirements). Taken together, these results are consistent with the hypothesis that differences in sequence among cocoonases are primarily driven by a requirement for diversity in solvation-related properties, and that this requirement is broadly shared among pollen-feeding lepidopterans.

3. Concluding remarks

Several of the *cocoonase* gene duplication events appear to predate the divergence of heliconiine butterflies with most of the resulting copies being retained, at least certainly in *Heliconius* butterflies. The tree topology and selection tests suggest the neofunctionalization of these genes, with *cocoonase 3*, a *Heliconius*-specific duplication, showing evidence of episodic diversifying selection. *Cocoonase 2* may also have undergone selection in *E. isabella*, however this gene is likely to be inactive due to a mutation in the active site. Overall, this non-pollen feeding sister group to *Heliconius* has fewer expressed or active copies of *cocoonase*, suggesting the importance of *cocoonase* genes specifically within *Heliconius* butterflies.

Cocoonase is highly expressed in the mouthparts of butterflies and the function of this protease in *Heliconius* is likely to involve feeding, specifically pollen feeding. Solvent exposed surface residues tend to have a higher mutation rate than slower evolving interior residues, which are constrained by their important roles in structures such as the active site [34]. While surface residues may play a role in substrate and modulatory ligand recognition when associated with the active site and/or specialized regions such as binding pockets [35], we see no evidence of systematic variation in such residues among the heliconiine cocoonases. Variation is instead seen in residues comprising the bulk of the protein surface, which are more typically implicated in stability and solvation within particular chemical environments [36,37]. Cocoonase surface residue diversity may therefore be important for solvation and stability in the heterogeneous mix of chemical microenvironments comprising pollen grains [38], which is further increased by biochemical differences in the plants on which they feed. *Heliconius* butterflies visit many different plant species within their home range for nectar and pollen, and plant diversity varies geographically across Mexico and Central and South America [39,40]. The duplication, retention and maintenance of cocoonase genes thus might have a buffering effect, allowing effective proteolytic activity in diverse chemical environments, evolving alongside, or perhaps playing a role in, the successful diversification of these butterflies throughout the neotropics.

Some evidence for pollen-associated specialization exists within *Heliconius*, however this may be related more to pollen collection than digestion. Analyses of pollen load show the most common feeding plants include *Psiguria*, *Gurania*, *Lantana* and *Psychotria* species [40–42]. Some species groups collect pollen from specific plants; for example, *melpomene* group species (*H. hecale*, *H. ismenius*, *H. melpomene*, *H. cydno*) collect larger amounts of pollen from *Psiguria* and smaller amounts from *Lantana* species, compared with the *erato* group (*H. erato*, *H. sapho*, *H. sara* [40–42]). This is likely to be connected to pollen grain size, and perhaps proboscis morphology, as *melpomene* group species prefer large pollen grains, and *Psiguria* pollen is larger [43]. However, evolutionary generalization in terms of chemical environment may exist because butterflies feed on multiple plant species within an assemblage, regardless of whether the assemblage has large or small pollen.

The expansion of *cocoonase* genes is likely to be linked to the evolution of pollen processing behaviour, which is unique to *Heliconius*. Further work is needed to understand the origins of this very specialized behaviour, and should focus on the combination of morphological, behavioural and digestive changes that underlie it. Further, little information exists on the variation in pollen grain composition of plants on which *Heliconius* species feed. Such data would help to elucidate the environments in which cocoonase is expressed. Here, we have shown that a combined sequence analysis and molecular modelling approach can uncover insights into the evolutionary history and functional diversity of highly duplicated candidate genes. More widespread use of such structural modelling techniques could help shed light on these and other questions that hinge on the interplay of evolutionary and structural or biochemical factors.

4. Material and methods

(a) *Heliconius melpomene cocoonase* gene annotations

Cocoonase genes were annotated as per Smith *et al.* [3] using the recently updated *Heliconius melpomene* genome assembly (Hmel2 [21]).

