Evolutionary Origins and Molecular Mechanisms of Hostplant Adaptation in Lepidopteran Herbivores

Dissertation
zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)

Vorgelegt dem Rat der Biologisch-Pharmazeutischen Fakultät
der Friedrich-Schiller-Universität Jena

von Diplom-Biologin
Hanna Marieke Heidel-Fischer

geboren am 14. Juli 1977 in Berlin

Contents

1. General introduction 5
2. Chapter I: Evolutionary origins of a novel hostplant detoxification gene in butterflies 14
2.1 Introduction 15
2.2 Material and Methods 17
2.3 Results 22
2.4 Discussion 31
3. Chapter II: Microevolutionary dynamics of a macroevolutionary key innovation in a Lepidopteran herbivore 40
3.1 Introduction 41
3.2 Material and Methods 44
3.3 Results 46
3.4 Discussion 52
4. Chapter III: Gene expression in a generalist butterfly upon feeding on different hostplants 59
4.1 Introduction 60
4.2 Material and Methods 62
4.3 Discussion 70
5. General discussion 76
6. Summary 85
7. Zusammenfassung 87
8. Acknowledgements 90
9. References 91
10. Selbständigkeitserklärung 99
11. Curriculum Vitae 100
12. Publications 101

1. General introduction

The intimate interactions between herbivorous insects and their food plants have resulted in their coevolution wherein phytophagous insects overcome plant defenses followed by plants counter-adapting to herbivorous insect feeding damage. This constant arms race between plants and insect herbivores was first postulated by Ehrlich and Raven in 1964 and is one of the foundations of insect-plant interactions research. Although the ecology of this co-evolutionary arms race is mostly well understood, an understanding of the molecular mechanisms is still lacking. In particular, molecular insight remains scarce on the insect side of this interaction with little known about the mechanisms used by insects to adapt and metabolize plant allelochemicals. Thus, developing a molecular understanding of insect coevolutionary innovations is very important to understanding the evolution of plant-insect interactions and adaptive processes in general.

Molecular mechanisms for adaptive mutations

As most plants have evolved chemical antiherbivory defenses, successful feeding on plants requires an efficient detoxifying mechanism. Adaptive mutations allowing an insect to utilize a new food plant can have different molecular origins, affecting the regulatory regions as well as the coding sequence of genes. Mutations in the *cis*-regulatory regions can either alter the expression level of genes or it can result in expression in different tissues or developmental stages. As a consequence, the organismal phenotype can be dramatically affected. Schlenke and Begun (2004) showed that an insertion of a transposable element in the *cis*-regulatory region of *Cyp6g1* is associated with increased expression in a *Drosophila simulans* population. *Cyp6g1* has been shown to be responsible for DDT resistance in *Drosophila melanogaster* (Daborn et al. 2002). Surveys of *D. simulans* populations show that lineages with the transposable insertion exhibit evidence of strong directional selection suggesting selection for resistance to an insecticide, a natural toxin or an environmental contaminant (Schlenke and Begun 2004). In *Anopheles gambiae* the expression of one quarter of the detoxification genes is developmentally regulated, indicating the importance of *cis*-regulation for the specificity of detoxification genes in this species (Strode et al. 2006).

Point mutations, insertions or deletions within the coding region can result in novel gene function, allowing for rapid adaptation to new environments. In *D. melanogaster* the insertion

of a transposable element within the coding region of a gene resulted in a truncated gene product that nevertheless generated a functioning protein. The truncated protein appears to increase resistance to an organophosphate pesticide and population surveys indicate that this novel gene product has spread across *D. melanogaster* populations (Aminetzach, Macpherson, and Petrov 2005). Li *et al* (Li, Schuler, and Berenbaum 2003) provided evidence of how shifts in host plant utilization in two *Papilio* species were associated with the evolution of the corresponding P450 sequence which facilitated hostplant specialization.

Gene duplications also play a very important role in the evolution of detoxification mechanisms, either by tandem or relocation duplication of a gene fragment, a whole gene, or a whole chromosomal fragment (Force et al. 1999; Lynch 2007). Although genomic resources suggest that duplication events arise at a very high rate of about 0.01 per gene per million years (Lynch and Conery 2000), the most common fate of duplicated genes will be silencing and loss. Only a small amount of duplicates are retained as functional genes. Such paralogous genes could facilitate adaptive evolution in two ways. Either one paralog could acquire a new function (neofunctionalization), or both duplicates could divide the existent function of the gene (subfunctionalization). In *Papilio polyxenes* for example, a duplicated P450 gene underwent subfunctionalization resulting in two paralogous genes, *CYP6B1* and *CYP6B3* which show different efficiencies in metabolizing plant allelochemicals (Wen et al. 2006). Such gene duplication events, followed by neofunctionalization or subfunctionalization, are likely to be the origin of many detoxifying enzymes in insects.

Phase I and phase II detoxifaication

Metabolism and thereby detoxification of lipophilic toxins into more hydrophilic products typically occurs in two phases. In phase I, a primary product is formed, that might in some cases be more toxic than the parent molecule but in other cases might already be ready for excretion. In the phase II, the primary products are metabolized into secondary products that can be directly excreted (Brattsten 1992). This type I and II categorization is primarily applied to the metabolization of drugs in animals and humans and also provides a useful perspective for considering the metabolism of plant allelochemicals in phytophagous insects.

Although to date knowledge is scarce about the mechanisms applied by insect to metabolize plant allelochemicals, in general it is assumed that phase I and phase II detoxifying enzymes play a major role detoxifying plant compounds in insects. Many studies have found the phase

I enzyme cytochrome P450 monooxygenase to be important across lepidopteran and other insects in the detoxification of plant allelochemicals and other toxins (Petersen et al. 2001; Daborn et al. 2002; Li, Berenbaum, and Schuler 2002; Li et al. 2004; Zeng et al. 2007). In the generalist corn earworm, *Helicoverpa zea*, P450's are upregulated upon larval exposure to toxic plant allelochemicals (Li, Berenbaum, and Schuler 2002; Sasabe et al. 2004; Zeng et al. 2007). The black swallowtail (*Papilio polyxenes*) and the parsnip webworm (*Depressaria pastinacella*) are both specialized on plants containing high levels of the plant allelochemical foranocoumarin. In both species specific cytochrome P450 monooxygenases are induced after furanocoumarin ingestion and appear to be their primary detoxification mechamism (Petersen et al. 2001; Li et al. 2004). Gutathione-S-transferase (GST) belongs to the phase II enzymes. GSTs have been shown to be elevated in generalist caterpillars feeding on various hostplants and are known to be involved in many instances of insecticide resistance (Yu 1982; Nylin 1988; Wadleigh and Yu 1988; Huang et al. 1998; Vontas et al. 2002). In sum, P450's and GST's are important and common insect detoxification mechanisms representing both type I and II modes of action.

Plant defense mechanisms

Plants have several different lines of defense against phytophagous insects including constitutive, steady state, induced and activated defenses. Many plants utilize constitutive physical barriers against herbivory, such as wax layers and trichomes. For example, *Arabidopsis thaliana* has trichomes that are effective in reducing herbivory since trichome density is negatively correlated with herbivore damage in natural field populations (Mauricio 1998). 'Steady state' chemical defenses, that are always present in an active form within the plant tissue, such as terpenes, alkaloids, nicotine or phenolic compounds, are widespread across the plant kingdom and are shared among different plant families. Terpenes are not only present in conifers, but also in cotton plants (*Gossypium hirsutum*) and *A. thaliana* and countless other plants. They serve mostly as a defense mechanism against herbivorous insects but also contribute to plant-plant, plant-fungus and plant-insect interactions (Mumm and Hilker 2006; Stipanovic, Puckhaber, and Bell 2006; Herde et al. 2008). These compounds can be found throughout the plant, at multiple life stages. In conifers, terpenoids are abundant in large quantities and located throughout the wood bark, roots and needles of the trees in specialized resin ducts (Mumm and Hilker 2006). In cotton plants, terpenes, such as gossypol, are present in the seeds, leaves, stems and roots of the plants (Stipanovic, Puckhaber, and Bell 2006).

Upon herbivory, defensive compounds can be further induced, and expressed at higher amounts or even in different compositions. The production and accumulation of the toxic alkaloid nicotine is rapidly increased after a herbivore attack in *Nicotiana attenuata* (Steppuhn and Baldwin 2007). Such induced responses are an effective way for plants to reduce costly responses until an actual herbivory attack. A different type of herbivore defense is the activated defense system in plants. Here the chemical compounds are stored inactive in a plant's tissue and only release their harmful properties upon tissue damage during herbivory. This enables a faster response to herbivore attacks than induced defenses, where recognition of attack, transcription and translation will slow down responses. Furthermore, it allows plants to store the defensive compounds inactive in a nontoxic form within the plant's tissue. Harmful effects on the plant tissue can be avoided and will only appear upon tissue damage when the compounds are activated. Two examples of an activated defense system are the cyanogenic glucosides in Fabaceae and the myrosinase-glucosinolate system in the Brassicaceae that will be discussed in detail later.

Questions addressed in this thesis

The explanations above illustrate the complexity of plant-insect interactions. Understanding the evolution of these interactions on a molecular level is of great interest because this provides important insight into general evolutionary mechanisms of adaptation. In this thesis two plant-insect systems are studied on a molecular level. In chapter I the molecular origins of a novel detoxifying enzyme used by an insect are investigated using the Pieridae butterflies that feed solely on glucosinolate containing plants. In chapter II the same insect-plant system is used to investigate the ongoing evolutionary origins and dynamics of a detoxifying gene. In chapter III the polyphagous comma butterfly *Polygonia c-album* is used as a model species to address the molecular mechanisms involved when insects feed on plant species containing different chemical defenses.

The activated defense system of the Brassicacae plant family

In chapter I and chapter II the coevolutionary system of the Brassicaceae plants and the Pieridae butterflies are used as a model system. The Brassicaceae plant family has been thoroughly studied, most notably in the model species *A. thaliana* (Kliebenstein et al. 2002; Halkier and Gershenzon 2006) which therefore offers unique opportunities to mechanistically understand its coevolution with herbivorous insects. The Brassicales evolved about 90 – 85 million years ago (Wikstrom, Savolainen, and Chase 2001), presenting a novel and effective

two-component induced defense system. Within the intact plant tissue the substrate glucosinolate and the enzyme myrosinase are spatially separated. Upon tissue disruption, myrosinases hydrolyze the glucosinolate substrate generating biologically active hydrolysis products (Figure 1). The major outcome of this hydrolysis are isothiocyanates, that have been shown to possess antimicrobial and insecticidal activities (Wittstcok et al. 2003), but also other products such as epithionitriles and thiocyanates are formed. Although more than 120 glucosinolates are known to exist in brassicaceous plants, they all share the same core structure and differ only in their side chain braching (Benderoth et al. 2006). Glucosinolate diversity is created by the methylthioalkylmalate synthase (MAM) genes encoded by the MAM gene cluster. Gene duplication, neofunctionalization and positive selection of the MAM gene cluster are the driving forces generating glucosinolate diversity (Benderoth et al. 2006). In addition a variety of different myrosinases exist among plant species as well as in different plant tissues, which together with various cofactors also influence the biochemical outcome of the hydrolysis reaction (Rask et al. 2000).

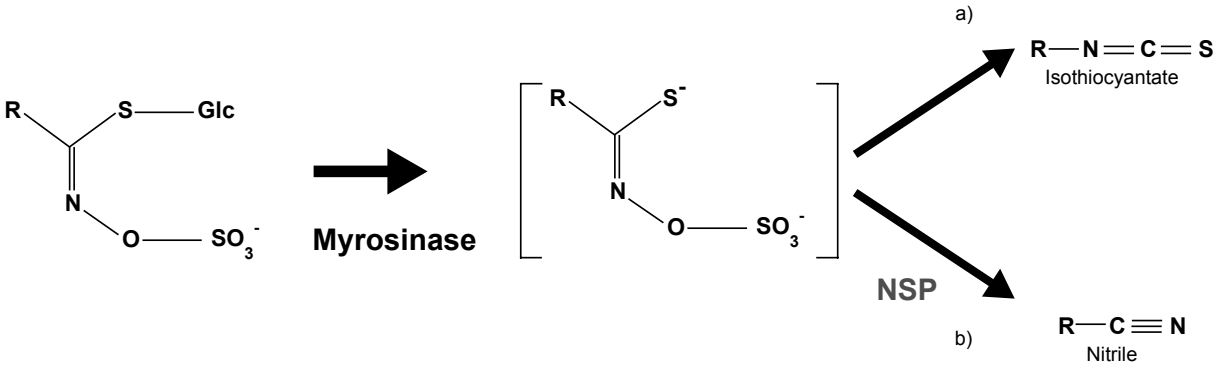


Figure 1: The glucosinolate-myrosinase system in plant defenses of the Brassicaceae. In damaged plant tissue the glucosinolate is hydrolyzed by the myrosinase, via an unstable midproduct the major out come of this hydrolysis is the toxic isothiocyanates (a). In the presences of NSP in the midgut of the Pierini caterpillars the major outcome of this hydrolysis is nitriles (b).

Insect adaptation to the Brassicaceae family with a focus on NSP in the Pieridae butterfly family

In spite of this biochemical complexity, some lepidopteran herbivores have adapted successfully to the brassicaceous plants by circumventing the toxic effects of the glucosinolate myrosinase system. The diamondback moth *Plutella xylostella* inhibits the hydrolysis of the glucosinolates completely by using an enzyme called glucosinolate sulfatase (GSS) that forms nontoxic desulfo-glucosinolates rather than the toxic isothiocyanate.

However, the most widespread mechanism in terms of species numbers within the lepidopterans to disarm the activated defense system of the Brassicaceae has emerged in members of the Pieridae family. The basal part of this lepidopteran family feeds on plants of the Fabaceae family. There has been a major host shift to brassicaceous plants about 80 million years ago, hence shortly after the appearance of the Brassicaceae (Figure 2) (Wheat et al. 2007). The Pieridae owe their ability to feed on brassicaceous plants to the Nitrile-specifier protein (NSP) that is expressed in their midgut and redirects the hydrolysis of the glucosinolates to the less harmful nitriles (Figure 1) (Wittstock et al. 2004). NSP is a novel enzyme that shows no similarities to any of the known detoxifying enzymes used by insects. Different to the GSS, which is only acquired by one member of the small Plutellidae family, NSP activity is widespread in the Pieridae family and the appearance of NSP caused a significantly elevated species number (Wheat et al. 2007). While the mode of action of NSP is unknown, experiments suggest that NSP has a catalytic role that is mostly independent of iron supply. The unstable intermediate of the glucosinolate hydrolysis appears to serve as a direct substrate for the nitrile formation by NSP (Burow et al. 2006a).

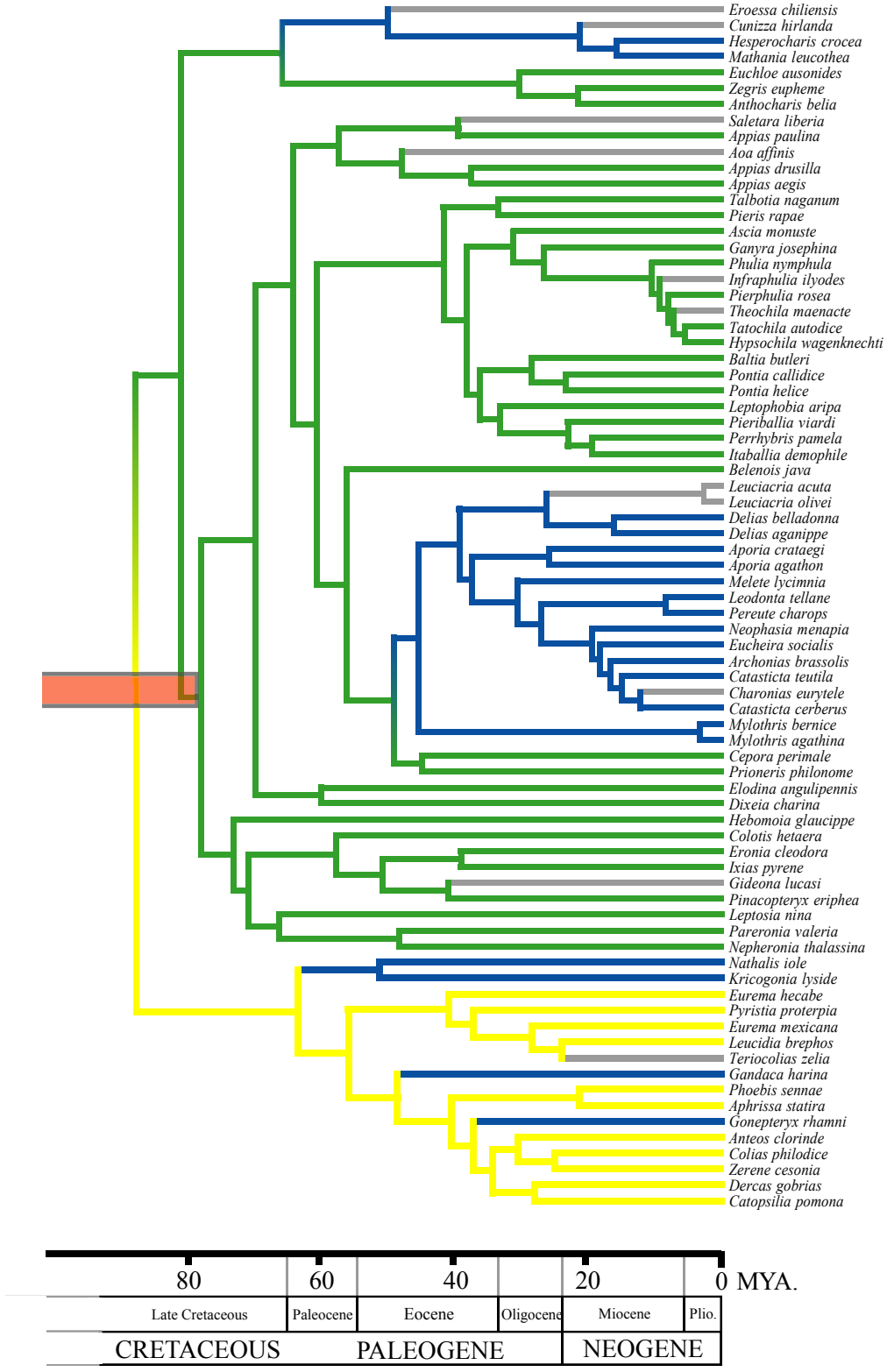


Figure 2: Phylogeny of the Pieridae family. Branches are scaled relative to a divergence time of 85 million years, as shown on the x-axes. Yellow branches refer to Fabales feeding species, green branches to glucosinolate feeding species and blue branches for derived non glucosinolate feeding Pieridae species. Modified from Wheat *et al* 2007.

Due to the evolutionary success of NSP within the Pieridae and the well understood phylogeny of the Pieridae family, NSP provides an ideal system to study the molecular evolution of novel detoxifying enzymes. This will be the focus of chapter I. In addition, the activated defense system of Brassicaceae and its molecular mechanisms are well understood. The Pieridae butterflies, such as for example the Small Cabbage White, *Pieris rapae*, has a wide host plant spectrum within the brassicaceous plants and encounters a variety of different glucosinolate-myrosinase systems during feeding. This makes *P. rapae* an ideal species to investigate the ongoing microevolutionary dynamics of a detoxification gene that needs to adapt to a variable plant defense system. Different evolutionary hypotheses can be tested in population study on this species and this will be the focus of chapter II.

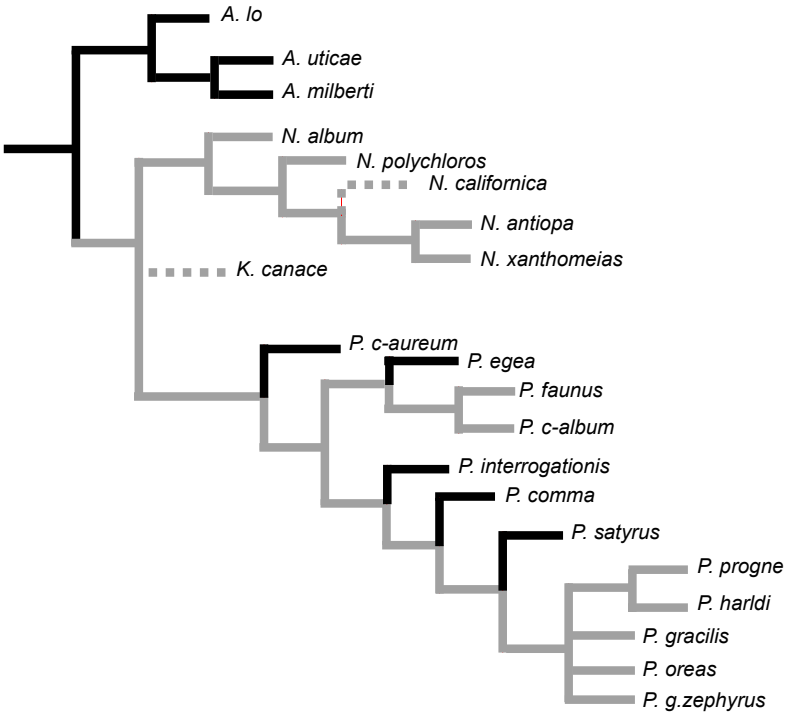


Figure 3: Most parsimonious phylogeny of *Polygonia*, *Nymphalis* and *Aglais* with host plant use. The use of urticalean rosids as food plants is depicted in black branches, a wide host plant range is depicted with grey branches and specialization on plant families outside the urticalean are depicted with dotted branches. Modified from Weingartener *et al* 2005.

The polyphagous butterfly *Polygonia c-album*

Generalist herbivores need different defense strategies than specialized herbivores, as generalist herbivores face an array of different plant defenses and consequently secondary plant compounds and therefore need to invest in broad detoxifying strategies. *P. c-album* is a

member of the Nymphalidae family that underwent several incidences of host plant expansion and constriction along the phylogeny with *P. c-album* being on the far end of polyphagy (Weingartner, Wahlberg, and Nylin 2006) (Figure 3). *P. c-album* has a range including nearly all of Eurasia, and the larvae feed on a diverse spectrum of hostplants from several distantly realted taxa. Although there are some overlaps of secondary plant compounds within the different hostplants, many species contain unique defensive compounds within the host range such as nicotine in *Urtica dioica* (see chapter III). In contrast to the specialist *P. rapae*, that has to adapt to a flexible defense system of one plant family, *P. c-album* is successfully feeding on many different plant families. Although the ecology of *P. c-album* is well understood, the molecular mechanism it applies to feed on such a wide array of food plants is not known and will be the focus of chapter III.

2. Chapter I: Evolutionary origins of a novel hostplant detoxification gene in butterflies

Abstract

Chemical interactions between plants and their insect herbivores provide an excellent opportunity to study the evolution of species interactions on a molecular level. Here we investigate the molecular evolutionary events that gave rise to a novel detoxifying enzyme (nitrile-specifier protein: NSP) in the butterfly family Pieridae, previously identified as a coevolutionary key innovation. By generating and sequencing ESTs, genomic libraries, and screening databases we found NSP to be a member of an insect specific gene family that we characterized and named the NSP-like gene family. Members consist of variable tandem repeats, are gut expressed, and are found across Insecta evolving in a dynamic, ongoing birth-death process. In the Lepidoptera, multiple copies of single domain genes (SDMAs) are present and originate via tandem duplications. Multiple domain genes are found solely within the brassicaceous feeding Pieridae butterflies, one of them being NSP and another called major allergen (MA). Analyses suggest that NSP and its paralog MA have a unique single domain evolutionary origin, being formed by intragenic domain duplication followed by tandem whole-gene duplication. Duplicates subsequently experienced a period of relaxed constraint followed by an increase in constraint, perhaps after neofunctionalization. NSP and its ortholog MA are still experiencing high rates of change, reflecting a dynamic evolution consistent with the known role of NSP in plant insect interactions. Our results provide direct evidence to the hypothesis that gene duplication is one of the driving forces for speciation and adaptation, showing that both within and whole gene tandem duplications are a powerful force underlying evolutionary adaptation.

2.1 Introduction

Understanding the evolutionary origins of genes with well understood adaptive roles is a goal of functional genomics. Attaining such a goal requires both an understanding of gene function and its ecological consequences, along with deep taxonomic sampling of homologous genes. Plant-insect interactions, exemplified by the constant arms race between host plants and their specialist insect herbivores, provide an opportunity to study such adaptive gene evolution on a molecular level. However, developing a functional genomics understanding on both sides of a plant-insect interaction can be difficult. A good example is the glucosinolate–myrosinase system of brassicaceous plants, which is an activated plant chemical defense system. Thorough study in many plant species, the most notable being the genomic-model plant species *Arabidopsis thaliana* and relatives, has identified clear molecular targets of adaptive evolution in the formation of diverse glucosinolate compounds (Koroleva et al. 2000; Kroymann et al. 2001; Kroymann et al. 2003; Windsor et al. 2005; Halkier and Gershenzon 2006; Heidel et al. 2006). Understanding insect detoxiciation of these brassicaceous defensive compounds has also recently advanced, with different mechanisms identified in two non-model lepidopteran species which enabled their caterpillars to feed on brassicaceous plants with impunity (Ratzka et al. 2002; Wittstock et al. 2004). Here we focus on the evolutionary origins of one of these detoxification mechanisms, the nitrile-specifier protein (NSP), present in *Pieris rapae* (the Small White) and its relatives. Our research exemplifies how functional genomic tool development across several non-model species can facilitate evolutionary insights into novel gene evolution.

NSP is structurally different and has no amino acid homology to any known detoxifying enzymes. Only the Pierinae butterflies, members of the Pieridae butterfly family (Lepidoptera), specialized feeders on brassicaceous plants as larvae, possess this protein. NSP disarms the activated glucosinolate–myrosinase defense system of the Brassicales by shifting the hydrolysis of glucosinolates to nitriles instead of the more toxic isothiocyonates (Wittstock et al. 2004; Wheat et al. 2007). Wheat et al. (2007) have argued that NSP was a key innovation in Pieridae butterfly evolution, as it has a single evolutionary origin, appeared shortly after the origin of the Brassicales, and was followed by significantly increased butterfly diversification rates. Here, we pursue this idea by comparing DNA and protein sequences of NSP and its relatives, within the Pierinae and across other insect species (Tabel1), with the goal of developing insight into how novel herbivorous insect detoxification mechanisms evolve. Within this plant-insect interaction, the molecular understanding of plant

chemical defense biosynthesis is well developed, providing a unique opportunity to develop similar insight for the herbivorous molecular counteractions.

As we will show, NSP appears to have arisen by a process of domain and gene duplication, from a sequence of unknown function that is widespread in insect species. The origin of gene diversity is thought to be primarily driven by gene duplication, either in tandem or via large chromosomal duplications (Force et al. 1999; Lynch and Conery 2000). While there is a high death rate of such duplicates, duplicated genes can also undergo subfunctionalization (specializing existing functions) and/or neofunctionalization (giving rise to new functions) (Briscoe 2001; Jordan, Wolf, and Koonin 2004; Spaethe and Briscoe 2004; He and Zhang 2005; Rastogi and Liberles 2005; Benderoth et al. 2006). In addition to whole gene duplications, internal tandem duplications of specific exons or structural domains also occur (Bjorklund, Ekman, and Elofsson 2006). Domain duplication and reorganization may enhance existing functions, promote protein stability or modify functions, for example, by altering substrate specificity (Ponting et al. 2001; Pearson et al. 2004). We aim to understand how domain and gene duplication have formed the NSP gene, facilitating adaptation to new environments and possibly even speciation.

Table 1: List of all insect species from which sequence information was used in this study.

ScientificName (used abbreviation)	Common name	phylogenetic Affiliation Order; Family	Sequence source
Pieris rapae (Pra)	Small Cabbage White	Lepidoptera; Pieridae	cDNA Library
Pieris brassicae (Pbr)	Large cabbage White	Lepidoptera; Pieridae	cDNA Library
Pieris napi (Pna)	Green-veined White	Lepidoptera: Pieridae	PCR-product
Anthocharis cardamines (Aca)	Orange Tip	Lepidoptera; Pieridae	cDNA Library
Pontia daplidice (Pda)	Bath White	Lepidoptera; Pieridae	cDNA Library
Pontia protodice (Ppr)	Checkered White	Lepidoptera; Pieridae	cDNA Library
Dixeia pigea (Dpi)	Antheap White	Lepidoptera; Pieridae	cDNA Library
Eucheira socialis (Eso)	Madrone Caterpillar	Lepidoptera; Pieridae	cDNA Library
Colias eurytheme (Ceu)	Orange Sulphur	Lepidoptera; Pieridae	cDNA Library
Gonepteryx rhamni (Grh)	Brimstone	Lepidoptera; Pieridae	cDNA Library
Bombyx mori (Bmo)	Silkworm	Lepidoptera; Bombycoidae	Butterfly Base, Silk Base
Helicoverpa armigera (Har)	Cotton Bollworm	Lepidoptera; Noctuidae	cDNA Library
Spodoptera frugiperda (Sfr)	Fall Armyworm	Lepidoptera; Noctuidae	cDNA Library
Trichoplusia ni (Tni)	Cabbage Looper	Lepidoptera; Noctuidae	cDNA Library
Plutella xylostella (Pxy)	Diamondback Moth	Lepidoptera; Plutellidae	cDNA Library
Anopheles gambiae (Aga)	Malaria Mosquito	Diptera; Culicidae	NCBI
Aedes aegypti (Aae)	Yellow Fever Mosquito	Diptera; Culicidae	NCBI
Periplaneta americana (Pam)	American Cockroach	Blattodea; Blattidae	NCBI
Blattella germanica (Bam)	German Cockroach	Blattodea; Blattidae	NCBI
Tenebrio molitor (Tmo)	Mealworm	Coleoptera; Tenebrionidae	NCBI
Tribolium castaneum (Tca)	Flour Beetle	Coleoptera; Tenebrionidae	NCBI

Note.- Shown are species name, abbreviation used, common name, phlyogenetic affiliations and sequence source.

2.2 Material and Methods

Specimen Collection

Colias eurytheme, *Pontia daplidicae, Anthocharis cardamines*, *Gonepteryx rhamni, Pieris rapae*, *Pieris brassicae,* and *Pieris napi* butterflies were collected in the field around Jena (Thuringia, Germany) and kept as short term lab cultures. Both adults and larvae which hatched from eggs laid by field-collected butterflies were partly used for tissue samples. Several *Pontia protodice* larvae were field collected in Montana (USA), larvae of *Dixeia pigea* were field collected in South Africa, larvae of *Delias nigrina* were from Southern

Australia and *Eucheira socialis* larvae were collected in Mexico and were dissected and used for extraction of RNA and DNA. *Plutella xylostella*, *Trichoplusia ni,* and *Helicoverpa armigera* larvae and adults were reared in the lab as permanent cultures on artificial diets described elsewhere (Ratzka et al. 2002; Maischak et al. 2007).

RNA Isolation and Reverse Transcription

Larval guts were dissected, immediately submersed in liquid nitrogen and stored at -80 °C. TRIzol Reagent (Invitrogen) was used to isolate the RNA according to the manufacturer's protocol with the following modifications. After adding chloroform to separate the phases, the tubes were stored for 15 minutes at 4 °C before centrifugation. To precipitate the RNA the solution was stored at -20°C overnight. The dried pellet was dissolved in 90 µl RNA storage solution (Ambion), and any remaining genomic DNA contamination was removed by DNAse treatment (TURBO DNAse, Ambion). The DNAse enzyme was removed and the RNA was further purified by using the RNeasy MinElute Clean up Kit (Qiagen) following the manufacturer's protocol and eluted in 20 µl of RNA storage solution (Ambion). To transcribe the mRNA into cDNA the SuperScript III First-Strand Kit (Invitrogen) was used according to the manufacturer's protocol.

Preparation of cDNA libraries

RNA from the gut of the lepidopteran larvae was isolated as described above and poly(A)+ mRNA was isolated using the Poly(A)Purist mRNA Purification Kit according to the manufacturer's protocol (Ambion). For *T. ni*, *P. daplidicae, P. protodice*, *P. rapae, P. brassicae*, *A. cardamines*, *G. rhamni* and *E. socialis* double-stranded, full-length enriched cDNA from dissected larvae were generated by primer extension with the SMART cDNA library construction kit (Clontech) according to the manufacturers protocol but with several modifications. In brief, 2 µg of poly(A)+ mRNA was used for each cDNA library generated. cDNA size fractionation was performed with SizeSep 400 spun columns (GE Healthcare) that resulted in a cutoff at ~200 bp. The full-length-enriched cDNAs were cut with SfiI and ligated to the SfiI-digested pDNR-Lib plasmid vector (Clontech). Ligations were transformed into *E. coli* ELECTROMAX DH5α-E electro-competent cells (Invitrogen). Furthermore, for *C. eurytheme, A. cardamines, D. nigrina, H. armigera* and *P. xylostella* normalized full length-enriched cDNA libraries were generated using a combination of the SMART cDNA library construction kit and the Trimmer Direct cDNA normalization kit (Evrogen) following the manufacturer's protocol.

Generation of EST sequence databases

Plasmid minipreparation from bacterial colonies grown in 96 deep-well plates was performed using the 96 robot plasmid isolation kit (Eppendorf) on a Tecan Evo Freedom 150 robotic platform (Tecan). Single-pass sequencing of the 5' termini of cDNA libraries was carried out on an ABI 3730 xl automatic DNA sequencer (PE Applied Biosystems). Vector clipping, quality trimming and sequence assembly was done with the Lasergene software package (DNAStar Inc.). Blast searches were conducted on a local server using the National Center for Biotechnology Information (NCBI) blastall program. Sequences were aligned using ClustalW software (Thompson et al. 1997).

Cross taxon gene identification

Genes of interest were recovered by different methods. For the species *P. xylostella*, *T. ni*, *H. armigera*, *D. nigrina*, *C. eurytheme*, *P. daplidice, P. protodice*, *P. rapae, P. brassicae*, *A. cardamines*, *G. rhamni* and *E. socialis* cDNA libraries were generated and screened as described below. Degenerate primers were used to amplify genes that were not found in the EST sequences and to identify the desired genes in *P. napi* and *D. pigea* for which no cDNA libraries were created. 5' and 3' RACE methods were used to obtain full length clones when necessary.

Search of existing databases.

Sequences from other insect species were extracted from the NCBI non redundant (nr), and whole genome shotgun sequence (wgs) databases and the ButterflyBase v2.9 (http://heliconius. cap.ed.ac.uk/butterfly/db/).

Genomic DNA Isolation and PCR

For the isolation of genomic DNA from *P. rapae*, *P. brassicae* and *C. eurytheme* the abdomens of adult butterflies kept at 4°C were ground to a fine powder in liquid nitrogen and DNA was isolated using the genomic tip 100/G and genomic DNA buffer kit following the manufacturer's protocol (Qiagen). To amplify the genomic sequences of a gene, specific primers for the genes of interest were used and the desired product amplified by PCR. Positive PCR bands of the correct size were cut out from the agarose gels, column purified (Zymogen), ligated into the pCR II TOPO vector (Invitrogen) and sequenced.

Fosmid library generation and screening

Genomic DNA was isolated from several pupae of *Colias eurytheme* and *Pieris rapae*, using the genomic tip 500/G isolation kit (Qiagen) as described above. Genomic DNA quantity was measured photospectrometrically on a Nanodrop ND1000 and DNA quality and size was checked by pulsed-field gel electrophoresis on a CHEF Mapper XA (Bio-Rad). As the mean size range of both the *Colias* and *Pieris* genomic DNA was ~ 150 kb, the DNA was sheared to the desired 40 kb range with a Hydroshear device (Molecular Devices). For the generation of a fosmid library ~ 3 µg of sheared genomic DNA was used as starting material in a CopyControl fosmid library production kit protocol (Epicentre). This resulted in a library of ≈ 120000 *E. coli* EPI300 clones, each carrying a ~ 40 kb DNA fragment of *Colias* and ≈ 90000 *E. coli* EPI300 clones, each carrying a ~ 40 kb DNA fragment of *Pieris* in the pCC1FOS vector. Appr. 28000 colonies for each species were picked into 384well microtiter plates with a QPix II robotic colony picker (Genetix) and subsequently spotted onto large Performa II nylon membranes (Genetix). Colony picking, replicating, membrane spotting and quality testing was performed by the RZPD (German Resource Center for Genome Research). The library was stored as −80°C glycerol stocks in 384well microtiter plates.

The randomly picked (n = 28000) clones represent a 2-3 fold genome coverage, assuming a genome size G of 450 Mbp and an insert size i of 40 kb. A first quality check of the library for DNA insert size and clone diversity was done by restriction analysis of fosmid DNA isolated from twelve randomly selected library clones for each library. This revealed a total of 24 different restriction patterns and an average insert size of 34 kb.

Overnight cultures of *E. coli* EPI300 clones were diluted 10× in LB containing 12.5 µg ml^{-1} chloramphenicol and 1× induction solution (Epicentre) and incubated for 5 h at 37°C, 300 rpm. Fosmids were isolated with the Nucleobond Xtra Midi Kits according to the manufacturers´ instructions (Macherey-Nagel).

Fosmid library nylon filters were washed, blocked and hybridized with horseradish peroxidase (HRP)-labelled DNA fragments containing the *Colias* SDMA genes. Labelling, hybridization and probe detection were done according to specifications in the ECL DNA labelling and detection kit (GE Healthcare).

Sequence Analysis

Nucleotide sequences were analyzed using the commercial Lasergene Software package and the freeware BioEdit program. All sequences were submitted to Genbank (Accession numbers EU13117-EU13739 identified in Figures). Genes were aligned by their amino acid

sequences using ClustalW (Thompson et al. 1997), sequence identity calculated in the BioEdit 7.0.2 program (Hall 1999), and dN/dS ratios calculated in the DnaSP program (Rozas et al. 2003). If necessary, alignments were then corrected by eye and reverted back to the nucleotide sequences for the phylogenetic analyses. Predictions of the secondary protein structure were generated by the freeware PredictProtein program (http://www.predictprotein.org/) (Rost, Yachdav, and Liu 2004)

Phylogenetic Analysis

The phylogenetic reconstruction implemented four methods. The PAUP 4.0b10 package (Swofford 2003) was used for 1) the construction of a parsimony tree with 500 bootstraps, 2) Neighbor Joining distance analysis, with reconstructed distances estimated using either Jukes-Cantor model or a general time-reversible model with 5000 bootstraps (Felsenstein 2004), 3) maximum likelihood estimations using Model Test 3.7 (Posada and Crandall 1998) to identify the best fitting nucleotide substitution model for each dataset analyzed, with 100 bootstraps performed. Our final method of analysis used Bayesian inference with a GTR + I + G nucleotide substitution model implemented in Mr. Bayes 3.1 (Ronquist and Huelsenbeck 2003). Markov Chain Monte Carlo runs were carried out for 1,000,000 generations after which log likelihood values showed that equilibrium had been reached after the first 400 generations in all cases, and those data were discarded from each run and considered as 'burnin'. At least two runs were conducted per dataset showing agreement in topology and likelihood scores.

Analysis of evolutionary rate differences

Analysis of variation in molecular evolution rate among SDMA (Single Domain MA) clades and the NSP + MA clade was performed using likelihood ratio tests of different branch rate models as implemented in codeml of the PAML software package (version 3.15) (Yang 1997). In this model both synonymous and nonsynonymous changes are allowed and are independent, but are constrained in their ratio, ω. A likelihood value is then calculated for the fit of the DNA data to the ω constraint. For these analyses we used the Bayesian phylogenetic reconstruction of these groups. Likelihood ratio tests were used to compare different hypotheses regarding evolutionary rate (ω) change among clades in the Bayesian phylogenetic tree of the NSP-like gene family. A model assuming one rate of molecular evolution among branches (one ratio model) was compared with models where specific clades as a whole were allowed to have differing rates, resulting in two, three, and four ratio models (Table 2). The

best fitting model allowed four independent rates of molecular change, one each among the SDMA clades of moths (Bombyx + Noctuidae), Coliadinae, Pierinae, and (NSP + MA). Although this model was not significantly different from another model which solely differed by allowing the evolutionary rate to be identical for both Coliadinae and Pierinae SDMAs, the conclusion is the same and highly significant (Table 2).

Table 2: dN/dS (ω) estimates among different evolutionary hypotheses, assessing variation in NSP-like single domain evolutionary rates, with LRT comparisons to best fit model H_6.

	Estimated ω							
Models	$SDMA_M$	$SDMA_C$	$SDMA_P$	NSP/ MA	lnL	LRT vs H_6	d.f	P value
H_0: $SDMA_M = SDMA_C = SDMA_P = NSP/MA$	0.1790	=	=	=	-13048.47	81.256	3	<0.001
H_1: $SDMA_M \neq SDMA_C = SDMA_P = NSP/MA$	0.1512	0.1860	0.1860	0.1860	-13047.31	78.932	3	<0.001
H_2: $SDMA_M = SDMA_C \neq SDMA_P = NSP/MA$	0.1231	0.1231	0.2046	0.2046	-13039.47	63.246	3	<0.001
H_3: $SDMA_M = SDMA_C = SDMA_P \neq NSP/MA$	0.1065	0.1065	0.1065	0.2530	-13014.59	13.494	2	0.001
H_4: $SDMA_M \neq SDMA_C \neq SDMA_P = NSP/MA$	0.1555	0.0652	0.2050	0.2050	-13034.30	52.906	1	<0.001
H_5: $SDMA_M \neq SDMA_C = SDMA_P \neq NSP/MA$	0.1538	0.0803	0.0803	0.2519	-13008.10	0.514	1	0.473
H_6: $SDMA_M \neq SDMA_C \neq SDMA_P \neq NSP/MA$	0.1544	0.0687	0.0849	0.2519	-13007.84	0	n/a	n/a

Note.- Phylogenetic tree and SDMA subscripts M, C, and P stand for Moths, Coliadinae, and Pierinae clades found in Figure 5. Likelihood ratios (lnL), the calculated likelihood ratio test (LRT) vs H_6 lnL, degrees of freedom (d.f.) and significance are reported (P).

2.3 Results

EST databases and Fosmid libraries

For many lepidopteran species, including *Pieridae*, only a very limited number of sequences are available in public databases. To identify NSP-like genes in larval tissues, cDNA libraries were constructed from several Lepidopteran species, generally using larvae of different instars. DNA sequencing from the 5' ends of (directionally cloned) clones followed by clustering resulted in a variable number of ESTs which were further processed. A series of filtering steps were applied to identify and remove reads that did not contain any or very short inserts and each sequence was automatically edited to remove primer sequence and exclude contaminants (from *E. coli*). Sequences were assembled with the Lasergene Software using moderately stringent parameters (*i.e.*, match size was chosen to be at least 25 nucleotides, and match percentage was 94). Using these parameters we obtained the following number of contigs and singletons (represented by single reads) for the different species (total number of

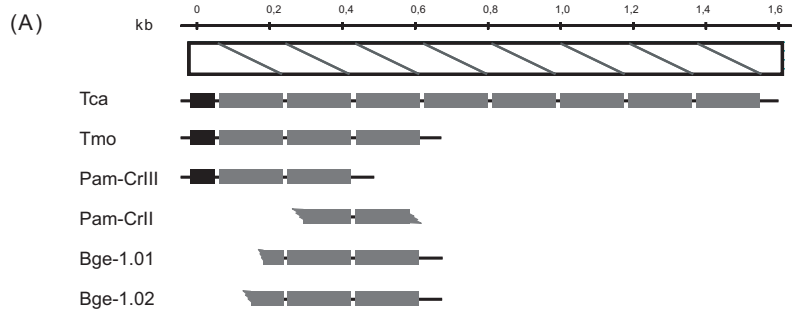
high quality reads/resulting assembled contigs/singletons): *P. xylostella* (15730/2072/5495), *T. ni* (6986/1065/2740), *H. armigera* (35389/3385/8135), *D. nigrina* (4638/840/2286), *C. eurytheme* (5171/944/2145), *P. daplidice* (463/34/158), *P. protodice* (1206/133/320), *P. rapae* (18599/2596/5418), *P. brassicae* (2780/256/668), *A. cardamines* (4634/551/2267), *G. rhamni* (2165/170/548), and *E. socialis* (1946/192/467). For putative functional assignments, the assembled sequences were compared against protein and nucleotide NCBI databases, using a locally installed BLAST search tool. Sequences from these12 cDNA libraries, containing both full-length and partial cDNA sequences, resulted in databases used in subsequent analysis.

Genomic organization insights (exon-intron structures and putative flanking regions) of NSP-like genes were facilitated by in house generated Fosmid libraries from a Pierinae representative (*Pieris rapae*) and an evolutionarily distant Pieridae (Coliadinae, *Colias eurytheme*). These large-insert genomic libraries were probed with genes/gene fragments obtained from screening the EST databases (see above).

Search results and general gene similarity

When blasting the *P. rapae* NSP amino acid against the NCBI nr database, the highest significant hit is to the Cr-PII allergen from *Periplaneta americana*, a major allergen like protein in the american cockroach (23% identity, 42% similarity). Similar to NSP, Cr-PII consists of a signal peptide resulting in excretion into the gut lumen and has repeats of an approximately 200 amino acid domain, which we hereafter call the major allergen (MA) domain. Similar proteins with different numbers of MA domain repeats exist in *Aedes aegypti* (AEG12) and *Blatella germanica* (Bla g1) and have been shown to be induced after food intake (Pomes et al. 1998; Wang, Lee, and Wu 1999; Yang and Bielawski 2000; Chad Gore and Schal 2005; Shao et al. 2005). Searching the publicly available databases and whole genome sequences across Insecta only identified major allergen like proteins, with several non-lepidopteran species having large MA proteins consisting of many repeating domains (Figure 1a). The 8 MA domain repeats of *Tribolium castaneum* are depicted against one of the three repeats of *Tenebrio molitor* in Figure1a showing the similarity of the domains. In contrast, Dipteran species, for example *Anopheles gambiae* (ANG12 (CAA80505)) and most screened lepidopteran species only possess MA genes consisting of a signal peptide and a single MA domain, which we therefore named single domain MA (SDMA). Amino acid alignment of the SDMA from *P. brassicae* and other lepidopterans and the first domains of the NSP and MA from *P. brassicae* and a single domain of the *P. americana* MA illustrates

the similarity of these sequences, implying a shared evolutionary history (Figure 1b). Pierinae butterflies are the Lepidopteran exception, as they additionally possess two genes with a signal peptide and three MA domains each (Figure 2). Comparison between one MA domain and the NSP sequence without signal peptide of *P. rapae* in a dot plot illustrates the three repeat domain structure of these proteins (Figure 2a). By searching our cDNA libraries with these sequences we identified one SDMA gene in *P. xylostella*, *T. ni*, *A. cardamines, E. socialis, D. pigea, P. rapae, P. brassicae, G. rhamni* and *P. napi* and two SDMA genes in *H. armigera* and *C. eurytheme,* and screening public databases we found two SDMA genes in *Bombyx mori* and *Spodoptera frugiperda* (Figure 3a and 3c).



(B)

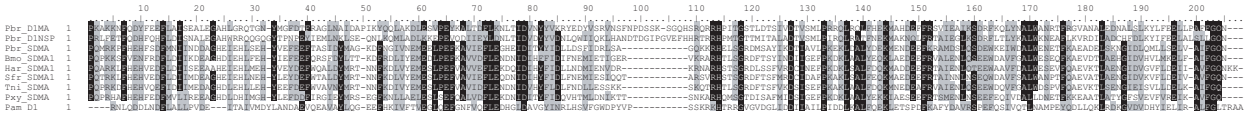


Figure 1: (A) A comparison of amino acid sequence homology and domain repeat structure, shown as a dotplot, between a single domain of *Tenebrio molitor* Insect allergen-like protein (vertical axis) and the *Tribolium castaneum* 8-domain repeat insect allergen-like protein (horizontal axis). Dots represent ≥ 40% identity across a sliding window of 15 amino acids. Below is the comparison of amino acid domain repeats of insect allergen-like sequences in beetles and cockroaches, aligned with the sequences above. Tca = *Tribolium castaneum* (XM_966479) Tmo = *Tenebrio molitor* (AY327800), Pam = *Periplaneta americana* (CrIII=PAU69957; CrII= AAC34736), Bge = *Blattella germanica* (1.01= AF072219, 1.02=AF072220). Pam-CrII, Bge-1.01 and Bge-1.02 are partial cDNA sequences. Black and grey bars respectively represent signal peptide and MA domains. (B) Amino acid alignment of the SDMA from *Pieris brassicae* (Pbra Acc: EU137128), *Bombyx mori* (Bmo), *Helicoverpa armigera* (Har Acc: EU137124), *Spodoptera frugiperda* (Sfr Acc: EU137137), *Trichoplusia ni* (Tni, Acc: EU137139) and *Plutella xylostella* (Pxy, Acc: EU137131), and the first domains of the NSP and MA from *P. brassicae* (NSP Acc: EU137127, MA Acc: EU137126) with a single domain of the *P. americana* MA. 80 % similarity is shaded in grey.