(b) RNA-Seq library preparation and sequencing

RNA was extracted from the head (excluding antennae and mouthparts) of a male *H. aoede* collected in La Merced, Peru by TRIzol extraction (Life Technologies, Grand Island, NY) and purified using the NucleoSpin RNA II kit (Macherey-Nagel, Bethlehem, PA). Libraries were prepared using the TruSeq RNA Preparation kit (Illumina, San Diego, CA) and sequenced at the Princeton Core Facility on a HiSeq 2000 (Illumina, San Diego, CA).

(c) *Cocoonase* tree reconstruction

Cocoonase sequences of four *Heliconius* species and *E. isabella* were identified in transcriptome assemblies from Smith *et al.* [3] (Dryad doi:10.5061/dryad.8d724), and in the non-pollen feeding *H. aoede* from a *de novo* transcriptome assembled using Trinity (version r2012-06-08 [44]). Sequences were aligned with orthologues from the genome assemblies of outgroup species: *Manduca sexta*, *Bombyx mori*, *Plutella xylostella*, *Danaus plexippus* and *Melitaea cinxia*. Tree reconstruction was performed in MEGA v. 5.2.2 [45] from coding nucleotide sequences using maximum likelihood with partial deletion (85% site coverage cut-off), the Tamura 3-parameter nucleotide substitution model (gamma distributed with invariant sites), and 1000 bootstrap replicates. The nucleotide substitution model was selected after testing the fit of 24 different models in MEGA. The amino acid tree was reconstructed using the same methods, under the WAG + G + I substitution model.

(d) Branch-site REL tests

The method of Kosakovsky Pond *et al.* [26] permits the unbiased identification of branches that have a class of site where $dN/dS > 1$, indicating episodic diversifying selection. Briefly, codons in the alignment were removed if any other sequence was missing that codon. Three sequences were excluded entirely due to their short length. All branches were tested using likelihood ratio tests against a model in which no branch has sites with $dN/dS > 1$ (see electronic supplementary material, figure S7 for test tree topology). The resulting *p*-values were corrected using the Holm–Bonferroni methods as implemented in DataMonkey [46].

(e) *Cocoonase* expression levels

Gene expression count data for *H. melpomene cocoonases* were obtained by mapping RNA-Seq libraries from Briscoe *et al*. [47] (ArrayExpress accession: E-MTAB-1500; six individuals) and Macias-Muñoz *et al*. [48] (ArrayExpress accessions: E-MTAB-6249 and E-MTAB-6342; eight individuals) to coding nucleotide sequences using RSEM [49]. Counts were converted to counts per million (CPM) using the total number of reads in each library. Libraries included four biological replicates of males and four females from four tissues: antennae, head, legs (all six) and mouthparts (proboscis and labial palps). Mean CPM was plotted on a $\log_2$ scale using the package ggplot2 in R [50].

(f) *Cocoonase* copy number variation in *Heliconius melpomene*

Reads for 18 resequenced *H. melpomene* genomes generated by Martin *et al*. [51] (European Nucleotide Archive: ERP002440) were aligned to the reference genome [21] using bwa [52] for 4 subspecies (6 *H. melpomene melpomene* samples, 4 *H. melpomene rosina*, 4 *H. melpomene amaryllis*, 4 *H. melpomene aglaope*). Samtools [53] was used to index and sort the read mapping results. Pindel [54] was used to detect potential gene deletions and tandem duplications, and results verified manually (electronic supplementary material, figure S8). Full results are given in electronic supplementary material, tables S3 and S4.

(g) Comparative modelling and protein structure analysis

Sequences were aligned using ClustalOmega [55] (gap open penalty = 10.0, gap extension penalty = 0.05, hydrophilic residues = GPSNDQERK, weight matrix = BLOSUM). The presence and cleavage sites for N-terminal secretion signal sequences were predicted using SignalP 4.1 [56]. Structure prediction was performed in three stages as described by Butts *et al*. [57,58]. Simulation was performed using the CHARMM36 forcefield [59], with each model being energy-minimized for 10 000 iterations and then simulated at 293 K for 500 ps; the final protein conformation was retained for subsequent analysis. For reference sequences for which an experimentally determined structure was available, this was used as the initial starting model (following removal of heteroatoms and protonation using REDUCE [60]). PDB files for all proteins are available from Dryad [61] and listed in electronic supplementary material tables S6 and S7.