Exon intron structure for the formation of SDMA, MA and NSP

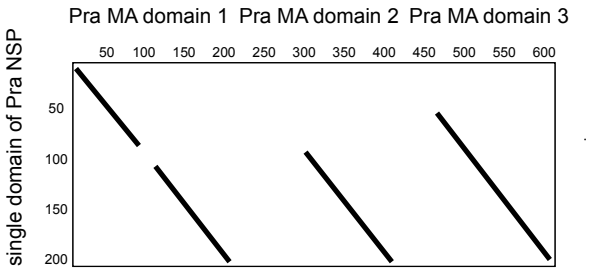
Based on genomic sequence, the exon intron structure of the MA domains appears to be conserved across the NSP, MA and SDMA genes (Figure 2b). Each domain consists of three exons, with the first consisting of about 200 base pairs (bps) (range 192-204 bps), the second containing 69-70 bps and the third being more variable at around 337 bps (range 302-362 bps). The only exception is given by the first domain of the MA gene, which consists of only two exons, with the first and second exons being 192 and 408 bps respectively.

Defining the core domain of the NSP-like gene family

The core domain of the NSP-like gene family was defined by a combination of Dot Plot visualization (Figure 2b), the length and amino acid sequence of the SDMA (Figure 3a), the exon-intron structure of the SDMAs in comparison to the exon-intron structure of NSP and MA (Figure 2b), and the predicted secondary protein structure generated by PredictProtein (Figure 3a). All four analyses suggested a division of NSP and MA into three domains. Strong support comes especially from the predicted secondary protein structure. All SDMAs show the same motif of two times 6 Alpha-Helixes, separated by a turn, exemplified by the the predicted secondary protein structure of *P. xylostella* SDMA (Figure 3a).

We defined domains to end with the amino acids G Q N as does the SDMA protein. However, there are slight inconsistencies within the length of the SDMA protein versus the domains of MA and NSP. While SDMA is about 190 amino acids long, NSP and MA domains are both about 210 amino acids long. This produces a gap of 11-15 amino acids in the domain alignment at position 91 of the SDMA Pieris rapae SDMA (minus signal peptide). The observed amino acids lacking in the SDMA sequence correlate with the beginning of the third exon of the *Pieris rapae* NSP genomic sequence. These additional amino acids in NSP and MA are thus simple terminal exon additions and not changes to the core domain.

(A)



(B)

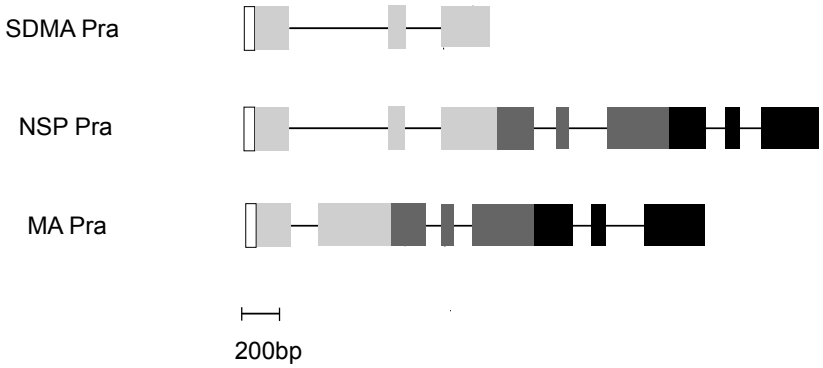


Figure 2: (A) Dotplot depicting a comparison of amino acid homology and domain repeat structure between a single domain of NSP (vertical axis) and the 3 domain MA of *Pieris rapae* (horizontal axis). Dots represent ≥ 56% identity across a sliding window of 30 amino acids. (B) Structural comparison among the SDMA (EU265819), MA (EU265818) and NSP (EU265817) of *P. rapae*. Shaded bars and lines respectively represent gene exons and

Phylogenetic Analysis of the SDMAs within the Lepidoptera

Figure 3 shows an alignment of insecta SDMAs (Figure 3a) and the phylogenetic relationships of the lepidopteran SDMA genes with *Plutella xylostella* as out-group due to its more basal position in the Lepidoptera (Figure 3c). Parsimony, ML, and Bayesian phylogenetic reconstruction methods found nearly identical topologies with only moderate variation in node support, with NJ distance reconstruction supporting these groupings albeit with several polytomies. For some lepidopteran species we could identify multiple SDMA copies in the genome. There is a clear separation of the Noctuidae from the Pieridae SDMAs. SDMA duplicates are of variable ages, if we assume a constant rate of evolution among them. Two very recent duplicates can be seen in both *Bombyx mori* and *Helicoverpa armigera*. *Spodoptera frugiperda* duplicates appear to have arisen from an older duplication event while *C. eurytheme* duplicates have a potentially much deeper origin. Analysis of genomic sequence finds that *B. mori* and *C. eurytheme* SDMAs share exon intron structure and these duplicates in both species are tandem (Figure 3b). These observations suggest conservation of gene structure with tandem duplications being common at least across the

higher Lepidoptera. In addition, the phylogeny of the SDMAs follows the expected species phylogeny of the Pieridae and their relation to the Noctuidae (Braby, Vila, and Pierce 2006).

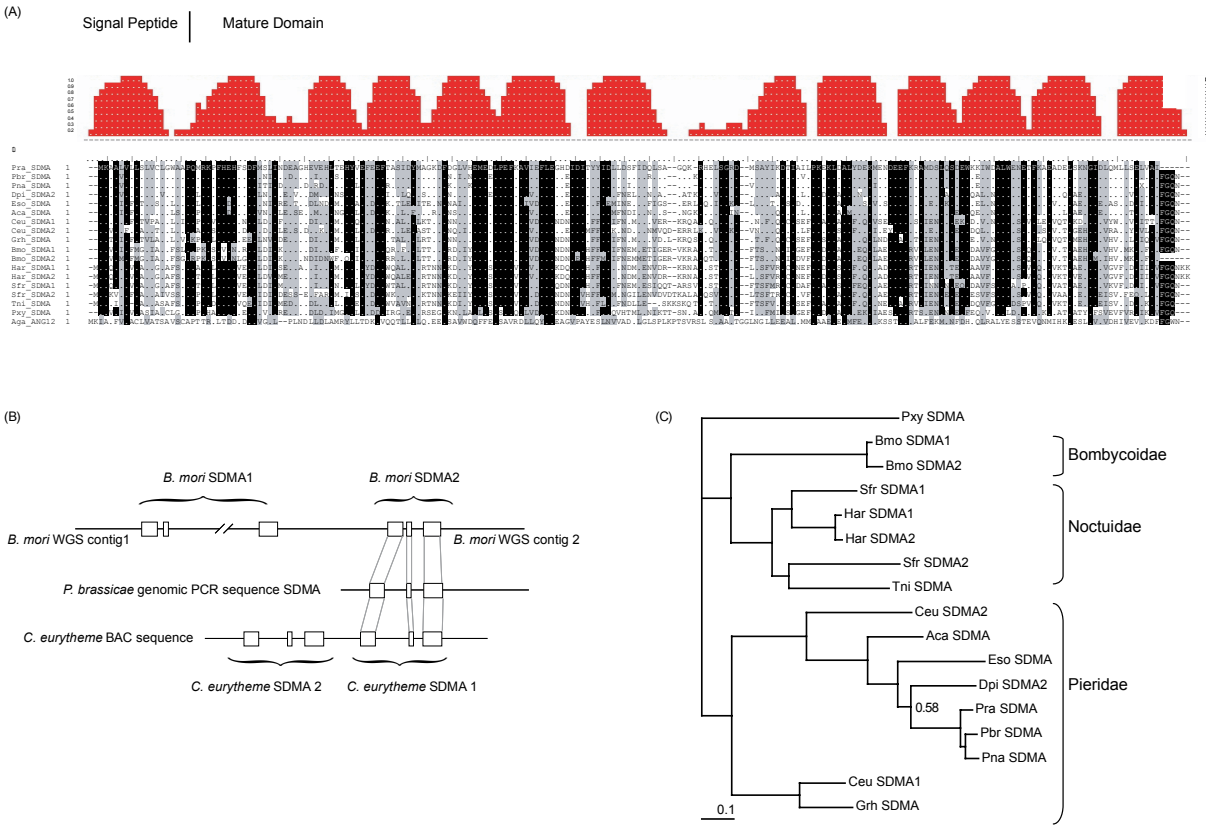


Figure 3: (A) ClustalX alignment of identified SDMA amino acid sequences in Lepidoptera and *Anopheles gambiae*. Identical residues shaded black, with 80 % similarity shaded in grey. Species name abbreviations (with accession number): Pra – *Pieris rapae* (EU137136), Pbra – *Pieris brassicae* (EU137128), Dpi – *Dixeia pigea* (EU137121), Eso – *Eucheira socialis* (EU137122), Aca – *Anthocharis cardamines* (EU137118), Ceu – *Colias eurytheme* (EU137122), Grh – *Gonepteryx rhamni* (EU137123), Bmo – *Bombyx mori*, Har – *Helicoverpa armigera* (SDMA1: EU137124, SDMA2: EU137125), Sfr – *Spodoptera frugiperda* (SDMA1: EU137137, SDMA2: EU137138), Tni- *Trichoplusia ni* (EU137139), Pxy – *Plutella xylostella* (EU137131), Aga – *Anopheles gambiae*. Predicted secondary structure of the Pxy SDMA is depicted above alignment, showing the probability of alpha-helix formation from 0.2 to 1.0 (vertical axis). (B) Genomic organization and intron exon structure of the *B. mori* and *C. eurytheme* SDMA (with locus homology indicated), and the intron exon structure of the *P. brassicae* SDMA. (Accession numbers: *C. eurytheme* genomic SDMA1: EU265820, *C. eurytheme* genomic SDMA2: EU265821, P. brassicae genomic SDMA: EU265822) (C) Bayesian gene phylogeny of all identified lepidopteran SDMAs with *P. xylostella* as an outgroup. The sole nodes with a posterior probability below 0.89 is indicated.

Phylogenetic Analysis of the Pierinae MAs and NSPs

Reconstructed phylogeny of MA genes is robust to analysis methods, as all methods agree on tree topology with good support. Sequence divergence of the MAs is high between Pierinae species *(P. rapae, P. brassicae, Pieris napi*) and *(Pontia daplidice*, *Pontia protodice*, *A. cardamines*) (Supplementary Figure 1a). This strong deviation between *Pieris* and the grouping of *Pontia* + *Anthocharis* is not in accordance with the resolved species phylogeny.

Pieris and *Pontia* species are nearly sister genera and very derived within Pieridae, while *A. cardamines* is more basal (Braby, Vila, and Pierce 2006; Chew and Watt 2006; Wheat et al. 2007). In order to resolve this issue, we calculated and graphed these species pair-wise divergences for MA and two additional genes commonly used in phyogenetic studies, downloaded from Genbank (Supplementary Figure 1b). Both elongation factor 1 alpha (EF1a) and cytochrome oxidase I (COXI) show expected evolutionary distance relationships, with *Pieris* vs. *Pontia* comparisons being much less than either compared to *A. cardamines*. However, MA comparisons between *A. cardamines* and *Pontia* are much lower than either compared to *P. rapae*. These observations indicate that the *P. rapae* MA is not an ortholog to the *A. cardamines* and *Pontia* MA, but rather a different locus originating from an old duplication event.

An unrooted phylogenetic tree (not shown) of the NSP domains from *P. rapae*, *P. brassicae* and *Pontia daplidice* shows that the three domains of the NSP genes cluster together and the distance between the species in each domain is in accordance with their phylogenetic species distances suggesting they are true orthologs. These domain relationships can be seen in the domain phylogeny presented below (Figure 4).

Phylogenetic Analysis of domains: SDMA, MA and NSP

The phylogeny of all the lepidopteran SDMAs and the single domains of all identified NSPs and MAs of the Pieridae family is shown in Figure 4. The SDMAs give the same grouping as in the phylogeny in Figure 3c, showing a clear divergence between Noctuidae and Pieridae species. MA and NSP show a common ancestry to the Pierinae SDMAs, which are in turn sister to one of the two *Colias* SDMAs. The domains of NSP and MA group together forming a well supported monophyletic group. Each of the MA and NSP domains in turn group together, generally according to domain and locus (i.e. all third domains of both MA and NSP are a derived clade, and within this clade, all third domains of NSP group together). Overall tree topology is consistent among methods supporting a shared origin of MA and NSP and their clustered domains, but only Bayesian reconstruction gives good support for the relative grouping among these domain clusters. Our Bayesian method is also the most robust to dramatic changes in rates among codon positions as it is partitioned by codon position. Other methods support only a polytomy for these relative clade relations. Bayesian reconstruction places the first domain as basal and the third domain as derived.

Sequence comparison of MA and NSP domains and SDMAs

The three MA and NSP domains differ considerably and consistently within and between each other (Supplementary Table 1). First, sequence identity values calculated among the domains of MA, beginning with the two orthologous comparisons (*P. napi* MA vs. *P. rapae* MA and *P. daplidice* MA vs. *A. cardamines* MA) followed by the interlocus comparison (*P. rapae MA* vs. *A. cardamines* MA), indicate that domain 1 is more closely related to domain 2 than domain 3. This is also observed in comparisons within *P. rapae* MA and NSP, as well as between these two genes. Domain 3 is more closely related to domain 2 in all these comparisons as well.

Comparative analysis of molecular evolutionary rate

Likelihood ratio tests were used to compare different hypotheses of evolutionary rate change along branches in the Bayesian phylogenetic tree of the NSP-like gene family domains (Figure 4). A model assuming one rate of molecular evolution among all branches (one ratio model) was compared with models where specific clades as a whole were allowed to have differing rates, resulting in two, three, and four ratio models (Table 2). The best fitting model (H_6) allowed a total of four independent rates of molecular change, one each among the SDMA clades of moths (Bombyx + Noctuidae), Coliadinae, Pierinae, and NSP + MA. Although this model was not significantly different from another model which differed only by allowing the evolutionary rate to be identical for both Coliadinae and Pierinae SDMAs (H_5), the conclusion is the same and highly significant (Table 2). First, the rate of molecular change is slower in the Pieridae SDMAs compared to the other Lepidopteran SDMA representatives we have in our dataset, at nearly half the rate of molecular change. Second, the molecular evolution rate of NSP + MA is more than three times that found in Pieridae SDMAs. Thus, the rate of molecular evolution in the SDMA lineage was relatively slow, whereas once NSP and MA appeared, the rate increased significantly in the Pierinae.

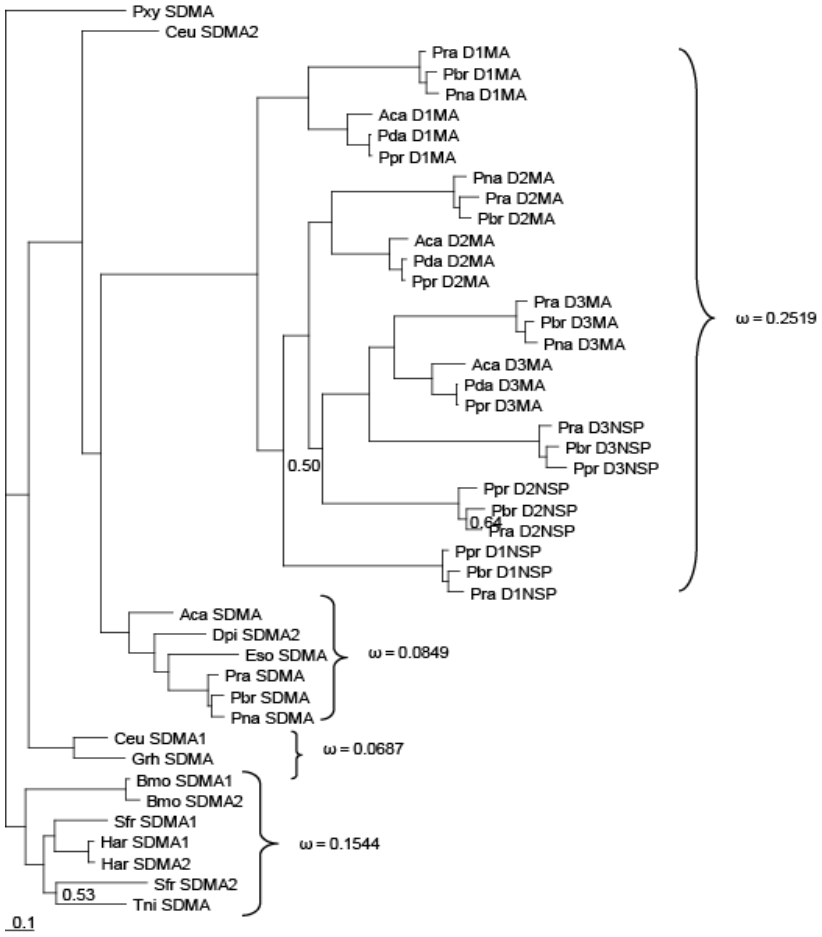


Figure 4: Bayesian gene phylogeny of all lepidopteran SDMAs and the single domains of all identified NSPs and MAs from the Pierinae butterflies. Posterior probability shown for all nodes below 0.8. The evolutionary rate for each clade measured as dN/dS (= **ω**). (Accession numbers: *P. rapae* NSP: AAR84202, *P. rapae* MA: EU137135, *P. brassicae* NSP: EU137127, *P. brassicae* MA: EU137126, *P. protodice* NSP: EU137134, *P. protodice* MA: EU137133, *P. protodice* MA: EU137132, *P. napi* MA: EU137129, *A. cardamines* MA: EU137117)

2.4 Discussion

Here we present a novel insect gene family and focus on the evolution of a member of that family, NSP, that acquired a new, ecologically important function. NSP appears to have evolved via gene duplication and neofunctionalization from a three domain MA only present in Pierinae butterflies. Their common ancestor seems to have evolved via both domain and gene duplication from an ancestral SDMA. While the function of both SDMA and MA expressed in the gut of insect species still remains unknown, SDMA was an exaptation, as defined by Gould and Vrba (1982) for the NSP function, which facilitated a key host plant shift with the macroevolutionary consequence of increased Pierinae butterfly speciation rates (Wittstock et al. 2004; Wheat et al. 2007).

The NSP-like gene family

The naming and current grouping of MA genes needs further development. Initial and ongoing research and expression studies on MAs in cockroaches have focused on the allergic reaction they cause in humans and the proposed name Major Allergen (MA) is solely derived based on these findings. This current naming infers from human health impacts, not reflecting anything related to the intrinsic biological function of MA. (Pomes et al. 1998; Wang, Lee, and Wu 1999; Yang and Bielawski 2000; Chad Gore and Schal 2005; Shao et al. 2005). The InterPro online database of protein families, available through the EBML-EBI website (http://www.ebi.ac.uk/interpro/DisplayIproEntry?ac=IPR010629), contains an accession entry for insect specific allergen repeats (IPR010629). This InterPro accession summarizes several potential major allergen like proteins in insects, listing NSP as well. While many of these computer annotations indicate domain repeats in several genera of Diptera (i.e. *Drosophila, Anopheles, Aedes*), we can find no such tandem repeat structures in dot plot analyses (data not shown). This InterPro accession divides the proteins into domains of 100 amino acids as proposed by Pomes et al (1998). However, our analysis using detailed Dot Plot analysis, predicted protein structure, comparisons of amino acid sequence, and analysis of the exon-intron structure of the genes provides robust support for the division of NSP and MA into three 200 amino acid domains (Figure 1, 2, 3 and see also Results). With this additional comparative genomic insight, we use the 200 amino acid domain designation in our analyses.
Pomes et al. (Pomes et al. 1998) proposed that the repeating units of the allergenic family of proteins from cockroaches were evolutionarily derived by duplication of an ancestral amino acid domain in the "mitochondrial energy transfer proteins", based on the match of some (but

not all) allergen family members to the motif P-x-[DE]-x-[LIVAT]- [RK]-x-[LRH]-[LIVMFY]. However, recent information on these mitochondrial proteins reveals additional differences that make this evolutionary scenario much less likely. First, the characterization of this motif has expanded to a much larger, 78-amino acid region that is described as "pfam00153.13. Mito_carr, Mitochondrial carrier protein" (http://www. ncbi.nlm.nih.gov /Structure/cdd/cddsrv.cgi?uid= pfam00153). This larger region shows no evident similarity to the cockroach allergenic proteins or the single-domain proteins (Figure 3a). Second, the motif best describing the region corresponding to the shorter motif in the single-domain proteins (positions 142- 150 in Fig 3a) is P-K-[DEATS]-K-L-[DAS]-A-L-[YF] which shows some similarities but also several differences from the originally identified motif. Third, a three-dimensional structure has been determined for one of the pfam00153 family members (Protein Data Bank Accession 1okc, (Nury et al. 2006)), and it is rather different from the predicted structure of the NSP-like proteins as generated by PredictProtein (Figure 3a). Therefore, the origin of the repeating units can no longer be confidently ascribed to the mitochondrial solute carrier proteins as originally proposed, and we suggest that this hypothesis of Pomes et al. (1998) should be removed from the descriptions of pfam06757 Ins_allergen_rp and InterPro IPR010629 (http ://www. ncbi.nlm.nih. gov/Structure/cdd/ cddsrv.cgi?uid =pfam06757).

Furthermore, there is a need to both clarify and distinguish these insect gut expressed, multiple domain repeat proteins now that a function in an insect is well established (Wittstock et al. 2004). We therefore propose naming this gene family as the NSP-like gene family with regard to its only functionally characterized protein, while keeping the historically developed names for the proteins of unknown function MA and SDMA until their function is better understood.

Evolutionary origins

The alignment of a single domain of the *P. americana* MA with a single domain each of MA and NSP and the SDMA of *P. brassicae* demonstrates the similarity of the amino acid sequences (Figure 1c). Further support comes from their expression profiles. *P. rapae, P. americana* and *A. aegypti*, each from a different taxa, express their respective forms of MA solely in gut tissue of the primary food consuming life stage, with the latter two being adult stages (Chad Gore and Schal 2005; Shao et al. 2005). These findings strongly suggest that SDMA genes found from across Lepidoptera, as well as MA and NSP of the Pierinae

(Pieridae) butterflies, share a common evolutionary origin with other MA genes found across insecta and are involved in digestive functions.

Proposed model of NSP evolution

The only lepidopterans that appear to possess multiple domain MAs are members of the butterfly subfamily Pierinae, possessing two classes of three domain genes, MA and NSP. Genetic divergence among the different MAs of Pierinae species suggests more than one MA locus, but with only one of the hypothesized paralogs found in to date within any one species (Supplementary Figure 1). In contrast, our data to date suggest NSP to be a single locus gene (Figure 4). However more NSP genes need to be identified as well as their chromosomal regions sequenced to directly answer this question as tandem duplications are likely to occur in this gene family. Our robust phylogenetic reconstruction of the SDMA, MA and NSP show a single shared origin of MA and NSP within the Pierinae. Within the Pieridae, only in *Colias* butterflies have we identified a duplicated SDMA, with one of the genes more closely related to the SDMAs found in Pierinae. While this *Colias* SDMA locus itself appears to be the Pierinae ortholog, this is a tentative assumption (Figure 4). Regardless, using the phylogeny and the gene structure we can postulate the molecular origin of MA and NSP from an SDMA locus most recently shared with the Pierinae, rather than a locus shared with the Coliadinae.

The SDMA common to Lepidoptera seems to have been evolving under a normal birth-death like process with different stages of this process seen in both moths and butterflies (Figure 4). According to our proposed model the SDMA underwent two within gene tandem duplication events, effectively duplicating each time the MA domain, but not the signal peptide. This within gene tandem duplication is similar to what has likely happened in the other insect taxa possessing multiple MA domain repeats (Figure 1). These duplications led to the formation of the common three domain ancestor to MA and NSP. Either a gene duplication event of this ancestor then led to the current MA and NSP genes in the Pierinae, followed by tandem duplication of the MA gene, or this ancestral gene duplicated, and only one of these two MA genes duplicated again to give rise to NSP (Figure 5). One if not both MA loci appear to have lost the second intron along this process (Figure 2). Low, but consistently different interdomain amino acid identities within MA and NSP, within and between species suggest a specific domain duplication scenario. Domain 1 consistently shows a closer similarity to domain 2 over domain 3, while domain 3 is always more similar to domain 2 than 1 (Supplementary Table 1). Thus a possible scenario is that domain 2 originated from domain 1

and is the origin of the domain 3 by a series of within gene tandem duplication events (Figure 5). However, in this scenario, domain 3 must have experienced a considerably higher rate of evolution after emergence from duplication than domain 2, this we cannot explain with our current data set. Nevertheless, the Bayesian phylogenetic analysis of the NSP-like gene Family domains shows within the MA and NSP clade, domain 1 is basal while domain 3 is derived (Figure 4) supporting our proposed duplication scenario, which is also the most parsimonious explanation.

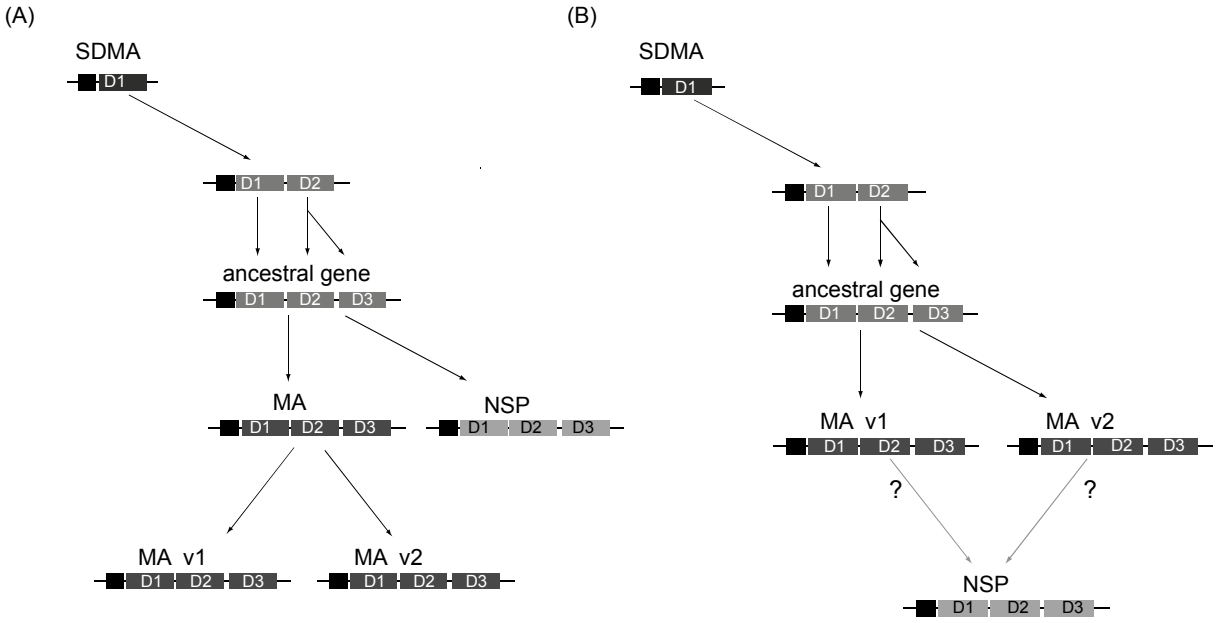


Figure 5: Two possible scenarios for the origin of NSP and MA. In both scenarios the ancestral SDMA underwent a duplication event forming a two-domain gene. This was followed by another duplication of the second domain which formed the ancestral gene to NSP and MA. (A) In the first scenario, the initial duplication allowed one paralog to undergo neofunctionalization, thereby acquiring NSP function. The MA gene, fulfilling the original function, underwent then a second duplication event, likely subfunctionalizing the two gene copies. (B) In a second scenario, the first duplication event produced two MA copies, followed by a duplication of one of the two copies allowing the duplicate to undergo neofunctionalization into NSP (b). Abbreviations stand for domain numbers: D1= Domain1, D2= Domain2, D3= Domain3

Changed rates of dN/dS ratios and subsequent functional divergence of duplicates in the NSP-like gene family

The fate of duplicated genes is a current subject of debate. The traditional neofunctionalization hypothesis suggests that one copy of a duplicated gene preserves the original function leaving the other copy free to accumulate mutations, which subsequently can lead to a new function or the loss of function of one gene copy (Ohno 1970). In contrast, the subfunctionalization hypothesis argues that the ancestral function of the gene is partitioned by the duplicates (He and Zhang 2005). Recent studies provide evidence supporting the second

scenario. For example, in the black swallowtail (*Papilio polyxenes*) subfunctionalization of Cytochrome P450 duplicates led to different efficiencies of metabolising furanocoumarins, the secondary plant compounds present in the caterpillar's food plants (Wen et al. 2006). Computer simulations and genome analysis showed that subfunctionalization might be a first, rapidly occurring step in a duplication event, keeping a second gene copy active and preserved and opening the possibility for a latter lengthy neofunctionalization, suggesting a possible combination of both post-duplication scenarios (He and Zhang 2005; Rastogi and Liberles 2005). Our results provide additional insights into the relative role of sub- and neo-functionalization in duplication dynamics.

Determining the changes in functional role between SDMA, MA, and NSP within the Pierinae lineage is difficult without understanding the biological role of the former two. However, we can gain some insights into how these three groups may differ by assessing relative changes in evolutionary rate. The rate of change shifted substantially between moth SDMA and Pieridae SDMA compared to Pierinae MA + NSP (Table 2, Figure 4), showing a substantial increase in purifying selection followed by a significant relaxation within the MA + NSP clade.

Within the MA and NSP clade we can gain insight into evolutionary processes through comparisons of amino acid identity within and between these gene's domains. Although these domains are very divergent from each other, there clearly was a reappearance of functional constraints after the likely neofunctionalization forming NSP and MA, which is seen in the high sequence identity within each domain across species, loci and genes (Supplementary Table1). We know that NSP and MA have diverged in function as earlier experiments showed that heterologously expressed MA does not show NSP function (H. Vogel, unpublished data). This is also reflected in very low identity value between the NSP and MA domains. Interestingly, the two hypothesized MA paralogous loci are very different, suggesting that these two MA loci may have diverged in functional role as well. We do not know the role of MA in the insect midgut though and experiments are planned in the future to unravel the function of MA giving insight into the neofunctionalization of these gene pairs. The high dN/dS ratio found among MA and NSP genes suggests that one if not both proteins are still experiencing a high rate of evolutionary change. Thus both intragenic and whole gene duplication of the progenitor of MA and NSP facilitated increased evolutionary change compared to the SDMAs, which presumably resulted in neofunctionalization leading to NSP activity.

Mechanisms of duplication and their evolutionary consequences

The mechanism of duplication events is important for answering the question of evolutionary constraints acting on a duplicate. Tandem duplication will make crossing over events between two copies more likely and will retain both genes in the same genomic region, therefore letting them experience the same local recombination or substitution rates, while non tandem duplication may lead to very different evolutionary rates of the duplicates (Zhang and Kishino 2004). Comparative analysis of the NSP-like gene family members in the Lepidoptera indicates a birth-death process involving both within gene and whole gene tandem duplication. The close physical proximity of the recently diverged SDMA duplicates found in *B. mori* and the more divergent ones in *C. eurytheme* suggest an origin via tandem duplication (Figure 3b). It is therefore likely that many lepidopterans and other insects may possess more than one SDMA locus following recent tandem duplication but prior to death of one of the copies.

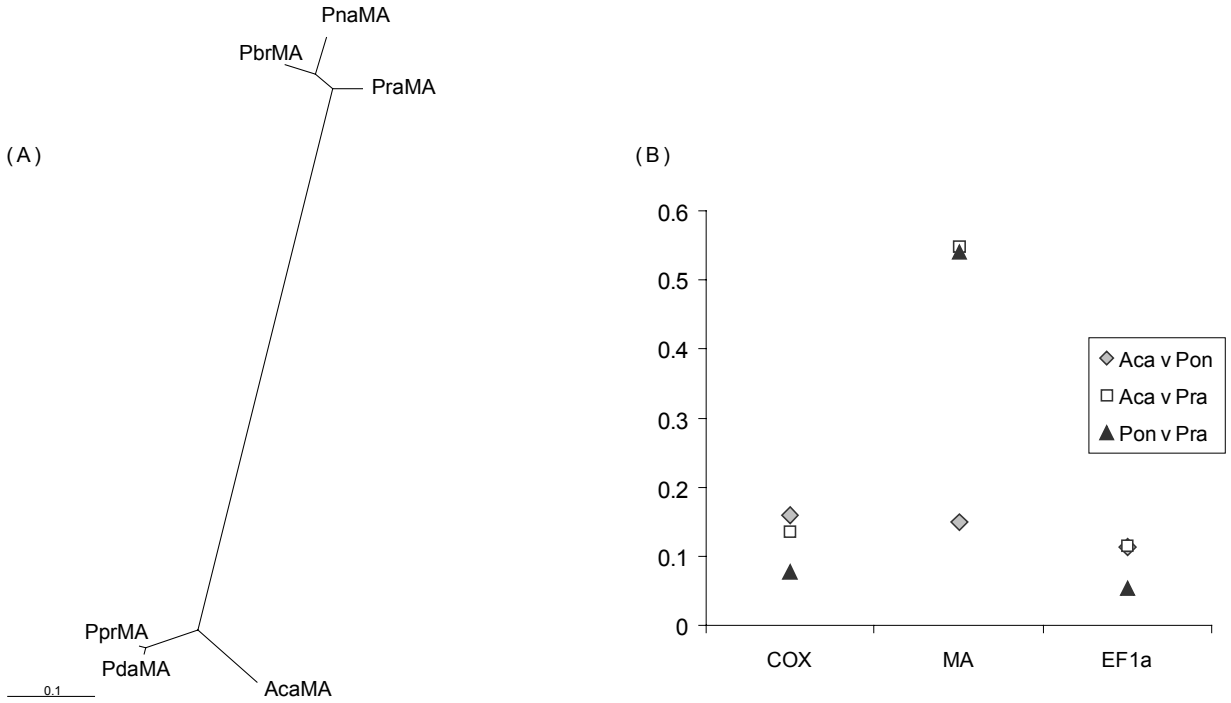
Which molecular mechanism caused the duplication of the ancestral gene of NSP and MA can not be answered with our currently limited dataset. However, the high microsynteny within the lepidopteran genomes allows us to draw some conclusions. Sequencing of the neighbouring genomic area of SDMA, NSP and MA in *P. rapae* and comparison of those regions with the *B. mori* genome suggest that SDMA and MA of *P. rapae* are located on the same chromosome, about 100 kb apart from each other and only separated by one open reading frame. NSP however, appears to be located on a different chromosome. Hence, a relocational duplication event most possibly gave rise to those two paralogs.

Gene evolution insights from non-model species

Using sequenced ESTs from multiple species, degenerate PCR, and database information we were able to find many members of the NSP-like gene family across the higher Lepidoptera. With the exception of *B. mori*, lepidopteran genomic resources are poor. Incomplete sequence data make differentiation of paralogs from orthologs, and by that the understanding of the evolutionary origin of a gene, problematic. Although we can show that MA, SDMA and NSP have a common origin, and we can determine when genes are paralogous and not orthologous via phylogenetic comparisons, we cannot satisfactorily infer orthologous gene copies except when they are very similar in sequence. In sum, although our utilization of the molecular tools available for the weaker molecular half of this insect plant interaction has greatly facilitated our understanding, more resources are needed for further molecular evolutionary insights in ecologically well characterized species, such as Lepidoptera.

Molecular evolution of an ecologically important gene

Gene duplications have been shown to frequently occur in genomes. These studies on duplicated genes, however, have mainly been restricted to genes in model species, whose impact on the evolution of that species were mostly not well understood (Force et al. 1999; Zhang and Kishino 2004; He and Zhang 2005; Nei and Rooney 2005; Kawahara and Nishida 2007). Here, in contrast, we have generated a broad sketch of the evolutionary origins of an adaptive trait, which facilitated an ecologically important host shift and the diversification of the Pieridae butterfly family (Wheat et al. 2007). Thus, we provide direct evidence to the hypothesis that gene duplication is one of the driving forces for speciation and adaptation (Lynch 2002), showing that both within and whole gene tandem duplications are a powerful force underlying evolutionary adaptation.



Supplementary Figure 1: (A) Bayesian gene phylogeny of the known Pierini MA genes. All nodes are supported by a posterior probability of 0.95 or more.(B) Pair wise divergence among *P. rapae*, *P. protodice* and *A. cardamines* across three genes. The x-axis is divided among the three genes (cytochrome oxidase (COX), Elongation Factor 1 alpha (EF1a), major allergen (MA)). Y-axis is pairwise divergence for *A. cardamines* vs. *P. protodice* (gray diamond), *A. cardamines* vs. *P. rapae* (open square), *P. protodice* vs. *P. rapae* (black diamond), MA pairwise divergences involving *P. rapae* are much greater than expected based on control genes (COX and EF1a), suggesting *P. rapae* MA is not orthologous to the other species.

Supplementary Table 1: Amino acid sequence identity values based on clustalW alignments of the single domains of MA and NSP. Shown is a comparison of each MA domain from *Pieris rapae* with *Pieris napi* (a) and a comparison of each MA domain from *Antocharis cardamines* with *Pieris rape*(b) and *Pontia daplidice* (c), an intragene domain comparison of the *P. rapae* MA (d) and NSP (e) domains and an intergene comparison of the *P rapae* MA domains with the NSP domains (f). Boxes comparing identical domain positions are shaded in grey.

a) Recent within MA locus domain comparison

	Pna MA D1	**Pna MA D2**	**Pna MA D3**
Pra MA D1	0.901	0.408	0.257
Pra MA D2	0.375	0.859	0.338
Pra MA D3	0.271	0.347	0.861

b) Divergent within locus domain comparison

	Pda MA D1	**Pda MA D2**	**PdaMA D3**
Aca MA D1	0.81	0.46	0.336
Aca MA D2	0.436	0.886	0.441
Aca MA D3	0.321	0.465	0.699

c) Between MA locus domain comparison

	Pra MA D1	**Pra MA D2**	**Pra MA D3**
Aca MA D1	0.584	0.396	0.302
Aca MA D2	0.424	0.504	0.362
Aca Ma D2	0.315	0.357	0.492

d) Within MA domain comparison

	Pra MA D1	**Pra MA D2**
Pra MA D1	ID	-
Pra MA D2	0.375	ID
Pra MA D3	0.271	0.347

c) Within NSP domain comparison

	Pra NSP D1	**Pra NSP D2**
Pra NSP D1	ID	-
Pra NSP D2	0.401	ID
Pra NSP D3	0.292	0.322

e) Between MA and NSP by domain comparison

	Pra NSP D1	**Pra NSP D2**	**Pra NSP D3**
Pra MA D1	0.446	0.352	0.315
Pra MA D2	0.352	0.399	0.318
Pra MA D3	0.292	0.399	0.378

3. Chapter II: Microevolutionary dynamics of a macroevolutionary key innovation in a Lepidopteran herbivore

Abstract

Understanding the microevolutionary dynamics of genes influencing plant-herbivore interactions can help elucidate their role in the coevolutionary process. Previous work documents the macroevolutionary importance of the nitrile-specifier protein (NSP) in hostplant detoxification which facilitated the hostshift of Pierid butterflies onto Brassicacae hostplants ~80 Myr ago. Here we assess the microevolutionary dynamics of the NSP gene, by studying the within and among-population variation at NSP and reference genes in the butterfly *Pieris rapae* (Little Cabbage White). NSP exhibits unexpectedly high amounts of amino acid polymorphism, unequally distributed across the gene, with little to no genetic differentiation among four populations on two continents. A comparison of synonymous (dS) and nonsynonymous (dN) substitutions in 70 randomly chosen genes among *P. rapae* and its close relative *Pieris brassicae* (Large Cabbage White) finds NSP to be evolving much faster than the genomic average. In addition, multiple NSP haplotypes in *P. rapae* have a dN/dS ratio in excess of 1 even though some portions of the gene exhibit strong purifying selection. While these microevolutionary insights are consistent with diversifying selection at the NSP gene, functional studies are necessary to infer the action of selection upon NSP variation within populations.

3.1 Introduction

Studying plant insect interactions provides an opportunity to investigate the coevolution of species on a molecular, ecological, and evolutionary level. While ecologists are interested in the overall dynamics and interactions between plants and their insect herbivores, biochemical and molecular level studies focus on the genes and gene products that actually interact between these species groups (Berenbaum 2002). Evolutionary understanding is aided by combining both approaches, investigating the origins of genes and understanding their fitness level impacts. In this light, developing a molecular population genetics understanding of the likely selective dynamics acting on candidate genes can greatly facilitate ecological studies, providing markers for genetic variants whose ecological performance can be characterized in the field. Here we present the results of our population genetic study of a novel butterfly detoxification gene, extending previous biochemical, molecular, and macroevolutionary insights to the microevolutionary dynamics between Pieridae butterflies and their host plants, the Brassicaceae.

Brassicaceous plants present a formidable anti-herbivore defense system, where the enzyme myrosinase upon tissue damage catalyzes the hydrolysis of its glucosinolate substrates to toxic end products (Rask et al. 2000; Wittstock and Halkier 2002; Halkier and Gershenzon 2006). Studies of this plant family, most notably on the model species *Arabidopsis thaliana* and relatives, have identified a complex array of molecules involved in this activated chemical defense system (Kliebenstein et al. 2002; Halkier and Gershenzon 2006). A diversity of myrosinases exist in some brassicaceous plants (Rask et al. 2000), which can be accompanied by a variety of cofactors and coenzymes, resulting in the hydrolysis of glucosinolates to variable end products which can influence feeding behavior (Barth and Jander 2006; Burow et al. 2007a; Burow et al. 2008). Additionally, myrosinase concentration in a given plant tissue has been shown to affect herbivore feeding (Kliebenstein et al. 2002; Barth and Jander 2006). Glucosinolate diversity is also an important factor driving adaptive evolution. Methylthioalkylmalate synthases (MAM), encoded by the MAM gene cluster, control an early step in the synthesis of glucosinolates and are responsible for a major part of the glucosinolate diversity within a given plant tissue (Kroymann et al. 2001; Kroymann et al. 2003; Kliebenstein, Kroymann, and Mitchell-Olds 2005; Benderoth et al. 2006; Heidel et al. 2006). Within the MAM gene family, gene duplication, neofunctionalization and positive selection drive biochemical adaptation (Benderoth et al. 2006).

While our understanding of the plant side of this plant insect interaction is well developed, we lack a similar depth of insight on the insect side. We previously identified the enzyme enabling the Pieridae butterfly larvae to circumvent the activated defense of brassicaceous plants (Wittstock et al. 2004). This enzyme, designated a nitrile-specifier protein (NSP), is expressed in the midgut of the caterpillar and promotes the formation of nitriles rather than toxic isothiocyanates upon the myrosinase breakdown of glucosinolates. NSP is a unique detoxifying gene that shows no homology to any known detoxifying enzyme (Fischer et al. 2008).

Macroevolutionary studies allowed us to demonstrate that NSP is a key biochemical innovation in the Pieridae family, evolving shortly after the appearance of the brassicaceous plants and associated with an significantly increased speciation rate (Wheat et al. 2007). More recently we showed that NSP belongs to an insect specific gene family designated the NSP-like gene family and that domain and gene duplication are the driving forces enabling the molecular evolution of NSP (Fischer et al. 2008). NSP has a distinct three-domain structure and is only found in the brassicaceous-feeding Pieridae species (Figure 1). Thus, although we have made some advances in developing an understanding for the insect side of this system, much information is still lacking. While previous macroevolutionary study indicates that the evolution of NSP was likely a key event in the host shift of pierid ancestors from Fabaceae to Brassicaceae, we know nothing about the population level dynamics of NSP with respect to the highly variable and complex activated plant defense system of the Brassicaceae. Here, we begin to address these microevolutionary questions by conducting a population genetic study of *P. rapae* butterflies, the species in which we originally identified the NSP gene.

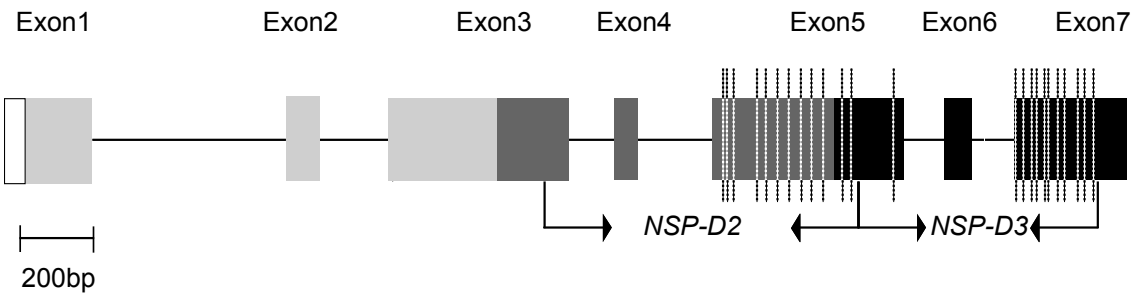


Figure 1: Structural overview of NSP (EU265817) from *P. rapae*. Shaded bars and lines respectively represent gene exons and introns to scale. The signal peptide region is indicated by a blank box while the three domains are shaded to different degrees. Depicted are also the approximate annealing sites of the primer pairs used to amplify ~1 kb large segments of the gene and the corresponding names of the fragments. The two segments studied were NSP-D2 (in domain2) and NSP-D3 (in domain3). Dashes show approximate sites of amino acid substitutions located in the amplified coding regions as listed in Figure 2.

Pieris rapae (small cabbage white) is a highly abundant pierid butterfly species. It is native to Europe and has up to four generations per year in temperate zones. A high dispersal ability coupled with feeding on common agricultural plants (e.g. alfalfa and cabbage) has enabled it to spread rapidly and successfully colonize Australia, New Zealand and North America within the last 120 years (Ohsaki 1979; Jones et al. 1980; Ohsaki 1980). *P. rapae* caterpillars have over 17 reported host plants within the Brassicaceae and thus encounter a high diversity of glucosinolate-myrosinase systems which vary in all the previously discussed components.

Several hypotheses emerge when considering the possible microevolutionary dynamics and patterns of diversity at the NSP gene. For comparative purposes, NSP and a set of reference genes (likely to be experiencing normal purifying selection) were sequenced from the same individuals: two nuclear coding enzymes, as well as a mitochondrial gene, from ten individuals from each of four populations (Italy, France, Germany, and U.S.A.). Additionally, the divergence among 70 random genes between *P. rapae* and *Pieris brassicae* was compared with the divergence at NSP. These datasets allow us assess the relative support for alternative hypotheses of selection at the NSP locus, with a null hypothesis of no adaptive selection and patterns of genetic variation solely reflecting demographic effects (H0). Hypothesis one (H1) expects NSP to be involved in local host plant adaptation, showing unique alleles in each population with greater variation among than within populations. Hypothesis two (H2) proposes a high diversity of NSP across all populations due to *P. rapae* being a highly dispersive generalist, encountering a diverse spectrum of host plants across its range. This hypothesis thus predicts a greater diversity within populations than among them. An additional hypothesis (H3) assumes low diversity in NSP both within and across

populations, due to strong purifying selection on the NSP locus coupled with selective sweeps since diverging from a recent ancestor.