Relative solvent accessibility (RSA) values were calculated for all equilibrated structures using DSSP 2.2.1 [62]; residues with RSA values less than 0.2 were regarded as buried, with other residues classified as solvent exposed. For the set of solvent exposed residues within each structure, the fraction of polar residues and charged residues, mean residue charge, and mean hydrophobicity (using the scale of Kyte & Doolittle [63]) were calculated. All data analysis and visualization was performed using R [64]. The first two principal components jointly accounted for 87% of the standardized variance. Projections of the original variables into the PCA space were also calculated to assist with interpretation.

Data accessibility. *H. melpomene* RNA-Seq libraries: ArrayExpress: E-MTAB-6249 and E-MTAB-6342; PDB files for all proteins: Dryad doi:10.5061/dryad.355qk [61]. *H. melpomene* sequencing libraries from Martin *et al*. [51]: European Nucleotide Archive: ERP002440. Transcriptome assemblies: Dryad doi:10.5061/dryad.8d724 [3]. RNA-Seq libraries from Brisoce *et al*. [47]: ArrayExpress: E-MTAB-1500. Additional datasets supporting this article have been uploaded as part of the electronic supplementary material.

Authors' contributions. A.D.B., C.T.B. and R.W.M. designed the research, G.S., A.D.B. and A.M.-M. collected data, G.S., A.D.B., A.M.-M., R.W.M., C.T.B. and J.E.K. analysed the data, and G.S., A.D.B., A.M.-M., R.W.M., C.T.B. and J.E.K. wrote and edited the paper.

Competing interests. We declare we have no competing interests.

Funding. This work was supported by National Science Foundation grants IOS-1656260 to A.D.B. and DMS-1361425 to C.T.B. and R.W.M.

Acknowledgements. We thank Ying Zhen and Furong Yuan for RNA-Seq technical assistance and Robert Reed for the *H. aoede* specimen.

References

1. Zhang J. 2003 Evolution by gene duplication: an update. *Trends Ecol. Evol.* **18**, 292–298. (doi:10.1016/S0169-5347(03)00033-8)
2. Kafatos FC, Williams CM. 1964 Enzymatic mechanism for the escape of certain moths from their cocoons. *Science* **146**, 538–540. (doi:10.1126/science.146.3643.538)
3. Smith G, Macias-Muñoz A, Briscoe AD. 2016 Gene duplication and gene expression changes play a role in the evolution of candidate pollen feeding genes in *Heliconius* butterflies. *Genome Biol. Evol.* **8**, 2581–2596. (doi:10.1093/gbe/evw180)
4. Berger E, Kafatos FC. 1971 Immunochemistry of an insect protease, cocoonase, and its zymogen. *Immunochemistry* **8**, 391–403. (doi:10.1016/0019-2791(71)90502-7)
5. Kramer KJ, Felsted RL, Law JH. 1973 Cocoonase. V. Structural studies on an insect serine protease. *J. Biol. Chem.* **248**, 3021–3028.
6. Hruska JF, Felsted RL, Law JH. 1973 Cocoonases of silkworm moths: catalytic properties and biological function. *Insect Biochem.* **3**, 31–43. (doi:10.1016/0020-1790(73)90016-4)
7. Kafatos FC, Kiortsis V. 1971 The packaging of a secretory protein: kinetics of cocoonase zymogen transport into a storage vacuole. *J. Cell Biol.* **48**, 426–431.
8. Kafatos FC, Law JH, Tartakoff AM. 1967 Cocoonase. II. Substrate specificity, inhibitors, and classification of the enzyme. *J. Biol. Chem.* **242**, 1488–1494.
9. Kafatos FC, Tartakoff AM, Law JH. 1967 Cocoonase. I. Preliminary characterization of a proteolytic enzyme from silk moths. *J. Biol. Chem.* **242**, 1477–1487.
10. Felsted RL, Law JH, Sinha AK, Jolly MS. 1973 Properties of the *Antheraea mylitta* cocoonase. *Comp. Biochem. Physiol. B* **44**, 595–609. (doi:10.1016/0305-0491(73)90033-3)
11. Yamazaki Y, Ogawa K, Kanekatsu R. 1995 N-terminal amino acid sequence of cocoonase in the silkworm, *Bombyx mori*. *J. Seric. Sci. Jpn.* **64**, 467–468. (doi:10.11416/kontyushigen1930.64.467)
12. Eguchi M, Iwamoto A. 1975 Role of the midgut, crop, and maxillae of *Bombyx mori* in the production of cocoon-digesting enzyme. *J. Insect Physiol.* **21**, 1365–1372. (doi:10.11416/kontyushigen1930.44.314)
13. Law JH, Dunn PE, Kramer KJ. 1977 Insect proteases and peptidases. In *Adv. Enzymol. Relat. Areas Mol. Biol.* (ed. A Meister), pp. 389–425. Hoboken, NJ, John Wiley & Sons.
14. Rawlings ND, Waller M, Barrett AJ, Bateman A. 2014 MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.*. **42**, D503–D509. (doi:10.1093/nar/gkt953)
15. Geng P, Lin L, Li Y, Fan Q, Wang N, Song L, Li W. 2014 A novel fibrin(ogen)olytic trypsin-like protease from Chinese oak silkworm (*Antheraea pernyi*):