3.2 Material and Methods

Sampling

Ten *P. rapae* adults were collected in the wild at each of three different locations in Europe in the summer of 2002. In Germany (DE) samples were taken 1 km north of Jena, in France (FR) from 50 km northeast from Lyon, and in Italy (IT) from 15 km south of Modena. An additional ten *P. rapae* adults were collected in Ithaca, New York, USA (US) in the summer of 2007. Thus, a total of 40 butterflies were kept at -20°C until their DNA was isolated.

DNA Extraction and PCR

Abdomens of the adult butterflies were homogenized with a TissueLyser (Eppendorf) in the buffer system provided by the genomic DNA extraction kit (Qiagen), and the genomic DNA isolated using genomic tip 20/G columns and the genomic DNA extraction Kit following the manufacturer's protocol (Qiagen). The Eppendorf Master Mix (Eppendorf) was used for the amplification of the desired gene. The PCR products were extracted using a DNA purification kit following the manufacturer's protocol (Zymogen) and cloned into the pCR II TOPO vector (Invitrogen). Eight clones were picked per individual per gene and sequenced.

Amplified genes

Two segments of the NSP gene located directly downstream of each other were amplified from genomic DNA, here referred to as *NSP-D2* and *NSP-D3* (Figure 1). The three reference gene regions studied did not contain introns: isocitrate dehydrogenase (IDH), Glyceraldehyde dehydrogenase (Ga3pdh) and Cytochrome oxidase I (COI) .

Primer sequences were: Pra*NSP-D2*for: tcggctagtcctgctttcaa, Pra*NSP-D2*rev: tgtgttgtcaagggtgtcca, Pra*NSP-D3*for: tggacacccttgacaacaca, Pra*NSP-D3*rev: gtaaagggcaggcacgaagg, Pra*Ga3pdh*for: aaaagggagccaaggttgtt, Pra*Ga3pdh*rev: acgccacaattttcctgaag, Pra*IDH*for: tgctaccatcacaccagatga, Pra*IDH*rev: accaaattcctgcaccttca

Sequencing

Plasmid minipreparation from bacterial colonies grown in 96 deep-well plates was performed using the 96 robot plasmid isolation kit (Eppendorf) on a Tecan Evo Freedom 150 robotic platform (Tecan). Single-pass sequencing of the 5' termini of cDNA libraries was carried out on an ABI 3730 xl automatic DNA sequencer (PE Applied Biosystems).

Data analysis

Vector clipping, quality trimming and sequence assembly was done using the Lasergene software package (DNAStar Inc.). The resulting contig assemblies were aligned using the Clustal W (Thompson et al. 1997) program as implemented in the freeware BioEdit program and corrected by eye. Standard measures of DNA polymorphism, demographic analysis and selection, as well as the G-test, were calculated using DnaSP version 4.50.2 (Rozas et al. 2003) including nucleotide diversity (π) (Nei 1987), nonsysnonymous and silent site substitutions ns/nn (Nei 1987) within *P. rapae* as well as across species (ω) (Watterson 1975), number of segregating sites (S), theta per site from S (θ defined as $4Ne\mu$) (Watterson 1975), recombination rate using the 4 gamete test (Rm) (Hudson, Kreitman, and Aguade 1987), Tajima's D (Tajima 1989), the McDonald Kreitman Test (McDonald and Kreitman 1991) as well as Fay and Wu's H (Fay and Wu 2000) and Fu and Li's D with and without outgroup (Fu and Li 1993). For outgroup analysis *P. brassicae* sequence information was used. P values were determined using coalescent simulations (10,000 runs) of a standard neutral model as implemented in DnaSP. Finally, multilocus tests of selection used the maximum-likelihood-ratio Hudson-Kreitman-Aguadé test (ML-HKA-test) (Wright and Charlesworth 2004). Simulations found that 100,000 chains were sufficient for convergence and the starting value of divergence time for the Marcov chain (T) was obtained using a standard HKA test for the control genes, implemented in DnaSP.

For the following calculations the Arlequin Software package was used (Excoffier, Laval, and Schneider 2005). Population genetic structure in *P. rapae* populations were examined using an analysis of molecular variance (AMOVA) (Excoffier, Smouse, and Quattro 1992; Michalakis and Excoffier 1996). A population pairwise Fst was estimated by the AMOVA, which can be used as a measure of genetic distanced between populations, by linearizing the distance with population divergence time (Slatkin 1995). The significance of the estimated Fst was determined via Markov chain analysis (Raymond and Rousset 1995) using 10,000 permutations. Sequential Bonferroni adjustment was applied to control for Type I errors (Rice

1989). For AMOVA analysis samples were classified in two groups (USA vs Europe). For Fst estimation samples were classified as a single group. Migration rate (m) (Slatkin 1991), and from m the absolute number of migrants exchanged between two populations (M), were computed. An exact test for population differentiation was also computed. The exact test of population differentiation is equivalent to the Fisher's exact test, which tests the null hypothesis of identical allelic distribution across all populations. Significance was determined via Markov chain analysis with 400,000 steps and 100,000 dememorization steps, again applying Bonferroni adjustment when screening for significant values.

***P. rapae* vs. *P. brassicae* EST comparison**

Random sequencing of cDNA libraries made from *P. rapae* and *P. brassicae* gut tissue and the NSP sequence of *P. brassicae* have been described elsewhere (Fischer et al. 2008). 2593 number of unique EST contigs were identified for *P. rapae* from 8153 sequencing reads, while only 973 were found among 2560 reads of *P. brassicae*. The reciprocal best blast hits between each of these two cDNA libraries to the predicted genes of *Bombyx mori* was used to identify homologous genes in both *Pieris* EST collections. Identified sequences were aligned by Clustal X (Thompson et al. 1997) and each visually inspected for regions of high quality sequence and alignment. End regions of alignments were trimmed such that reading frame (i.e. amino acid sequence) was identical for 3 consecutive codons. Degenerate base pair calls were included. Maximum likelihood estimates of the number of pairwise nonsynonymous (dN) and synonymous (dS) substitutions were performed using codeml of the PAML software package (Yang 1997), with the estimates of codon frequencies set as free parameters (option F3x4). The ratio of dN/dS, ω, is indicative of strong purifying selection when $\omega << 1$ and is suggestive of diversifying selection when $\omega > 1$.

3.3 Results

Molecular variation

We examined variation in two segments of the NSP gene (*NSP-D2* and *NSP-D3*) in comparison to exons of reference genes isocitrate dehydrogenase (*Idh*) and glyceraldehyde dehydrogenase (*Ga3pdh*) and a portion of the mitochondrially-encoded Cytochrome oxidase I (*COI*) gene. All genes in all populations harbored genetic variation, with the *NSP* gene segments generally being the most diverse. Nucleotide diversity (π) was roughly 2 to 3 times

higher in *NSP-D2* compared to the control genes, while *NSP- D3* π was nearly double the control genes (Table 1). θ_W showed similar patterns of diversity as π. Levels of synonymous diversity (π_{ss}) are roughly similar across all the nuclear genes, except for *NSP-D2* which has about 50% higher diversity. Nonsynonymous diversity (π_{ns}) is highest in *NSP-D2*, followed by *NSP-D3*, followed by the control genes which have much lower levels of amino acid variation (Table 1). *NSP-D2* and *NSP-D3* have a π_{ns}/π_{ss} that is over twice that of *Idh* and more than 20 times that of *Gapdh* (Table1).

Table 1: Summary statistics for all sequenced genes in *Pieris rapae* for each population separately and across all populations.

		coding								whole gene			non coding		
		n	bp	Jall	Θall	S	Jss	Jns	ns/ss	bp	Л	Θ	bp	Jall	Θall
NSP-D2	DE	20		0.0108	0.00902	19	0.02182	0.00769	0.347378		0.02013	0.01823		0.0405	0.03938
	FR	20		0.0093	0.00795	17	0.02121	0.00593	0.275416		0.0164	0.01326		0.02925	0.02542
	IT	20		0.01042	0.00723	15	0.02307	0.00684	0.292393		0.01689	0.01209		0.02858	0.02087
	US	20		0.01145	0.00854	18	0.02783	0.00683	0.240799		0.01783	0.01319		0.02951	0.02342
	total	80	594	0.01093	0.01054	31	0.02473	0.00703	0.27995	943	0.01722	0.01797	349	0.03159	0.03494
NSP-D3	DE	20		0.00739	0.00732	12	0.01476	0.00531	0.356045		0.01337	0.01335		0.02735	0.02759
	FR	20		0.0062	0.00671	11	0.01262	0.00437	0.342969		0.01241	0.01335		0.02665	0.02759
	IT	20		0.00562	0.00549	9	0.01159	0.00392	0.335038		0.01247	0.01092		0.02809	0.02519
	US	20		0.00538	0.00549	9	0.0153	0.00256	0.16538		0.01114	0.0105		0.02444	0.02269
	total	80	462	0.00642	0.00918	21	0.01419	0.00422	0.294649	717	0.01255	0.01537	327	0.02681	0.03279
IDH	DE	20		0.00459	0.00613	9	0.01723	0.00084	0.048752						
	FR	18		0.00422	0.00492	7	0.01062	0.00233	0.219397						
	IT	18		0.00155	0.00281	4	0.00562	0.00035	0.062278						
	US	15		0.00394	0.00524	7	0.01144	0.00174	0.152098						
	total	71	291	0.00365	0.00755	15	0.01163	0.00131	0.11264						
Ga3pd	DE	18		0.0042	0.00496	6	0.01564	0.00042	0.026854						
	FR	20		0.00302	0.0016	2	0.01217	0	0						
	IT	18		0.00743	0.00743	9	0.01781	0.00042	0.023582						
	US	18		0.00418	0.00248	3	0.01684	0	0						
	total	74	352	0.00444	0.00583	10	0.0173	0.00021	0.012139						
COX	DE	8		0.0079	0.00715	14	0.03026	0.00075	0.024785						
	FR	8		0.01003	0.01022	20	0.03865	0.00088	0.022768						
	IT	8		0.00629	0.00715	14	0.02595	0	0						
	US	7		0.00164	0.00162	3	0.03026	0.00075	0.024785						
	total	31	755	0.00691	0.00963	28	0.02611	0.00076	0.029108						

Note.- Shown are the number of sequences (n), the number of base pairs (bp), the average pairwise differences (π), the pairwise differences for synonymous and nonsynonymous sites (πss and πns), θw, the number of segregating sites (S) and the rate of nonsynonymous to synonymous substitutions for the coding part of the genes. If introns are included in the sequence, the number of base pairs, the average pairwise differences and θw is given separately for the whole gene and the non coding part

The location of amino acid substitutions varied across the sequenced domains of NSP. Each domain is composed of three exons, with codon lengths of about 66, 23 and 110 respectively (Figure 1) (Fischer et al, 2008). The 12 amino acid polymorphisms in NSP domain 2 are only found in its terminal exon (exon five), while 11 of the 14 amino acid polymorphisms in NSP domain 3 are also found in its terminal exon; the other three are in the first exon of the domain 3 (Figures 1,2). The distribution of nonsynonymous polymorphisms across these domains

significantly departs from a random distribution based on the size of the exons, with a paucity of amino acid polymorphisms observed in the first and second exons, and an excess in the terminal exons, of both domains (G value = 5.99, P = 0.014; Supplemental Material). The distribution of synonymous polymorphisms does not show this trend (G value = 0.43, P = 0.512).

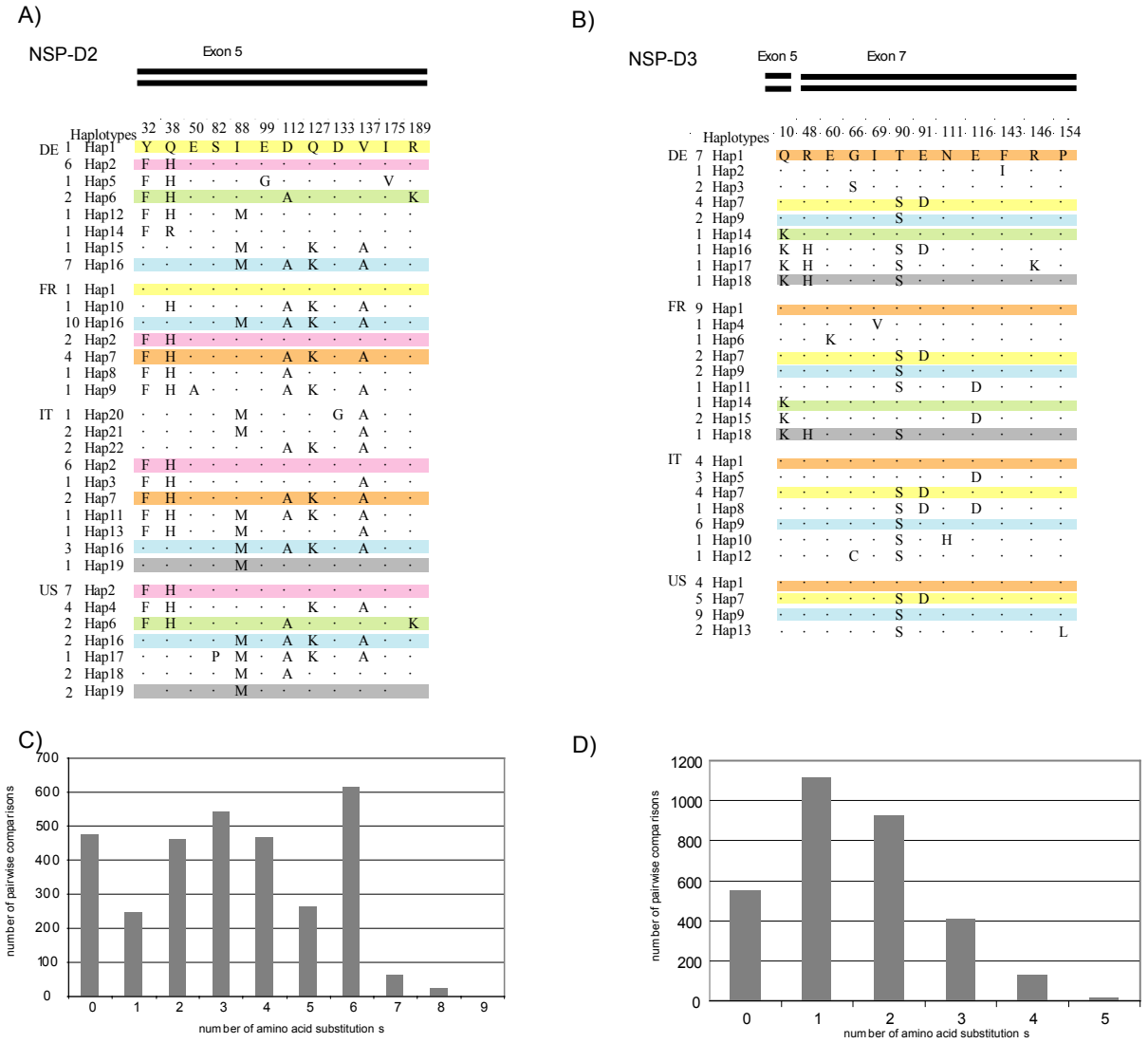


Figure 2: Overview of the haplotypes of NSP-D2 (A), NSP-D3 (B) present in each population. Amino acid variation with reference to first sequence is depicted for each unique allele, with shared alleles across populations highlighted with same color. Bars across the top of the sequences indicates the exon location (Figure 1). A pairwise distance comparison in C) and D) for NSP-D2 and NSP-D3 respectively. gives the number of pairwise comparisons (y-axis) that share the same number of differences (x-axis).

There was also variation among genes in the number and distribution of haplotypes across populations (Figure 2). Populations contained both distinct haplotypes as well as some haplotypes that were shared across populations (Figure 2a, b). There was also variation

across genes in terms of the age of alleles. Graphing the pairwise differences between all observed haplotypes (Figure 2c, d) reveals a non-normal distribution of pairwise differences in *NSP-D2* (Figure 2 c). The outlying peak indicates that there are two common alleles which differ from each other at 6 amino acids. These are found in the German population, where haplotypes two and sixteen are the two most common types, having 6 and 7 copies in the population respectively.

Population genetic structure

Population structure analysis using AMOVA on the coding region of all five gene fragments indicated no significant differentiation between populations and rather high variation within each population (Figure 3). Fst values show an overall low differentiation between populations, with most variation located within them (Table 2). After Bonferroni correction we detected significant differences only in NSP between Germany and the USA and in Ga3pdh between France and the other populations. COI shows Germany and the USA to be differentiated (Table 2). Similarly, across all genes and many population comparisons, the migration rate is high and in many cases indicative of complete gene flow.

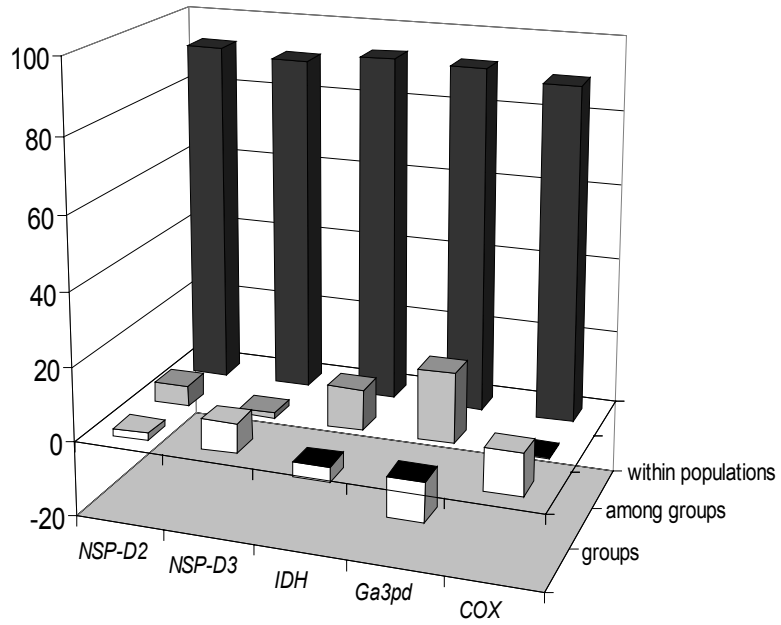


Figure 3: Results of the analysis of molecular variance (AMOVA) used as an estimate for genetic structure. On the y-axis the percentage of variation is graphed for each gene, as stated on x-axis. Comparisons are made within each population and between and within two groups, namely Europe and the USA.

In contrast the exact test for population differentiation suggests more population structure. While the AMOVA uses the number of genetic differences between haplotypes to assess structure, the exact test uses the haplotype identities themselves and is thus more sensitive to

recombinant haplotypes and recent gene flow. Exact test results indicate non-identical allelic distributions across more populations for NSP and Ga3pdh, hence more genetic structure at a finer scale (see also Table 2).

Table 2: **Estimates of population differentiation.**

	pop	NSP-D2	NSP-D3	IDH	Gap3	COX
Fst	DE-FR	0.03	-0.02	-0.01	0.14*	-0.05
	DE-IT	0.00	0.02	0.03	0.05	-0.03
	DE-US	0.01	0.09	-0.02	0.04	0.20*
	FR-IT	0.03	0.04	0.10	0.29**	0.00
	FR-US	0.06	0.13	-0.04	0.17*	0.14
	IT-US	0.03	0.07	0.07	0.02	0.11
Exact Test	DE-FR	0.0009*	0.0082*	0.4361	0.0279	0.4868
	DE-IT	0.0321	0.1859	0.3974	0.1286	0.5966
	DE-US	0.0373	0.0001*	0.9271	0.0177	0.0507
	FR-IT	0.1009	0.0911	0.0060	0.0022*	0.1747
	FR-US	0.0007*	0.1322	0.7671	0.0525	0.3480
	IT-US	0.0039	0.1107	0.0842	0.5326	0.2481
Migration	DE-FR	15.48	inf	inf	3.01	inf
	DE-IT	101.85	30.21	14.32	10.31	inf
	DE-US	45.25	4.92	inf	12.93	1.93
	FR-IT	16.88	11.76	4.65	1.22	inf
	FR-US	8.20	3.33	inf	2.40	3.04
	IT-US	15.79	6.49	6.50	32.73	4.23

Note.- Fst values, p-values for the exact test and the estimated absolute number of migrants between two populations (M) as implemented in the Arlequin program are given for every population comparison for every sequenced gene. Analysis always includes the whole sequenced fragments, thus both exons and intron in *NSP-D2* and *NSP-D3*. Values significant after Bonferroni corrections are marked with an asterisk, if significant value is < 0.0017 (after Bonferroni) two asterisks are used.

Tests for selection

We employed standard tests based on the null hypothesis of the standard neutral model. Tajima's D is not significant for any of the tested gene regions, but the most positive values are found in *NSP-D2* while all the other genes are negative or close to zero (Supplementary Table1). Fu and Li's D also show no significant values, either with or without *P. brassicae* as an outgroup. Analysis of the relationship of non-synonymous vs. synonymous polymorphism within species to non-synonymous vs. synonymous divergence between species used the McDonald-Kreitman test on data from *P. brassicae* as an outgroup (Supplementary Table 2). Results for all genes are not significant, although the numbers of fixed and polymorphic substitutions are high in NSP compared to the control genes with the exception of COI. Substitution rates range from 9 to 29 in NSP, compared to 0 to 13 in IDH and GA3pdh. COI has the highest synonymous substitution rate at 73. The multilocus HKA tests on either of the

NSP regions (*NSP-D2* and *NSP-D3*) showed no significant divergence from the standard neutral model. Both loci were tested individually against the control genes and in combination. Values of ω were calculated for each pairwise combination of the *P. rapae* NSP domains separately and in combination (i.e. as a full gene), with full gene comparisons finding ω values > 1 and the single domain analyses finding ω values > 2 (Figure 4).

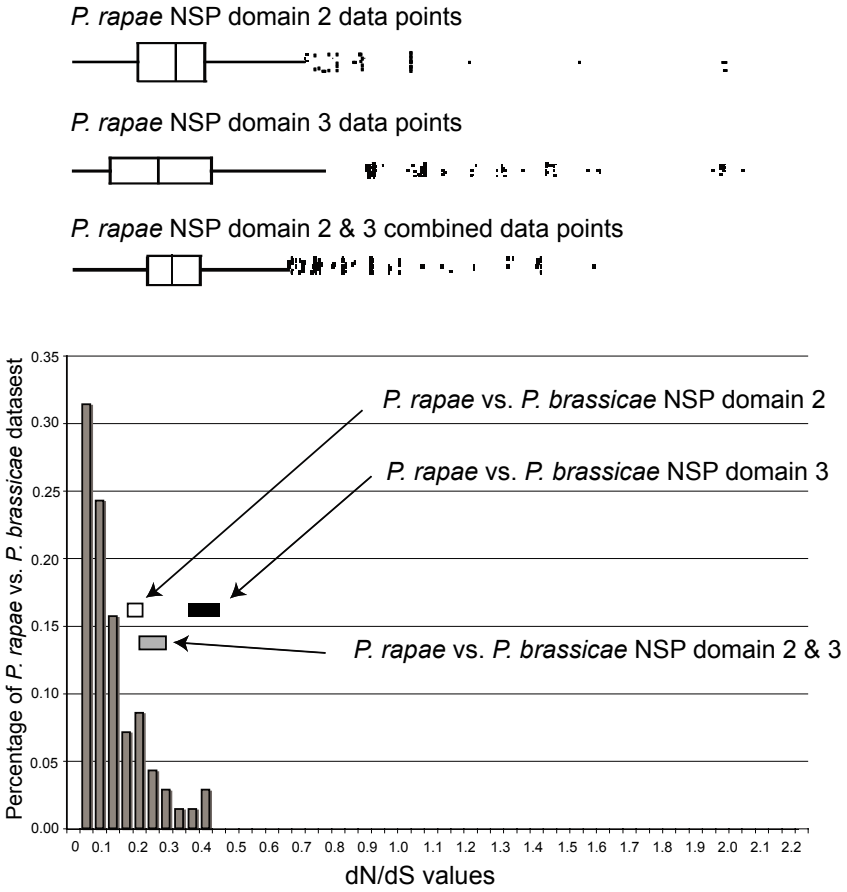


Figure 4: Distribution of dN/dS ratios from interspecific and intraspecific comparisons across genes. dN/dS values (x-axis) among 70 random genes compared between *P. rapae* and *P. brassicae* are represented as a gray histogram showing their frequency (y-axis). The range of dN/dS across all pairwise comparisons of *P. rapae* NSP domain 2 (white box) and 3 (black box) haplotypes with *P. brassicae* NSP sequence is shown to the right of the histogram. Above these, with respect to the x-axis, are shown the distribution of dN/dS ratios among all pairwise comparisons of *P. rapae* NSP domains 2 and 3. The datapoint row shows a stem and leaf plot for the data, where the box indicates the 25 % and 75 % quantiles, the line in the middle is the median, and the location of specific datapoints are represented as dots for the remaining outliers.

Interspecific divergence and dN/dS

P. rapae and its congener *P. brassicae* diverged approximately 11.75 Myr ago, based on temporal calibration of sequence divergence in the EF-1α gene as previously applied to Pieridae (Wheat et al. 2007). To compare the pattern of divergence at NSP with a genomic random sample of genes, 70 homologous gene sequences were identified in EST collections

of these two species. These ranged from a length of 183 to 792 bps, with a mean of 520.9 bps (std. dev. = 144) and 75% of sequences being > 430 bp long. This translates into a mean of 130 synonymous and 390 nonsynonymous sites per gene pair respectively (std. dev. 40.7 and 108 respectively). There was a range of between 5 to 71 bp differences between sequence pairs, with a mean of 27.2 bps (std. dev. = 13.2 bp).

Maximum-likelihood analysis of synonymous (dS) and nonsynonymous (dN) divergence between these *Pieris* species across these 70 genes finds substantial divergence, with the average dS = 0.189 (std. dev. = 0.073) and dN = 0.018 (std. dev. 0.018). However, these genes are, as expected, experiencing a fair amount of purifying selection with a mean ω = 0.097 (std. dev. = 0.091), with a range from 0 to 0.38.

The divergence and diversification at NSP between these species is much greater than the observed average genomic divergence. The mean dS and dN across the *P. rapae* combined NSP domains 2 & 3, when compared pairwise with *P. brassicae* NSP, is dS = 0.269 (std. dev. = 0.010) and dN = 0.071 (std. dev. = 0.002). The combined NSP domains 2 & 3 have a mean ω = 0.25, which is greater than 90% of the random gene ω values (Figure 4). Separate analysis of NSP domains 2 and 3 finds that domain 3 has a range of ω values overlapping and exceeding the largest values of ω in our random gene dataset, while domain 2 is lower than the combined domain NSP average (Figure 4).

3.4 Discussion

Our interest in the NSP gene originates from its role in host plant detoxification and the macroevolutionary consequences of its function (Wittstock et al. 2004; Wheat et al. 2007). With this functional and macroevolutionary insight the present study was focused at the population level, aiming to understand the microevolutionary dynamics of NSP in response to a complex host plant defense system. Here we use molecular tests of selection to help discriminate among alternative adaptive hypotheses and uncover segregating genetic variation upon which future ecological studies can focus.

Three alternative hypotheses were developed to assess the microevolutionary dynamics at NSP in the sampled *P. rapae* populations (Table 3). Hypothesis 1, positing local adaptation and a greater level of genetic diversity among than within populations, is not supported by our results. All populations contain a high diversity of NSP amino acid alleles with many alleles shared among populations (Figure 2). In addition, the NSP loci have low Fst values, high migration rates, and AMOVA results indicate greater variance within than among populations (Table 2 and Figure 3). While the exact tests of population differentiation in both *NSP-D2* and *NSP-D3* do give some hint of population structure (Table 2), this test is sensitive to the unique recombinant haplotypes found in each of the four populations which are at low frequency (Figure 2).

Table 3: Alternative hypotheses for the microevolution of NSP.

Hypothesis	Assumption	Expected pattern of variation
H0	No adaptive role	Reflects demographic history
H1	Unique local host plant adaptation	Variation within populations < variation among populations
H2	Generalist response to diverse host plant assemblages	Variation within populations > variation among populations
H3	Purifying selection upon optimal genotype	Little variation within and among populations

The high levels of amino acid polymorphism at NSP argue against Hypothesis 3, which posits purifying selection upon an optimal genotype with little variation among populations (Figure 2). In addition, NSP has ω values > 1 within *P. rapae* and ω values are in the 90th percentile when compared to a genomic average of interspecific comparisons (Figure 4). Thus, NSP is evolving at a faster rate compared to these reference genes and the observations of ω > 1 may be indicative of diversifying selection. These results as well as those from the AMOVA analysis are consistent with Hypothesis 2, where NSP diversity is expected to be higher within than among populations (Figure 3).

Let us now consider our null Hypothesis 0, which posits a demographic basis for the observed patterns of variation at NSP. First, we find a general pattern of greater genetic diversity within vs. among populations in the sequenced reference genes, which is consistent with the high dispersal abilities of *P. rapae*. Young females migrate long distances before egg laying (Jones et al. 1980), and human interference by long-distance transport of crop plants may lead to additional admixture in certain areas as suggested by an AFLP study of the genetic structure of urban and rural *P. rapae* populations in comparison to a native Japanese Pieridae species (Takami et al. 2004). In addition, the North American sample shows no clear distinction from the European populations, which may be indicative of recent and ongoing movement of *P. rapae* into the Americas instead of one historical introduction. Second, the other molecular tests of selection we have employed, with and without outgroups, do not detect departures from standard expectations of neutrality for any genes (Supplementary Table1). Together, these two observations suggest that Hypothesis 0 of selective neutrality cannot be rejected. However, demographic effects cannot account for the high level of amino acid diversity within NSP, the unequal distribution of this variation across the exons of NSP domains 2 and 3, or the fast rate of molecular evolution observed for NSP compared to a genomic average.

Our data suggest that the evolutionary dynamics acting at the NSP gene do not readily lend themselves to the standard molecular tests of selection. Hughes (2007) has argued that tests of neutrality only provide an appropriate test for very specific types of selection, which are not representative of selective events in general. For example, a large number of repeated selective sweeps are needed to generate a significant result in the McDonald-Kreitman test and they must occur over a short evolutionary time and in different areas of the protein (Hughes 2007). In addition, when the impact of biologically realistic conditions are used to assess the power of molecular tests of selection, such as when selection acts upon existing genetic variation in regions of moderate recombination rates and does not result in complete fixation of novel haplotypes, the power of tests to detect selective events is extremely weak (Nordborg and Innan 2003; Barrett and Schluter 2008). Our results presented here document that a large amount of existing variation, moderate levels of recombination, and no recent haplotype sweeps are all hallmarks of the NSP locus. Thus, if we posit that variation might be maintained within populations by the requirement of diverse hostplant use, we should expect to detect little if any molecular signature of selection from the standard tests we have employed here. With these issues in mind, we consider whether there are objective aspects of our data suggestive of a role in microevolutionary dynamics.

Compared to both our reference genes and level of genetic variation found within populations of diverse taxa in general, there is an unexpectedly high amount of amino acid polymorphism at NSP. The levels observed here are even greater than the well studied *Pgi* gene in *Colias* butterflies, which is remarkable in having 15 segregating amino acid sites spread across 556 codons (Wheat et al. 2006). Combining the information we have for NSP domains 2 and 3, we have identified 24 segregating amino acid polymorphisms across 346 codons. Considering that we have not even surveyed the first domain of NSP, it is very likely that NSP could harbor over 30 amino acid polymorphisms. Importantly, this diversity does not appear to be a general relaxation of constraint on amino acid variation randomly distributed across the gene, but a rather specifically restricted to the third exon of each NSP domain (Figure 1, Supplemental Material). This suggests a complex regime of selection pressure variation.

Inspection of the amino acid polymorphism within populations reveals very divergent haplotypes in NSP domain 2, some of which are at intermediate frequencies. In addition, numerous pairwise comparisons of NSP alleles in *P. rapae* find ω values > 1. While ω values > 1 are conservatively considered to be a hallmark of diversifying selection, ω values can be inflated due to the accumulation of deleterious nonsynonymous mutations when population size is low (e.g. (McBride and Arguello 2007)). However, given the high population densities and large range of *P. rapae*, and a genomic range of interspecific ω values consistent with relatively strong purifying selection (Figure 4), the accumulation of deleterious mutations at NSP due to demographic effects appears very unlikely.

Greater knowledge of the structure-function relationships of the NSP protein would facilitate understanding of the observed excess amino acid variation in the third exon of each of the domains. However, despite numerous efforts at heterologous expression, the secondary structure of NSP has not yet been experimentally determined (H.H-F. unpublished data; U. Wittstock pers. comm.) and structural prediction programs fail to produce consistent models. Nevertheless, the amino acid haplotypes identified in this study now provide the opportunity to study their relative functional performance in larval feeding assays across diverse hostplant species (Heidel-Fischer, ongoing work). While patterns of molecular variation are suggestive of a potential microevolutionary adaptive role for NSP, functional study is necessary to

determine whether the uncovered genetic variation has the potential for performance and fitness consequences in the wild (Dean and Thornton 2007).

NSP appeared more than 80 million years ago as an evolutionary novelty enabling a host range shift onto Brassicaceae and a subsequent increase in species diversity of the group (Wheat et al. 2007). The triple domain NSP was formed by two rounds of duplication of an original single domain enzyme (Fischer et al 2008). Our ability to reconstruct the evolutionary history of NSP (Fischer et al. 2008), and its maintenance in the genome is due at least in part to the purifying selection which we observe in the first two exons of domain 2 and 3 (Figure 1). In contrast, the observed high levels of intraspecific amino acid variation and fast divergence in the third exon of these domains suggests this region is experiencing very different selective dynamics more in the range of relaxed purifying selection and potentially diversifying selection (Figure 1). Although we do not know precisely the mode of action of NSP, the repeated domain structure suggests the potential for functional independence of these three individual domains. Such structural independence could allow for slightly deleterious mutations in one domain to be compensated by the independent functionality of the other two domains.

In conclusion, the microevolutionary dynamics at the NSP gene are complex. The NSP locus harbors an extremely high amount of amino acid diversity unequally distributed across its repeated domain structure which is suggestive of diversifying selection. Patterns of nucleotide diversity and molecular tests for selection rule out the potential for strong local adaptation, as well as directional and strong purifying selection. Patterns of genetic variation fail to provide a clear signature of historical selection at the micorevolutionary level, highlighting the necessity for functional study of the diverse set of NSP alleles we have uncovered in *P. rapae*.

Supplementary Table 1: Summary statistics for molecular tests for selection. No tests had a value of $P < 0.05$ under the standard neutral model. Tests employing an outgroup are indicated w out.

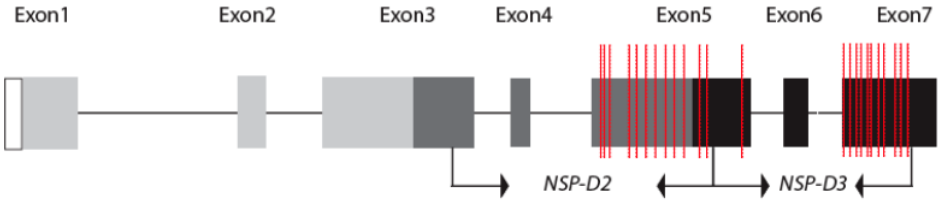
		Taj D	Fu &Li D	Fu &Li D w out	Fay & Wu H w out
NSP-D2	DE	0.52466	-0.01758	0.45161	-1.07368
	FR	0.62481	0.3006	0.63641	-1.95238
	IT	1.65115	0.84646	0.90748	-0.73684
	US	1.02003	1.01419	1.06926	-1.88421
NSP-D3	DE	0.0354	0.22999	0.18187	-0.01053
	FR	-0.27156	-0.3242	-0.45619	-0.58947
	IT	0.0785	-0.68651	-0.47456	-0.48421
	US	-0.07133	0.86241	0.9011	-0.71579
IDH	DE	-1.13975	-0.49086	0.9011	-3.29474
	FR	-0.48363	0.09399	0.02404	0.54902
	IT	-1.34736	-0.70114	0.16108	-1.30719
	US	-0.89286	-0.97212	-1.21763	0.2381
Ga3pd	DE	-1.26827	-0.84169	0.20307	-0.8366
	FR	0.68713	1.00649	1.01226	-0.30526
	IT	-1.53674	-1.7989	-1.64146	-1.33333
	US	0.79344	1.1232	1.14969	0.88889
COX	DE	0.53266	0.73372	1.08928	-2.21429
	FR	-0.3401	-0.21328	0.64362	-7.64286
	IT	-0.61245	-0.07256	0.09829	-5.28571
	US	0.05031	0.38925	0.23258	-0.19048

Supplementary Table 2: Summary statistics for the McDonald-Kreitman Test as implemented in DnaSP.

Gene	Substitution	fixed	polymorphic	Fisher's exact test
NSP-D2	syn	29	18	0.641663
	nonsyn	24	11	
NSP-D3	syn	16	9	1
	nonsyn	24	13	
MA-D2	syn	36	22	0.53425
	nonsyn	22	18	
IDH	syn	11	13	1
	nonsyn	3	4	
Ga3pdh	syn	7	10	0,508772
	nonsyn	0	2	
COX	syn	73	10	0,181992
	nonsyn	0	2	

Supplementary Material

Comparison of synonymous and non-synonymous site change across the NSP domains.



Exon	bps seq'd	codons seq'd	ss changes	ns changes
Domain 2 - exon 3	21	7	0	0
Domain 2 - exon 4	70	23	4	0
Domain 2 - exon 5	356	119	10	10
Domain 3 - exon 5	198	66	5	3
Domain 3 - exon 6	70	23	2	0
Domain 3 - exon 7	323	108	6	11
sum	1038	346		

totals	bps seq'd	codons seq'd	ss changes	ns changes
1st exon of Domains 2 & 3	219	73		
2nd exon of Domains 2 & 3	140	46		
1st + 2nd exons of Domains 2 & 3	359	119	11	3
3rd exon of Domains 2 & 3	679	227	16	21

Analysis by bp: ns changes

```
====== Input Tabla ======

   3    21 |   24
 359   679 | 1038
 ____________________
 362   700 | 1062

======== Output ========
```

G-test
G value: 5.994
P-value: 0.01435*
G with Williams' correction: 5.854
P-value: 0.01554*
G with Yates' correction: 4.777
P-value: 0.02884*

Analysis by bp: ss changes

```
====== Input Tabla ======

  11    16 |   27
 359   679 | 1038
 ____________________
 370   695 | 1065

======== Output ========
```

G-test
G value: 0.430
P-value: 0.51215 (not significant)
G with Williams' correction: 0.421
P-value: 0.51653 (not significant)
G with Yates' correction: 0.206
P-value: 0.65006 (not significant)

as implemented in DNAsp

4. Chapter III: Gene expression in a generalist butterfly upon feeding on different hostplants

Abstract

The mechanisms that shape the hostplant range of herbivorous insect are to date not well understood but knowledge of these mechanisms and the selective forces that influence them can expand our understanding of the larger ecological interaction. Nevertheless, it is well established that chemical defenses of plants influence the host range of herbivorous insects. While hostplant chemistry is influenced by phylogeny, also the growth forms of plants appear to influence the plant defense strategies as first postulated by Feeny (Feeny's "plant apparency" hypothesis). In the present study we aim to investigate the molecular basis of the diverse hostplant range of the comma butterfly (*Polygonia c-album*) by testing differential gene expression in the caterpillars on three hostplants that are either closely related or share the same growth form. The data suggest a complex interaction between the comma butterfly and its hostplants. On the one hand, each plant species appears to require a very specific subset of genes to be regulated in the midgut upon feeding, on the other hand species that share a growth form or are closely related have a higher agreement of gene regulation in the midgut of the caterpillar than species that do not share these traits. No known detoxifying enzymes were found to be differently regulated on different hostplants, suggesting the use of very broad acting continuously expressed detoxifying genes in *P. c-album* caterpillars.

4.1 Introduction

Chemical defenses of plants influence the host range of herbivorous insects (Dethier 1941, 1954, Fraenkel 1959, Thorsteinson 1960, Ehrlich and Raven 1964). Although by no means the only factor involved in shaping insect-host associations (Bernays 1989, Roy 2001), few researchers would argue against its general importance. However, there is an ongoing argument about why plant defense chemicals are similar. There are two ways for evolution to achieve similarity: either through shared ancestry or through evolutionary convergence (or parallelism). Ehrlich and Raven (1964) suggested that related insects tend to feed on related plants, and several other studies have continued to demonstrate a role of hostplant phylogeny (shared ancestry) on patterns of hostplant use (Futuyma et al. 1995, Menken 1996, Becerra 1997, Janz and Nylin 1998, Janz et al. 2001, Ronquist and Liljeblad 2001, Lopez-Vaamonde et al. 2003, Kergoat et al. 2005, Murphy and Feeny 2006). Hence, there is strong support for a historical component in patterns of hostplant use.

On the other hand, several authors have also pointed out that plant chemistry does not always follow phylogeny (Wahlberg 2001, Kergoat et al. 2005, Ohshima and Yoshizawa 2006). Feeny (1976) suggested that plant defense strategies should differ depending on their "apparency"; plants that are long-lived and/or physically large will always be found by attacking insects and should possess constitutive chemical defenses such as tannins, terpenes, and flavanoids (see also Futuyma 1976, Wasserman 1979, Chew and Courtney 1991, Miller et al. 2007) that have a quantitative, dosage-dependent effect. Unapparent plants, with lower risk of detection by herbivores, should instead tend to utilize induced chemical defenses. Trees should be the most apparent of plants as they are both physically large and long-lived. According to the apparency hypothesis, trees should then tend to have more convergent constitutive defenses than herbs, and as a consequence we should see more host shifts involving trees than herbs.

In a phylogenetic reanalysis of Ehrlich and Raven's (1964) study on butterfly and plant coevolution Janz & Nylin (1998) found strong effects of both plant phylogeny and growth form on patterns of host use among butterflies. An overwhelming majority of host shifts occurred while feeding on trees, giving support for Feeny's "plant apparency" hypothesis. Trees appeared to serve as a "bridge" that could facilitate host shifts between distantly related plants.

Hence, there is support for both shared ancestry and convergent evolution in the large-scale chemical structuring of insect-host associations, but the mechanistic basis remains largely unknown. However, recent years have seen great progress in understanding of the molecular mechanisms that enable insect to feed on certain hostplants. In general it is assumed that insects apply phase I and phase II detoxifying enzymes to metabolize secondary plant compounds. Several studies have for example revealed the important role of the cytochrome P450 enzyme family for detoxification of plant secondary compounds as well as insecticides. (Berenbaum et al. 1996, Daborn *et al.*, 2002, Zangerl and Berenbaum 2003, Li et al. 2004, Mao et al. 2006, 2007, Berenbaum and Feeny 2008). Glutathione S-transferases (GST) have also been shown to be induced in generalist and specialist lepidopteran larvae upon feeding on their hostplants (Wadleigh *et al* 1987, (Yu 1982). Wittstock et al (2004) identified the Nitrile-specifier protein (NSP) in *Pieris rapae*. NSP redirects the glucosinolate hydrolysis and by that enables the Pierinae butterflies to feed on the plant family Brassicaceae. Further research has been done on the evolution of NSP showing its evolution by domain and gene duplication from a gene of unknown function that is widespread in insect species (Wheat et al. 2007, Fischer et al. 2008). *Plutella xylostella* also feeds on glucosinolate containing plants. Here the Glucosinolate sulfatase (GSS) inhibits the hydrolysis of glucosinolates completely by forming desulfo-glucosinolates. In spite of progress in recent years, much is still to be discovered in the detoxification mechanisms of insects.

In the present study we aim to investigate the molecular basis of the diverse hostplant range of the comma butterfly (*Polygonia c-album*, Lepidoptera: Nymphalidae), by testing Feeny's "plant apparency" hypothesis. *P. c-album* is a widespread polyphagous butterfly species of the family Nymphalidae. It is found all over Eurasia, from England to Japan and from the center of Sweden to the northern tip of Africa. The larvae can be found on hostplants from several taxa: the "urticalean rosids" *Urtica, Humulus* and *Ulmus* and the distantly related *Salix* (Salicaceae)*, Ribes* (Grossulariaceae)*, Betula* and *Corylus* (Betulaceae) (Nylin 1988); hence the species is at the extreme end of polyphagy among butterflies, although by no means an indiscriminate generalist. For this study, we used a test array with three hostplants of *P. c-album* that are either closely related (Stinging nettle *Urtica dioica* and Wych Elm *Ulmus glabra* – both in Urticales) or share the same growth form (Great Sallow *Salix caprea* and *Ulmus glabra* – both trees). Following Feeny's "plant apparency" hypothesis we expected to find more similarities in the gene expression profiles of caterpillars that have been feeding on

plants that either have a shared ancestry (*U. dioica* and *U. glabra*) or belong to the same growth form (*S. caprea* and *U. glabra*).

4.2 Material and Methods

Larval rearing and preparation

The stock used in the experiments was the offspring of four female comma butterflies collected in early May 2007 in the area near to Stockholm in Sweden. The females had already mated in the wild with unknown males and were put into cages for oviposition. Each female was presented with the hostplants stinging nettle (*U. dioica*) and Great Sallow (*S. caprea*). Eggs were counted in the beginning of each day and were incubated in small jars on a sun-lit windowsill until hatching.

Larvae of each female were evenly spread across the three hostplants. They were raised on stinging nettle, Great Sallow or elm (*U. glabra*) in individual jars. The jars were placed in a climate room (temperature 20 °C, LD 12:12) where larvae were raised to the 4th instar before dissection of the midgut. Plants were changed when needed due to withering or feeding. To maintain humidity, water was sprayed over the jars twice a day. Jars were changed randomly to avoid position effects. Between 10 and 43 individuals from each family were dissected, for a total of 109 individuals across the three different diets. Midguts and the rest of the larval body were preserved separately in RNAlater.

RNA Isolation and Reverse Transcription

Larvae were dissected, and the midguts and restbodies were stored in RNA*later*® (Ambion). Tissue samples were pooled (10-13 individuals) according to the larval diet (*Salix*, *Urtica*, *Ulmus*). The guts were homogenized by a Ultra-Torax homogenizer (Beckman Coulter Scientific) in TRIzol (Invitrogen) reagent and restbodies were crushed in liquid nitrogen. For all samples TRIzol Reagent was used to isolate the RNA according to the manufacturer's protocol with the following modifications. After adding chloroform to separate the phases, the tubes were stored for 15 minutes at 4 °C before centrifugation. To precipitate the RNA, the solution was stored at -20°C overnight. After precipitation the RNA solution was centrifuged for 30 min at 4 °C. The obtained dried pellet was dissolved in 90 µl RNA storage solution (Ambion), and any remaining genomic DNA contamination was removed by DNAse treatment (TURBO DNAse, Ambion). The DNAse enzyme was removed and the RNA was

further purified by using the RNeasy MinElute Clean up Kit (Qiagen) following the manufacturer's protocol and eluted in 20 µl of RNA storage solution (Ambion).

Differential gene expression

To study differential gene expression between *P. c-album* larvae grown on different plants (*U.* glabra, *S. caprea* and *U. dioica*) the DEG GeneFishing Kit was used (SeeGene), following the manufacturer's protocol with a few modifications. The GeneFishing allows the amplification of the same set of genes from different samples due to a 10-mer core of arbitrary annealing control primers. By not exceeding the exponential phase of the PCR amplification, differentially expressed genes can be identified on an agarose gel.

In short, 3 µg of DNA-free total RNA was converted into single-stranded cDNA using annealing control primer one (dTACP1) and a mixture of different reverse transcriptases (Array Script, Ambion; Power Script, Clontech; Bioscript, Bioline). Second-strand cDNA synthesis and subsequent PCRs were performed as described in the DEG GeneFishing protocol. PCR products were separated and visualized on a 2% agarose gel. Differentially expressed bands were cut out from the agarose gels and PCR products extracted using Zymoclean Gel DNA Recovery Kit ™ (Zymo Research) according to the manufacturer's instructions. DNA fragments were cloned into the pCR II TOPO vector (Invitrogen). Eight clones were picked for each extracted band and further processed.