purification and characterization. *Biochem. Biophys. Res. Commun.* **445**, 64–70. (doi:10.1016/j.bbrc.2014.01.155)

16. Ovaere P, Lippens S, Vandenabeele P, Declercq W. 2009 The emerging roles of serine protease cascades in the epidermis. *Trends Biochem. Sci.* **34**, 453–463. (doi:10.1016/j.tibs.2009.08.001)
17. Hedstrom L. 2002 Serine protease mechanism and specificity. *Chem. Rev.* **102**, 4501–4524. (doi:10.1021/cr000033x)
18. Berger E, Kafatos FC, Felsted RL, Law JH. 1971 Cocoonase. III. Purification, preliminary characterization, and activation of the zymogen of an insect protease. *J. Biol. Chem.* **246**, 4131–4137.
19. Felsted RL, Kramer KJ, Law JH, Berger E, Kafatos FC. 1973 Cocoonase. IV. Mechanism of activation of prococoonase from *Antheraea polyphemus*. *J. Biol. Chem.* **248**, 3012–3020.
20. Fukumori H, Teshiba S, Shigeoka Y, Yamamoto K, Banno Y, Aso Y. 2014 Purification and characterization of cocoonase from the silkworm *Bombyx mori*. *Biosci. Biotechnol. Biochem.* **78**, 202–211. (doi:10.1080/09168451.2014.878215)
21. Davey JW *et al.* 2016 Major improvements to the *Heliconius melpomene* genome assembly used to confirm 10 chromosome fusion events in 6 million years of butterfly evolution. *G3* **6**, 695–708. (doi:10.1534/g3.115.023655)
22. Harpel D, Cullen DA, Ott SR, Jiggins CD, Walters JR. 2015 Pollen feeding proteomics: salivary proteins of the passion flower butterfly, *Heliconius melpomene*. *Insect Biochem. Mol. Biol.* **63**, 7–13. (doi:10.1016/j.ibmb.2015.04.004)
23. Gilbert LE. 1972 Pollen feeding and reproductive biology of *Heliconius* butterflies. *Proc. Natl Acad. Sci. USA* **69**, 1403–1407. (doi:10.1073/pnas.69.6.1403)
24. Cardoso MA. 2001 Patterns of pollen collection and flower visitation by *Heliconius* butterflies in southeastern Mexico. *J. Trop. Ecol.* **17**, 763–768. (doi:10.1017/S0266467401001572)
25. Rodbumrer P, Arthan D, Uyen U, Yuvaniyama J, Svasti J, Wongsaengchantra PY. 2012 Functional expression of a *Bombyx mori* cocoonase: potential application for silk degumming. *Acta Biochim. Biophys. Sin.* **44**, 974–983. (doi:10.1093/abbs/gms090)
26. Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K. 2011 A random effects branch-site model for detecting episodic diversifying selection. *Mol. Biol. Evol.* **28**, 3033–3043. (doi:10.1093/molbev/msr125)
27. Evnin LB, Vasquez JR, Craik CS. 1990 Substrate specificity of trypsin investigated by using a genetic selection. *Proc. Natl Acad. Sci. USA* **87**, 6659–6663. (doi:10.1073/pnas.87.17.6659)
28. Hedstrom L, Perona JJ, Rutter WJ. 1994 Converting trypsin to chymotrypsin: residue 172 is a substrate specificity determinant. *Biochemistry* **33**, 8757–8763. (doi:10.1021/bi00195a017)
29. Cao X *et al.* 2015 Sequence conservation, phylogenetic relationships, and expression profiles of nondigestive serine proteases and serine protease homologs in *Manduca sexta*. *Insect Biochem. Mol. Biol.* **62**, 51–63. (doi:10.1016/j.ibmb.2014.10.006)