DNA Sequencing and Analysis

Plasmid minipreparations from bacterial colonies grown in 96 deep-well plates were performed using the 96 robot plasmid isolation kit (Eppendorf) on a Tecan Evo Freedom 150 robotic platform (Tecan). Single-pass sequencing of the 5' and 3' termini of individual clones was carried out on an ABI 3730 xl automatic DNA sequencer (PE Applied Biosystems).

Vector clipping, quality trimming and sequence assembly was done with the Lasergene software package (DNAStar Inc.). BLAST searches were conducted on a local server using the National Center for Biotechnology Information (NCBI) blastall program and best hits were recorded. When two independently indentified differentially expressed sequences clustered in the assembly in the same contig, it was assumed to be a recent duplication.

Quantitative real-time PCR

500 ng of DNA-free total RNA was converted into single-stranded DNA using a mix of random and oligo-dT20 primers according to the ABgene protocol (ABgene). Real-time PCR

oligonucleotide primers were designed using the online Primer3 internet based interface (http://frodo.wi.mit.edu). Primers were designed by the rules of highest maximum efficiency and sensitivity rules were followed to avoid formation of self and hetero-dimers, hairpins and self-complementarity. Gene-specific primers were designed on the basis of sequences obtained for selected *P. c-album* genes and two additional genes as potential house-keeping genes (ribosomal protein subunit S 18 and elongation initiation factor 4 alpha) to serve as the endogenous control (normalizer). Both house-keeping primers were tested thoroughly. RPS 18 was the most consistent gene, and was therefore used for the further analysis. QRT-PCR was done in optical 96-well plates on a MX3000P Real-Time PCR Detection System (Stratagene) using the Absolute QPCR SYBR green Mix (ABgene) to monitor double-stranded DNA synthesis in combination with ROX as a passive reference dye included in the PCR master mix.

Results

Function and patterns of differentially expressed genes

P. c-album larvae were raised to the 4th instar on three different natural hostplants, namely the stinging nettle, (*U. diocia*), the great sallow (*S. caprea*) and the elm (*U. glabra*). Midgut and restbody RNA was then isolated, transcribed to cDNA and differences in gene expression analyzed using the GeneFishing method (SeeGene). For a selected number of sequences the expression profile obtained by the GeneFishing method was independently confirmed using qRT-PCR.

In total we identified 120 differentially expressed genes, 55 expressed genes in the midgut of *P. c-album* and 65 in the restbody (Table 1 and Table 2). In the midgut six sequences gave no hit in BLAST searches and in the rest body 2 sequences gave no hit. We identified the potential function of the genes found with the GeneFishing protocol using BLAST searches (Table1). Many of the differentially expressed genes in the midgut are likely to be involved in metabolism and digestion, ranging from protein degradation to starch and lipid breakdown for nutrient acquisition. We could also identify eight ribosomal genes and seven genes that are involved in translation regulation and maintaining the DNA structure in the cell nucleus. Furthermore, we found three genes involved in immunity and one gene with a predicted transmembrane transport domain (Table 1). In the midgut pairwise similarities in upregulation were higher between plants that shared either growth form (*U. glabra* and *S. caprea*) namely

10 or were phylogenetically closer related (*U. glabra* and *U. diocia*), namely 11 (Table 3 and Figure 1). For this all sequences were scored that showed an upregulation (+, or ++) in two hostplants and were absent (-) in the third, or vice versa. We identified differentially expressed proteases and four differentially expressed genes of unknown function. We did not identify any differentially expressed known detoxifying enzymes in the midgut.

Table 1: Differentially expressed genes identified from larvae fed on three different hostplants (*Salix*, *Ulmus*, *Urtica*) from the GeneFishing experiments of the midguts of *P. c-album*.

Contig	Hit	e value	Salix	Urtica	Ulmus
	Digestion				
40	chymotrypsin-like protease [Helicoverpa armigera], CAA72952.1	9.00E-84	++	+	+
123	chymotrypsinogen-like protein 3 [Manduca sexta], CAM84318.1	3.00E-35	-	++	-
15	trypsin Ia precursor [Sesamia nonagrioides], AAT95347.1	7.00E-20	++	-	-
42	trypsin-like protease [Helicoverpa armigera], CAA72955.1	3.00E-37	+	++	+
63	trypsin-like serine protease [Ostrinia nubilalis], AAX62033.1	8.00E-36	+	+	++
30	RE38869p, alpha-amylase [Drosophila melanogaster], AAL48973.1	2.00E-97	-	+	-
82	lipase-1 [Bombyx mori], NP_001036966.1	3.00E-57	+	-	-
12	lipase-1 [Bombyx mori], NP_001036966.1	4.00E-62	-	++	+
35	lipase [Bombyx mandarina], AAX39410.1	9.00E-49	+	-	-
14	lipase-1 [Bombyx mori], NP_001036966.1	3.86E-48	-	++	+
14	lipase-1 [Bombyx mori], NP_001036966.1	3.86E-48	-	+	+
155	beta-glucosidase precursor [Spodoptera frugiperda], AAC06038	6.00E-04	-	++	+
39	alpha-amylase 3 [Diatraea saccharalis], AAP97394.1	7.00E-26	++	+	+
68	serine protease precursor [Bombyx mori], NP_001036826.1	3.00E-61	++	+	+
1	serine protease [Bombyx mandarina], AAX39408.1	1.00E-18	+	++	+
68	serine protease precursor [Bombyx mori], NP_001036826.1	3.00E-61	+	++	+
98	serine protease [Bombyx mori], AAX39409.1	2.06E-05	++	+	+
135	35kDa protease [Bombyx mori], NP_001037037.1	3.00E-24	-	++	-
27	zinc carboxypeptidase A 1 [Culex pipiens quinquefasciatus], XP_001851495.1	1.00E-65	-	-	++
27	zinc carboxypeptidase A 1 [Culex pipiens quinquefasciatus], XP_001851495.1	1.00E-65	+	+	++
	Immunity				
20	immune related protein [Spodoptera frugiperda], AAZ94260.1	3.60E-01	-	+	+
133	cobatoxin long form B [Spodoptera frugiperda], AAQ18900.1	2.70E-01	++	-	-
11	gloverin [Trichoplusia ni], ABV68856.1	6.00E-11	-	++	-
	Metabolism				
140	proteasome 26S non-ATPase subunit 9 [Bombyx mori], NP_001093084.1	4.00E-44	++	+	+
140	proteasome 26S non-ATPase subunit 9 [Bombyx mori], NP_001093084.1	4.00E-44	++	+	++
13	PREDICTED: similar to CG3609-PA, oxidoreductase [Apis mellifera], XP_624408.1	3.00E-93	-	++	-
152	NADH dehydrogenase subunit 1 [Himantopterus dohertyi], CAH59762	3.00E-10	++	+	++
94	short-chain dehydrogenease/reductase 2 [Bombyx mori], NP_001040155.1	2.00E-03	-	++	+
13	PREDICTED: similar to myo-inositol dehydrogenase [Nasonia vitripennis], XP_001603982.1	2.74E-91	++	+	+
104	PREDICTED:similar to short-chain dehydrogenase [Tribolium castaneum], XP_001812912.1	8.00E-36	++	++	+
80	PREDICTED: similar to tafazzin CG8766-PA, isoform A [Apis mellifera], XP_623296.1	4.00E-27	++	+	+
95	peripheral-type benzodiazepine receptor [Bombyx mori], NP_001040343.1	9.00E-09	+	++	++
6	selenoprotein M [Litopenaeus vannamei], ABI93178.1	2.63E-18	+	++	+
27	zinc carboxypeptidase A 1 [Culex pipiens quinquefasciatus], XP_001851495.1	1.00E-65	++	+	+
36	PREDICTED: similar to GM14009p, Long-chain acyl-CoA synthetases (AMP-forming) [Nasonia vitripennis], XP_001606071.1	1.42E-95	-	++	-
141	PREDICTED: hypothetical protein, S-adenosyl-L-homocysteine hydrolase [Nasonia vitripennis], XP_001599389.1	2.00E-38	+	++	++
148	XP_001861763.1	3.00E-62	++	+	+
	Ribosomal proteins				
65	ribosomal protein L39 [Bombyx mori], NP_001037251.1	8.17E-23	++	-	++
137	ribosomal protein L27A [Bombyx mori]NP_001037522.1	6.72E-56	+	++	+
92	ribosomal protein S16 [Bombyx mori], |NP_001037508.1	6.45E-34	+	++	+
50	ribosomal protein S25 [Bombyx mori] NP_001037275.1	1.06E-38	-	++	-
44	ribosomal protein L15 [Bombyx mori],NP_001037162.1	1.06E-38	+	-	+
43	ribosomal protein S2 [Bombyx mori], NP_001037564.1	1.00E-109	+	++	+
	Translation regulation/DNA Structure				
131	hypothetical protein AaeL_AAEL004467, similar to Chromobox protein [Aedes aegypti], XP_001649228.1	2.00E-04	++	-	-
57	eukaryotic initiation factor 5A [Papilio xuthus], BAG30779.1	2.00E-41	+	++	+
119	argonaute 2 [Bombyx mori], NP_001036995.1	7.05E-72	++	+	++
93	Sui1 protein [Bombyx mori], NP_001037082.1	3.90E-16	++	-	-
33	PREDICTED: similar to GA18560-PA , Predicted cysteine protease (OTU family)[Posttranslational modification, protein turnover, chaperones][Nasonia vitripennis], XP_001602110.1	1.71E-95	+	+	++
55	XP_001808987.1	3.00E-03	++	+	+
	Transport				
151	transmembrane emp24 protein transport domain containing 9 [Bombyx mori], NP_001040538.1	4.37E-104	+	++	++
	Unknown				
46	AGAP003713-PA [Anopheles gambiae str. PEST], XP_001230950.2	1.10E-02	-	-	+
134	AGAP003713-PA [Anopheles gambiae str. PEST], XP_001230950.2	5.00E-03	-	-	++
61	unknown [Helicoverpa armigera], ABU98617.1	3.00E-07	+	++	+
102	PREDICTED: similar to CG11964 CG11964-PA [Tribolium castaneum], XP_967379.1	5.00E-32	+	++	+
142	PREDICTED: similar to DC2 protein [Tribolium castaneum], XP_001811714.1	4.00E-42	+	-	+

Note.- Shown are the accession numbers and the best BLAST hits with e-values for each gene. Differential gene expression visualized as band intensity on agarose gels is depicted (++ strong, + moderate, - absent).

In the restbody of the comma butterfly larvae, the majority of the gene products are again most likely involved in the metabolism of the caterpillar or translate into ribosomal proteins. Seven of the sequenced gene fragments appear to have a structural role, and are mostly involved in chitin binding. We could also identify translation regulation genes and some involved in cellular architecture (Table2). One sequence showed moderate similarities (e-value of 0,00000008) to potential detoxification genes, namely a cytochrome P450 of *Plutella xylostella* (Table 2). There were more similarities in upregulated genes between the "unrelated" plants (18) than between the trees (11) or urticalean rosid groups (1) (Table 3). In both insect tissues seven differentially expressed bands had homologies only to genes of unknown function (Table 1 and 2).

Table 2: Identified differentially expressed genes for larvae fed on three different plants (Salix, Ulmus, Urtica) from the GeneFishing experiments of the restbodies of *P. c-album*.

Contig	Hit	e value	Salix	Urtica	Ulmus
	Cell Signaling				
N17	Der1-like domain family member 1, [Bombyx mori], NP_001040297.1|	5.741E-100	+	+	-
	Detoxification				
44	cytochrome P450, [Plutella xylostella], BAF95609.1|	0.00000001	++	+	++
	Hormon synthesis				
86	similar to CG10638-PA, [Papilio xuthus], BAG30781.1|	1E-28	++	++	+
N1	juvenile hormone epoxide hydrolase, [Bombyx mori], BAF81491.1|	1E-113	++	+	+
	Immunity				
43	conserved hypothetical protein, [Culex pipiens quinquefasciatus], |XP_001869584.1	3E-36	+	++	-
212	mitochondrial aldehyde dehydrogenase, [Bombyx mori], NP_001040198.1|	7E-39	+	+	-
	Metabolism				
N 9	mitochondrial aldehyde dehydrogenase, [Bombyx mori], NP_001040198.1|	8E-50	-	+	-
72	methionine-rich storage protein, [Spodoptera exigua], EU259816	4E-81	+	++	++
72	methionine-rich storage protein, [Spodoptera exigua], ABX55887	4E-81	++	+	+
72	methionine-rich storage protein, [Spodoptera exigua],ABX55887	4E-81	+	-	++
69	aspartate aminotransferase, [Bombyx mori], NP_001040337.1|	2E-35	++	++	+
4	alpha-amylase, [Helicoverpa armigera], ABU98613.1|	2E-77	+	++	-
136	putative reverse transcriptase, [Zingiber officinale], ABK60177	1E-20	+	++	+
10	vacuolar proton atpases [Aedes aegypti] XP_001657344.1	4.26E-24	++	++	+
145	PREDICTED: similar to BcDNA.GH02921, [Nasonia vitripennis,] |XP_001600178.1|	6E-53	+	++	-
161	PREDICTED: similar to CG2656-PA, [Apis mellifera], XP_625026.1 :Conserved hypothetical ATP binding protein.	6E-23	++	++	+
143	cyclic beta 1-2 glucan synthetase [Xanthomonas campestris pv. campestris str. ATCC 33913] NP_637420.1	3.34E+00	-	-	++
20	PREDICTED: similar to alpha isoform of regulatory subunit A, protein phosphatase 2, [Apis mellifera], XP_001120202.1|	5.6058E-47	+	++	+
2	hypothetical protein PFE0755c, [Plasmodium falciparum 3D7], XP_001351708.1|, NADH dehydrogenase subunit	0.53	++	++	+
38	trypsin, [Choristoneura fumiferana], AAA84423.1|	1E-23	++	++	+
147, 89	ubiquinol-cytochrome c reductase, [Bombyx mori], NP_001106738.1	0.00000007	++	++	+
44	26S proteasome regulatory ATPase subunit 10B, [Bombyx mori], NP_001040484	0.00004	+	-	+
N 32	serine protease precursor, [Bombyx mori],NP_001036826.1|	3E-50	++	++	+
47	serine protease precursor [Bombyx mori], NP_001036826.1		-	+	-
N 49	phosphate transport protein [Bombyx mori]	6.00E-104	-	+	-
	Ribosomal protein				
50	ribosomal protein S7 [Bombyx mori] NP_001037261.1	4.12E-67	+	+	-
79	ribosomal protein L35A, [Bombyx mori], NP_001037243.1	2E-37	++	++	+
25	ribosomal protein S11-1 [Bombyx mori] AAV34867.1	7.47E-37	++	++	+
154	ribosomal protein S25 [Bombyx mori] NP_001037275.1	1.27E-31	-	+	-
23	ribosomal protein S10 [Bombyx mori] NP_001037524.1	2.42E-51	++	+	+
23	ribosomal protein S10 [Bombyx mori] NP_001037524.1	2.42E-51	+	++	+
N124	ribosomal protein L39, [Bombyx mori], NP_001037251.1	9E-23	++	+	+
N 48	ribosomal protein L5, [Bombyx mori], AAV34814	5E-76	+	+	-
N 18	ribosomal protein L27 [Bombyx mori] NP_001037235.1	2.91E-68	+	++	+
N 46	ribosomal protein L6, [Bombyx mori], NP_001037132.1	2E-26	++	++	+
N 38	ribosomal protein L27A [Bombyx mori] NP_001037522.1	1.37E-55	++	+	++
N 15	ribosomal protein S18 [Bombyx mori] NP_001037269.1	1.57E-42	+	++	+
	Silk production				
9	BAB39503.1| fibroin L-chain [Papilio xuthus]	1E-46	+	++	-
144	fibroin L-chain, [Papilio xuthus], BAB39503.1|	7E-47	+	++	-
	Stress				
150, 78	heat shock cognate 70 protein, [Sesamia nonagrioides], AAY26452.2|	2E-37	+	+	++
	Structure				
122	Kettin1 protein, [Helicoverpa armigera],ABU96746.1|	4E-91	++	+	++
13	obstractor B, [Tribolium castaneum], NP_001073566 ,Chitin binding Peritrophin-A domain	1E-104	+	+	-
8	CU15_MANSE Cuticle protein CP14.6 precursor (MSCP14.6), Q94984|	0.001	++	++	+
1	cuticular protein CPR41A [Papilio xuthus], BAG30737.1	6E-36	++	++	-
N 51, 22	cuticular protein 78, RR-1 family (AGAP009876-PA), [Anopheles gambiae str. PEST], XP_318996	1E-10	++	++	+
N 52	pupal cuticle protein [Bombyx mori], NP_001119729	1.7	+	++	-
N 37	mCG13192, isoform CRA_a, [Mus musculus], EDL05910.1|	5.8706E-05	+	+	-
N 16	basement membrane collagen,[Brugia malayi], AAC46611.1|	2E-39	+	+	-
	Translation regulation/ Cell structrure				
59	eukaryotic initiation factor 5A [Papilio xuthus], AB264704	2E-41	++	++	+
74	histone H3.3 type 2, [Culex pipiens quinquefasciatus],XP_001865500	5E-43	+	-	++
101	small nuclear ribonucleoprotein E, [Bombyx mori], NP_001040370.1|	1E-20	++	++	-
49	elongation factor 1 alpha, [Papilio xuthus],BAG30769	2E-52	+	+	-
24	PREDICTED: similar to exosome complex exonuclease RRP41, putative, [Tribolium castaneum], XP_975230.2	8E-20	++	++	-
109	ribophorin, [Aedes aegypti], XP_001663283.1|	8E-69	+	+	-
N 54	PREDICTED: similar to shroom family member 4, [Danio rerio],XP_687426	1.3	-	++	-
	Transport				
156	binding-protein-dependent transport systems inner membrane component [Roseiflexus sp. RS-1], ABQ88845.1	0.231895	+	+	++
92	binding-protein-dependent transport systems inner membrane component [Roseiflexus sp. RS-1], ABQ88845.1	2.28E-01	+	-	+
36	sodium-dependent phosphate transporter, [Aedes aegypti], XP_001658313.1|	3E-76	+	-	+
N14	sodium-dependent phosphate transporter, [Aedes aegypti],XP_001658313.1|	6E-76	++	++	+
15	signal sequence receptor beta subunit, [Bombyx mori], NP_001040332.1|	2E-21	++	++	+
223	transport protein Sec61 alpha subunit, [Bombyx mori], NP_001037628.1	0.79	++	-	-
	Unknown				
14	hypothetical protein, [Paramecium tetraurelia], XP_001428456.1|	3.9	++	++	+
27	hypothetical protein UM00309.1, [Ustilago maydis 521], XP_756456	0.46	++	++	-
152	PREDICTED: hypothetical protein [Homo sapiens], XP_001714781	2.6	+	+	++
17	unknown [Drosophila pseudoobscura pseudoobscura] XM_002133986	0.041	+	-	++

Note.- Shown are the accession numbers and the best BLAST hits with e-values for each gene. Differential gene expression visualized as band intensity on agarose gels is depicted (++ strong, + moderate, - absent)

Table 3: Number of sequences that showed pairwise similarities in upregulation.

	Midgut			**Restbody**		
	Urtica-Ulmus	Salix-Ulmus	Salix Urtica	Urtica-Ulmus	Salix-Ulmus	Salix Urtica
Digestion	6	3	1	0	ne	ne
Immunity	2	1	0	0	0	2
Metabolism	1	2	0	0	5	2
Ribosomal Protein	0	3	0	0	1	2
Tr. Regul./DNA Str.	2	0	0	0	2	4
Unknown	0	1	2	0	1	1
Cell Signaling	ne	ne	ne	0	0	1
Detoxification	ne	ne	ne	0	0	0
Silk production	ne	ne	ne	0	0	2
Structure	ne	ne	ne	0	0	5
Transport	ne	ne	ne	1	2	0
total	**11**	**10**	**3**	**1**	**11**	**19**

Note.- Sequences were scored when present (++or +) on two diets and absent (-) on the third diet or vice versa. The number of pairwise similarities in upregulation in each category and in total in the gut and in the restbody tissue is diplayed. ne: not existent in this tissue.

Confirmation of differentially expressed genes

Of the total of 122 genes originating from GeneFishing, we picked 27 genes (18 from midguts and 9 from restbodies) to confirm differential gene expression patterns with qRT-PCR. We were able to see identical expression patterns in the GeneFishing and in the qRT-PCR results for 14 of the 27 genes (10 in midguts and 4 in restbodies). Partial similarity (same in relation to one or two diets) in expression patterns between two independent methods could be observed in 7 genes (4 in midguts and 3 in restbodies) and 6 genes (4 in midguts and 2 in restbodies) behaved differently (Table 4).

In the midgut, genes that showed similar expression patterns as in the GeneFishing experiment included proteins involved in digestion, namely chymotrypsinogen-like protein 3, serine protease, chymotrypsin-like protease, alpha-amylase, trypsin-like protease, trypsin Ia precursor, long-chain acyl-CoA synthetase and short-chain dehydrogenease/reductase 2. In addition, a ribosomal protein S16 and the immune response related protein cobatoxin showed similar expression patterns in the qRT-PCR and in the differential gene expression study. In the restbodies, alpha-amylase and the serine protease precursor were similarly expressed. The stress related heat shock cognate 70 protein and a potential cytochrome P450 also showed similar expression patterns by both methods.

Table 4: qRT-PCR results for *P. c-album* midguts and restbodies fed on three different plants (Salix, Urtica, Ulmus).

GENE	relative fold gene expression		match with Gene Fishing data			best BLAST hit
	Salix	Ulmus	Urtica-Ulmus	Salix-Ulmus	Salix-Urtica	
Midguts						
MC 123	**-7.79 ± 0.62**	**-2.90 ± 0.28**				gi|146327862|emb|CAM84318.1| chymotrypsinogen-like protein 3 [Manduca sexta]
MC 92	-1.18 ± 0.15	**-2.58 ± 0.24**				gi|112984394|ref|NP_001037508.1| ribosomal protein S16 [Bombyx mori]
MC 17	-1.29 ± 0.11	1.36 ± 0.08				gi|112983142|ref|NP_001037037.1| 35kDa protease [Bombyx mori]
MC 36	**-2.03 ± 0.01**	**-2.32 ± 0.15**				gi|156550737|ref|XP_001606071.1| PREDICTED: similar to GM14009p [Nasonia vitripennis]
MC 44	1.27 ± 0.16	1.09 ± 0.04				gi|54609223|gb|AAV34827.1| ribosomal protein L15 [Bombyx mori]
MC 81	1.02 ± 0.02	**-2.18 ± 0.03**				gi|112983352|ref|NP_001036966.1| lipase-1 [Bombyx mori]
MC 159	**-6.78 ± 0.06**	**-6.14 ± 0.04**				gi:61191881 serine protease [Bombyx mandarina]
MC 40	**2.22 ± 0.04**	-1.34 ± 0.04				gi|2463064|emb|CAA72952.1| chymotrypsin-like protease [Helicoverpa armigera]
MC 42	**-4.76 ± 0.03**	-1.87 ± 0.002				gi|2463070|emb|CAA72955.1| trypsin-like protease [Helicoverpa armigera]
MC 65	-1.09 ± 0.05	-1.16 ± 0.15				gi|112984118|ref|NP_001037251.1| ribosomal protein L39 [Bombyx mori]
MC 133	-1.47 ± 0.06	**-6.76 ± 0.03**				gi:33439724] cobatoxin long form B [Spodoptera frugiperda]
MC 151	-1.55 ± 0.04	-1.37 ± 0.03				gi|114052711|ref|NP_001040538.1| transmembrane emp24 protein transport domain containing 9 [Bombyx mori]
MC 28	**-4.14 ± 0.06**	-1.15 ± 0.02				gi|108881060|gb|EAT45285.1| zinc carboxypeptidase [Aedes aegypti]
MC 93	-1.18 ± 0.04	-1.48 ± 0.04				gi|112983000|ref|NP_001037082.1| Sui1 protein [Bombyx mori]
MC 13	1.69 ± 0.06	-0.54 ± 0.04				gi|66514540|ref|XP_624408.1| PREDICTED: similar to CG3609-PA [Apis mellifera]
MC 30	**-2.12 ± 0.02**	-1.79 ± 0.03				gi|157126491|ref|XP_001660906.1| alpha-amylase [Aedes aegypti]
MC 94	-1.50 ± 0.06	1.10 ± 0.02				gi|114050773|ref|NP_001040155.1| short-chain dehydrogenease/reductase 2 [Bombyx mori]
MC 97	**10.02 ± 0.05**	1.56 ± 0.13				gi|157113343|ref|XP_001657786.1| trypsin [Aedes aegypti]
Restbodies						
RC 1	-1.65 ± 0.04	**-3.02 ± 0.00**				cuticular protein CPR41A [Papilio xuthus], gi:183979370
RC 13	-1.54 ± 0.20	-1.01 ± 0.04				obstractor B, [Tribolium castaneum], gi:121582324,Chitin binding Peritrophin-A domain
RC 2	-1.22 ± 0.15	1.14 ± 0.04				hypothetical protein PFE0755c, [Plasmodium falciparum 3D7], gi:124506221, NADH dehydrogenase subunit
RC 4	**-8.31 ± 0.05**	**-16.68 ± 0.25**				alpha-amylase, [Helicoverpa armigera], gb|ABU98613.1|
RC 43	**-2.11±**	-1.04 ±				ref|XP_001869584.1| conserved hypothetical protein, Destabilase [Culex pipiens quinquefasciatus] 3E-36
RC 44	1.27 ± 0.01	**2.36 ± 0.14**				cytochrome P450, [Plutella xylostella], dbj|BAF95609.1|
RC 47	**-33.24 ± 0.44**	**-9.45 ± 0.04**				serine protease precursor [Bombyx mori], NP_001036826.1
RC 78	-1.38 ± 0.04	**2.04 ± 0.11**				gi|157064217|gb|AAY26452.2| heat shock cognate 70 protein [Sesamia nonagrioides]
RC 9	-1.90 ± 0.44	1.04 ± 0.25				gi|13383201|dbj|BAB39503.1| fibroin L-chain [Papilio xuthus]

Note.- Relative expression of genes of interest were normalized using RPS18 as an expression control. The gene expression of larvae fed on Urtica was used as a reference to which relative expression in larvae fed on Ulmus and Salix was compared (values are mean ± SD). Consistency with GeneFishing results is depicted in color-code. (black bars – agreement, white – disagreement). Comparisons are always one diet relative to another diet.