30. Hedstrom L, Szilagyi L, Rutter WJ. 1992 Converting trypsin to chymotrypsin: the role of surface loops. *Science* **255**, 1249–1253. (doi:10.1126/science.1546324)
31. Kurth T, Ullmann D, Jakubke HD, Hedstrom L. 1997 Converting trypsin to chymotrypsin: structural determinants of S1′ specificity. *Biochemistry* **36**, 10 098–10 104. (doi:10.1021/bi970937l)
32. Ma W, Tang C, Lai L. 2005 Specificity of trypsin and chymotrypsin: loop-motion-controlled dynamic correlation as a determinant. *Biophys. J.* **89**, 1183–1193. (doi:10.1529/biophysj.104.057158)
33. Perona JJ, Craik CS. 1995 Structural basis of substrate specificity in the serine proteases. *Protein Sci.* **4**, 337–360. (doi:10.1002/pro.5560040301)
34. Zhang X, Perica T, Teichmann SA. 2013 Evolution of protein structures and interactions from the perspective of residue contact networks. *Curr. Opin. Struct. Biol.* **23**, 954–963. (doi:10.1016/j.sbi.2013.07.004)
35. Krem MM, Di Cera E. 2001 Molecular markers of serine protease evolution. *EMBO J.* **20**, 3036–3045. (doi:10.1093/emboj/20.12.3036)
36. Eijsink VG, Gaseidnes S, Borchert TV, van den Burg B. 2005 Directed evolution of enzyme stability. *Biomol. Eng.* **22**, 21–30. (doi:10.1016/j.bioeng.2004.12.003)
37. Ogino H, Uchiho T, Doukyu N, Yasuda M, Ishimi K, Ishikawa H. 2007 Effect of exchange of amino acid residues of the surface region of the PST-01 protease on its organic solvent-stability. *Biochem. Biophys. Res. Commun.* **358**, 1028–1033. (doi:10.1016/j.bbrc.2007.05.047)
38. Silvério A, de Araujo Mariath JE. 2014 Comparative structure of the pollen in species of *Passiflora*: insights from the pollen wall and cytoplasm contents. *Plant Syst. Evol.* **300**, 347–358. (doi:10.1007/s00606-013-0887-6)
39. Estrada C, Jiggins CD. 2002 Patterns of pollen feeding and habitat preference among *Heliconius* species. *Ecol. Entomol.* **27**, 448–456. (doi:10.1046/j.1365-2311.2002.00434.x)
40. Mallet J, Gilbert LE. 1995 Why are there so many mimicry rings? Correlations between habitat, behaviour and mimicry in *Heliconius* butterflies. *Biol. J. Linn. Soc.* **55**, 159–180. (doi:10.1111/j.1095-8312.1995.tb01057.x)
41. Murawski DA. 1986 *Pollination ecology of a Costa Rican population of* Psiguria warscewiczii *in relation to foraging behavior of* Heliconius *butterflies*. Austin, TX: University of Texas at Austin.
42. Salcedo C. 2011 Pollen preference for *Psychotria* sp. is not learned in the passion flower butterfly, *Heliconius erato*. *J. Insect Sci.* **11**, 25. (doi:doi.org/10.1673/031.011.0125)
43. Boggs CL, Smiley JT, Gilbert, LE. 1981 Patterns of pollen exploitation by *Heliconius* butterflies. *Oecologia* **48**, 284–289. (doi:10.1007/BF00347978)
44. Grabherr MG *et al.* 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652. (doi:10.1038/nbt.1883)
45. Tamura K, Dudley J, Nei M, Kumar S. 2007 MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599. (doi:10.1093/molbev/msm092)
46. Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. 2010 DataMonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26**, 2455–2457. (doi:10.1093/bioinformatics/btq429)
47. Briscoe AD *et al.* 2013 Female behaviour drives expression and evolution of gustatory receptors in butterflies. *PLoS Genet.* **9**, e1003620. (doi:10.1371/journal.pgen.1003620)
48. Macias-Muñoz A, McCulloch KJ, Briscoe AD. In press. Copy number variation and expression analysis reveals a non-orthologous *pinta* gene family member involved in butterfly vision. *Genome Biol. Evol*.
49. Li B, Dewey CN. 2011 RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323. (doi:10.1186/1471-2105-12-323)
50. Wickham H. 2009 *Ggplot2: elegant graphics for data analysis*. New York: NY: Springer Science & Business Media.
51. Martin SH *et al.* 2013 Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.* **23**, 1817–1828. (doi:10.1101/gr.159426.113)
52. Li H, Durbin R. 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760. (doi:10.1093/bioinformatics/btp324)
53. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009 The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. (doi:10.1093/bioinformatics/btp352)
54. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009 Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871. (doi:10.1093/bioinformatics/btp394)
55. Sievers F *et al.* 2011 Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539. (doi:10.1038/msb.2011.75)
56. Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011 SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786. (doi:10.1038/nmeth.1701)
57. Butts CT, Bierma JC, Martin RW. 2016 Novel proteases from the genome of the carnivorous plant *Drosera capensis*: structural prediction and comparative analysis. *Proteins: Struct. Funct. Bioinform.* **84**, 1517–1533. (doi:10.1002/prot.25095)
58. Butts CT, Zhang X, Kelly JE, Roskamp KW, Unhelkar MH, Freites JA, Tahir S, Martin RW. 2016 Sequence comparison, molecular

modeling, and network analysis predict structural diversity in cysteine proteases from the Cape sundew, *Drosera capensis*. *Comput. Struct. Biotechnol. J.* **14**, 271–282. (doi:10.1016/j.csbj.2016.05.003)

59. MacKerell AD *et al.* 1998 All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616. (doi:10.1021/jp973084f)

60. Word JM, Lovell SC, Richardson JS, Richardson DC. 1999 Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **285**, 1735–1747. (doi:10.1006/jmbi.1998.2401)

61. Smith G, Kelly JE, Macias-Muñoz A, Butts CT, Martin RW, Briscoe AD. 2017 Data from: Evolutionary and structural analyses uncover a role for solvent interactions in the diversification of cocoonases in butterflies. Dryad Digital Repository (https://doi.org/10.5061/dryad.355qk)

62. Kabsch W, Sander, C. 1983 Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637. (doi:10.1002/bip.360221211)

63. Kyte J, Doolittle RF. 1982 A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132. (doi:10.1016/0022-2836(82)90515-0)

64. R Core Team. 2017 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. See https://www.R-project.org/.