4.3 Discussion

Phylogenetically related plant species are expected to possess similar chemical defenses and might therefore demand similar detoxifying mechanisms from herbivores. Feeny's "plant apparency" hypothesis suggests that plants of the same growth form should also have similar defense strategies. Here, we analyze the gene expression of *P. c-album* caterpillars feeding on three different host pants that are either closely related (stinging nettle, *Urtica dioica* and wych elm, *Ulmus glabra* – both in Urticales) or share the same growth form (great sallow, *Salix caprea* and wych elm, *Ulmus glabra* – both trees) to find evidence for one or both plant defense hypotheses in the response of the caterpillars. We chose this butterfly species because it has an unusually diverse hostplant spectrum and is therefore at the extreme end of polyphagy among butterflies. In the GeneFishing approach we could identify a total of 120 differentially expressed sequences, 55 of them were expressed in the gut lumen of the caterpillars and 65 in the restbody. Independent verification by qRT-PCR experiments showed a good correlation of the expression profiles with the Gene Fishing data (Table 4).

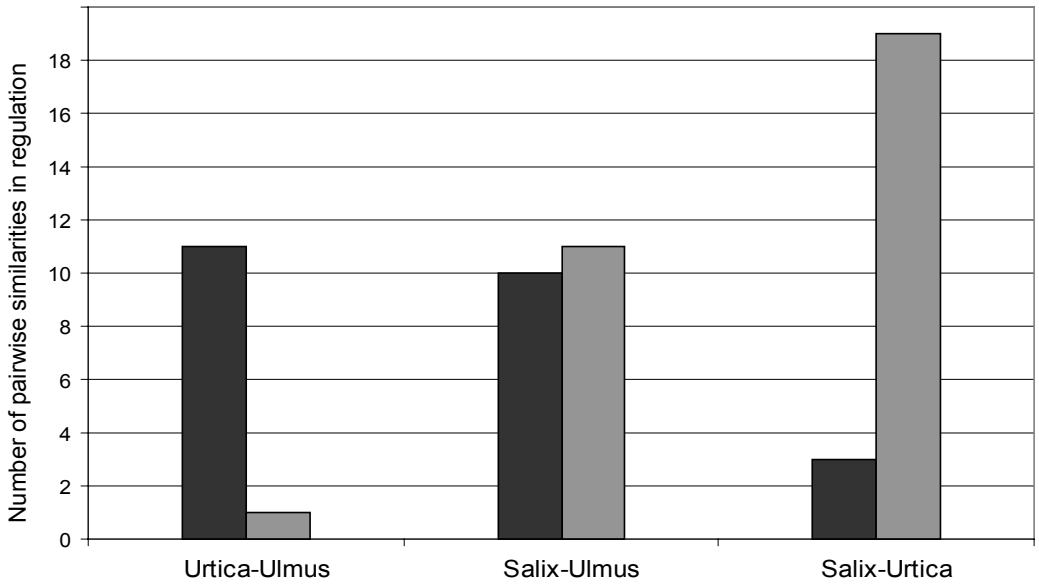


Figure 1: Number of sequences that showed pairwise similarities in regulation in the midgut (black bars) and the restbody (grey bars). Sequences were scored when present (++ or +) on two diets and absent (-) on the third diet or vice versa.

Secondary metabolites, leaf texture and nutrient contents vary greatly between different plant species and also depend on the plant age (Brenes-Arguedas et al. 2006). The three chosen *P. c-album* hostplants have different leaf architecture, with the stinging nettle possessing thin hairy leaves, and both trees waxy leaves, that differ in shape. Besides the prominent stinging trichomes that provide an effective defense against vertebrate herbivores (Pullin 1986) and also can inhibit small larvae movement, stinging nettle leaves contain the phenolic compound caffeic acid, tannins, nicotin in measurable amounts and flavonoid glycosides (Basaran et al 2000, Özen and Kokmaz 2003). In elm tree leaves flavonoid glycosides have also been shown to exist (Martin-Benito et al 2005), as well as the phenolic compound chlorogenic acid. In *Salix* leaves the contents of phenolic glycosides including salicin, chlorogenic acid and condensed tannins varies depending on age, with the highest concentration in young leaves (Jassib 2003). This leaves us with a complex and incomplete picture of secondary compounds in these three plant species (Table 5). Phenolic compounds, nicotine, tannins and flavonoids have been shown to affect insects feeding on them. Tannins for example are astringent, bitter plant polyphenols that either bind and precipitate or shrink proteins. Studies on the effect of tannins on the gut pH and redox potential of larvae found that many specialist and generalist insects have developed adaptations to cope with them (Johnson and Felton 1995, Johnson 2005). Tannin metabolites are for example oxidized during the gut passage in the aquatic caterpillar *Acentria ephemerella*. Resistance against polyphenols appears to be correlated in general with better repair mechanisms in the gut tissue that enables to cope with free radical

stress occurring during oxidation better (Gross et al. 2008). Nicotine acts as agonist of the postsynaptic nicotinic acetylcholine receptors of the insect central nervous system. Metabolism of nicotine has been attributed to the action of cytochrome P450 (Dowd et al 1983, Nauen and Denhol 2005). Flavonoids in contrast can be sequestered from the diet and used for protection and pigmentation. In *B. mori* for example three flavonoid glycosides could be isolated from the cocoon shell (Hirayama et al 2007). The role of these secondary plant compounds in the interaction between *P. c-album* and its hostplant and the adaptations *P. c-album* developed to them however, is not known.

Table 5: Short surveys of secondary plant compounds identified in *U. diocia*, *U. glabra* and *S. caprea*, that have been proven elsewhere to poses anti-herbivore attributes.

Compound group	*Urtica dioica*	*Ulmus glabra*	*Salix caprea*
Phenols	+[1) 2)]	+[5)]	+[4)]
Phenolic glycosides	–	–	+[4)]
Flavonoids	+[1) 2)]	–	–
Condensed tannins	–	+[5)]	+[4)]
Flavonoid glycosides	+[1) 2)]	+[3)]	+[5)]
Nicotine	+[1) 2)]	–	–

Note. – "+" indicates present in this species and "–" indicates not identified in this species
1) Basaran et al 2000, 2) Özen and Kokmaz 2003, 3) Martin-Benito et al 2005, 4) Jassib 2003, 5) Hegnauer 1973

Our expression data suggest a complex interaction between the comma butterfly and its hostplants. On the one hand, each plant species appears to require a very specific subset of genes to be regulated in the midgut upon feeding. This involves digestion proteins, immunity related genes, general metabolism genes and ribosomal proteins as well as translational and transport genes. In the restbody we also find cell signaling related domains, hormone synthesis genes and genes involved in silk production, suggesting gene regulation tuned specifically for each plant species. On the other hand, counting of upregulation versus downregulation of genes in the midgut shows a suggestive pattern (Table 3). Both species of the Urticales family (*Ulmus* and *Urtica*,) as well as both trees, have a higher agreement of gene regulation than do the stinging nettle and the great sallow (Table 1, Figure 1 and Table 3). Especially the digestion and ribosomal genes show clear differences here. This suggests, that phylogenetic and/or growth form relatedness demand more similar expression profiles in the midgut of the caterpillars. However, this only holds true for the midgut of the comma

butterfly; the restbody of the caterpillars show a completely different picture. Here the stinging nettle and great sallow diet in contrast share the highest similarity in gene expression, while the wych elm appears to require less gene upregulation in general (Table 2, Table 3, Figure 1). The midgut is the place of contact with the food bolus, and the location where the first detoxifying and digestive actions will take place, whereas the restbody is only indirectly involved in this process by receiving the solubilized products and only sometimes toxic or toxin degradation products. Here the actual nutrient content is of more importance. This might explain the converse picture we observe in those two tissues. While the defense compounds of the phylogenetically or growth form-related plants might be more similar, our data suggests that the nutritional value does not follow this line.

In the midgut we could not detect any differentially expressed known detoxifying genes, such as cytochromes P450s or glutathione-S-transferases (GST). However, those enzymes are with a high probability active in the gut of the comma butterfly larvae when encountering plant material. The GeneFishing method identifies solely differentially expressed genes. Hence, *P. c-album* might express such Phase I and Phase II detoxifying genes, but in a constant manner independent of the diet. Being a highly polyphagous insect species, the comma butterfly might be endued with very broad acting detoxifying enzymes that can cope with a wide variety of the compounds and hence are expressed continuously when feeding independent of the diet. The lack of differentially expressed detoxifying enzymes could also be due partly to an overlap of toxic chemistry on the three hostplants.
With the exception of cytochrome P450s and GSTs not much is known about detoxifying enzymes applied by generalist insects. Hence such expressed enzymes could not be identified in databases. In the midgut we identified 4 unknown differentially expressed genes, which could be unknown genes involved in detoxification. We could, however, identify a differentially expressed gene homologous to a cytochrome P450 of *P. xylostella* in the restbody of *P. c-album* larvae. Cytochrome P450s are also involved in the metabolism of many endogenous compounds, hence this P450 expresssed in the restbody is possibly not involved in the detoxification of plant defense compounds, but in general metabolism.

While the detoxifying mechanisms appear not to be differentially expressed we see genes belonging to other classes differentially expressed. Short-chain dehydrogenases (SCDH) for example are upregulated in the midguts of larvae feeding on *Urtica* and *Ulmus*. SCDH form a large protein family with highly different enzymes, which only share 15-30% identity among

each other (Jörnvall et al, 1995). They are present in all the life forms studied so far, have a wide substrate spectrum and are generally involved in cellular differentiation and signaling (Kallberg et al, 2002).

Not surprisingly also many digestive enzymes are differentially expressed in the midgut (Table 1). For example an alpha-amylase was down-regulated in both tree diets in comparison to the stinging nettle. Amylases are enzymes participating in carbohydrate digestion. It is known that insects possess different amylases for starch degradation (Terra & Ferreira, 2005). It is possible that the differential expression we observe is due to different starch contents in the stinging nettle compared to the other plants. We also found several serine proteases being differentially expressed upon feeding on the hostplants, namely trypsins and chymotrypsins. Plants possess proteinase inhibitors (PIs) that are insect inducible peptidases that can suppress insect proteinases and by that reduce the digestibility (Zavala, et al., 2004; Steppuhn & Baldwin, 2007). It has been shown that lepidopteran larvae adapt their proteinases expression profile to the PI content of their food plant, upregulating proteinase that are insensitive to the plant PIs (Terra & Ferreira, 2005). Our expression patterns suggest different proteinase requirements for the three hostplants. We also excised four bands that were identified as homologous to Bmlipase-1, of which three were highly expressed in larvae feeding on *Urtica* and one on *Salix*. Insect midgut lipases have been studied in few insects so far and little is known about differential expression of lipases in insects (Terra & Ferreira, 2005). In insects they form a gene family that underwent many duplication events (Horne et al. 2008) with resulting diverse and overlapping function. Bmlipase-1 from *Bombyx mori* shows a high antiviral activity against *B. mori* nucleopolyhedrovirus, although it is not inducible by viral infection. Its main function is probably as a digestive enzyme, as it is exclusively expressed in the gut tissue and has lipase activity (Ponnuvel, et al, 2003).

We also found many ribosomal proteins to be differentially regulated upon feeding on different hostplants, namely seven in the midgut and twelve in the restbody of *P. c.-album*. Ribosomal genes are considered to be stably expressed and have been suggested and used as housekeeping genes in expression analysis. However, there are a number of ribosomal proteins that have been found to be differentially expressed between tissues and developmental stages upon different treatments (Thorrez et al 2008), suggesting that there are major differences in expression patterns between different ribosomal proteins.

Different plants can possess very different microfloras (Meyling & Eilenberg, 2006), that affect immune response-related gene expression in the midgut of lepidopteran larvae (Freitak et al., 2007). We detected cobatoxin and gloverin homologous genes that are known to be inducible by bacterial challenge to be differentially expressed in the midgut tissue (Lundström, et al., 2002, Volkoff et al, 2003). We see gloverin expression only in *Urtica* fed larvae and cobatoxin in *Salix* fed larvae, suggesting different microbial environments or microbial load on those two species. In addition to differences in plant secondary metabolites, *P c.-album* must therefore also likely face different bacterial quantities and qualities on its various hostplants.

Our data suggest a complex picture of gene expression in response to hostplant feeding. While each plant evidently requires an unique set of genes regulation in the caterpillar, both phylogenetic relatedness and hostplant growth form appear to influence the expression profile of the polyphagous comma butterfly, in agreement with phylogenetic studies of hostplant utilization in butterflies.

5. General discussion

Although the defensive and counter-defensive molecules underlying many ecological interactions are known, the genetic mechanisms controlling these molecules are often unknown. Knowledge of these mechanisms, as well as the selective forces and adaptations that have shaped them, is necessary if we are to understand the evolution of ecological interactions. In this thesis the detoxification genes of two lepidopteran herbivores are characterized at different levels. The evolutionary origins of a detoxification adaptation are studied at the molecular level in a specialist lepidopteran herbivore (chapter I). Next, the microevolutionary dynamics of this detoxification gene in this specialist are investigated using molecular population genetics (chapter II). Finally, as a comparison to the study of a specialist herbivore, the molecular fundamentals of detoxification adaptation in a very broad generalist lepidopteran species are investigated (chapter III). While the specialist needs to avoid counter defenses of the main host plant family and can fine tune its detoxification against these specific plant species, the generalist herbivore needs to have a detoxifying system which can respond to a broad range of defenses presented by different plant families. Together these chapters provide important insights into the origins and ongoing evolution of detoxification mechanisms in specialist herbivores and compares these insights with the detoxification mechanisms used by a generalist herbivore.

The ability to adapt to new environments such as a new host plant can arise through different adaptive mutations. In chapter I the origins of the novel detoxifying enzyme NSP were investigated in the butterfly family Pieridae. It was found that NSP evolved through domain duplication followed by gene duplication from a single domain gene called SDMA. SDMA, NSP and MA, the paralog of NSP, are all members of the NSP-like gene family that appears to be widespread within the insects and underwent multiple gene and domain duplication events. In eukaryotic species, duplicated genes arise at a very high rate on an evolutionary time scale, with an average rate of about 0.01 duplications per gene per million years (Lynch and Conery 2000). While duplication generates potentially substantial molecular substrates for the origin of evolutionary novelties, the fate awaiting most gene duplicates appears to be silencing with selection only retaining favorable duplicates. Chapter I illustrates that after duplication, the resulting NSP gene was retained in the Pierinae and facilitated adaptation to feeding on Brassicaceae plants.

Gene duplication is widely found in the plant and animal kingdoms (Force et al. 1999; Briscoe 2001; Chinen et al. 2003; He and Zhang 2005; Loppin et al. 2005; Benderoth et al. 2006; Frentiu et al. 2007; Hoffmann et al. 2007) and in spite of the high rates of silencing it is now firmly established that duplication of genes is a major contributor to the evolution of novel adaptive function (Lynch and Conery 2000; Lynch 2007). However, the evolution of novel functions after duplication is difficult to trace. This is mostly due to problems identifying novel functions of a duplicated gene with no homologs serving similar functions. In the case of the NSP-like gene family, the function of the members other than NSP is unknown. However, research in *Periplaneta americana*, *Aedes aegypti*, *Blatella germanica* and most recently in *Tenebrio molitor* all show that members of the NSP-like gene family are expressed in the midgut of the insects after food intake (Pomes et al. 1998; Wang, Lee, and Wu 1999; Chad Gore and Schal 2005; Shao et al. 2005; Ferreira et al. 2008). In *A. aegypti*, AEG12 is only expressed in adult females and is strongly induced after a blood meal (Shao et al 2005) and in *B. germanica* the expression of Blag1 is also increased after food intake in adult females (Chad Gore and Schal 2005). It has also been suggested that PMAP in *T. molitor* has a role in peritrophic membrane formation (Ferreira et al 2007), but there is no evidence supporting this hypothesis so far. From these studies a digestive function of at least some members of the NSP-like gene family is suggested, but no detoxification function has been observed in any of these genes. Hence, while tissue and temporal expression patterns of the SDMA gene and the derived NSP gene are conserved (chapter I), the function that the multidomain NSP gene acquired, is to our knowledge very divergent from SDMA. This makes NSP a unique detoxifying gene with no known homolog serving similar functions. In this light, the results of chapter I studying the molecular evolution of NSP provide a rare example of proven neofunctionalization after duplication.

The variability of domain numbers in members of the NSP-like gene family that have likely retained the original function is striking. While the SDMA genes in all lepidopteran and AEG12 of *A. aegypti* have one domain, *T. molitor* has a three domain gene, *Tribolium castaneum* has a gene with eight repeats, and *B. germanica* and *P. americana* have at least two to three domain repeats in their NSP-like gene family genes. Hence, there appears to be an ongoing driving force for duplication even while fulfilling the original function. This force for duplication might have enabled the emergence of NSP in the Pieridae by favoring domain duplication of the original function before neofunctionalization.

How neofunctionalization occurs is still under debate. In the classical gene duplication model, one gene copy drifts neutrally and is able to accumulate mutations while the other copy retains the original function. In the majority of cases mutations will lead to the loss of function of one copy, but in rare cases mutations will by chance yield beneficial functions. When such beneficial mutations are fixed by selection, both the original and the neofunctionalized gene copies will subsequently remain active in the genome (Ohno 1970) (Figure 1a). Alternatively, under the DDC (Duplication-Degeneration-Complementation) model of Force et al. (1999) complementary degenerative mutations in different regulatory elements of duplicated genes facilitate the preservation of both duplicates and subsequently subfunctionalization. Neofunctionalization in this model can be acquired by one gene copy gaining a beneficial expression pattern at the expense of an ancestral subfunction (Figure 1b). With the available data it is not possible to decide which of these models might be more accurate for the evolution of NSP. No difference in regulatory elements between NSP and MA could be detected as both NSP and MA are expressed solely in the gut lumen of insect and are dependent on food intake (own preliminary data and Wittstock et al 2004). However, many other not so readily testable regulatory elements could and should show differences between NSP and MA. They could differ in expression across the fore-, mid-, or hind-gut of the caterpillar. Future research of location, function and mode of action of NSP and MA might shed light on the mechanism of neofunctionalization that gave rise to NSP and MA.

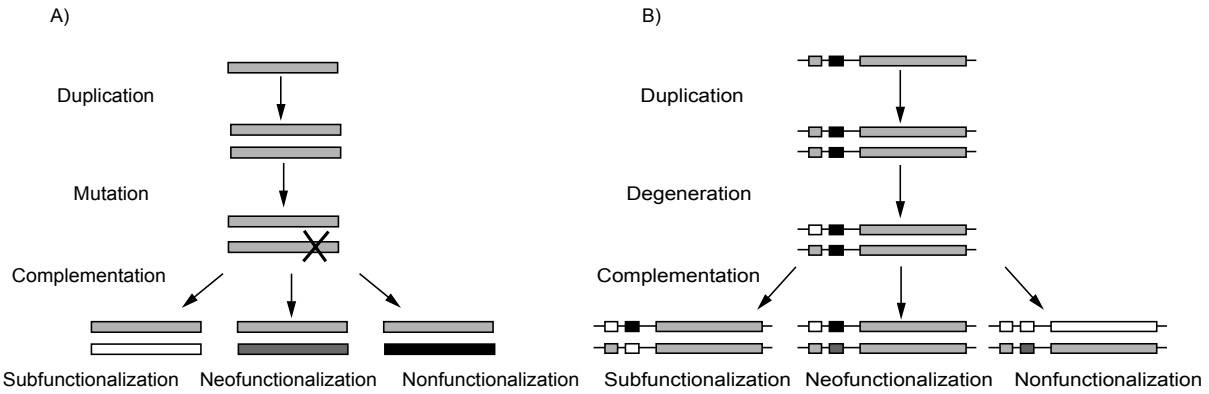


Figure 1: Alternative models for the fate of duplicated genes. A) In the classical model duplicated genes accumulate mutations in the coding area leading mostly to nonfunctionalization and rarely to neo- or subfunctionalization. B) In the DDC model the ancestral genes is here depicted with two mutable regulative regions upstream. After duplication through a degenerative mutation in one of the regulative region (open box) one subfunction is eliminated. The second mutational event dictates the fate of the duplicates. Subfunctionalization will occur when the gene loses the complementary part of its function, neofunctionalization will occur when one copy acquires a novel expression pattern and nonfunctionalization occurs when one copy looses all functional abilities. Modified from Force *et al* 1999 and Lynch 2007

While duplication mechanisms are the driving force in the evolution of NSP, the actual mode of duplication is not known as the origins of segmental duplications are diverse and the actual location of the three genes relative to each other not fully understood yet. Many duplicated genes appear in tandem to their parental copy, such as the multiple copy SDMA genes in *B. mori* (Chapter I). Such organization may arise by unequal crossing over events or replication slippage (Figure 2). Tandem duplications can produce whole gene duplications or duplication of gene parts depending upon the chromosomal region duplicated. Two domain duplication events were necessary to create the precursor gene of NSP and MA from SDMA as discussed in chapter I. Once duplicated, selection retained novel mutations in NSP and MA allowing their function to drift apart. Functional constraint appears to be lower in MA and NSP compared to SDMA, as MA and NSP have higher omega values (see chapter I). In addition, NSP and MA are highly divergent from each other as they only share 50 percent amino acid identity. Comparisons of genomic data from *P. rapae* with the newly released whole genome assembly of *Bombyx mori* suggest NSP to be located on a different chromosome than MA and SDMA, while SDMA and MA appear to be located on the same chromosome (chapter I). Movement of one gene to another chromosome may occur via sloppy transcription of non-LTR (long terminal repeat) retrotransposons followed by the replication of downstream genes and their reinsertion elsewhere in the genome (Lynch 2007), however, this would not replicate the introns as it happened in NSP. While the general mechanisms of duplication for the origin of NSP and MA can be retraced, it is not possible with the current dataset to decide on the mechanism of the segmental duplication that has taken place.

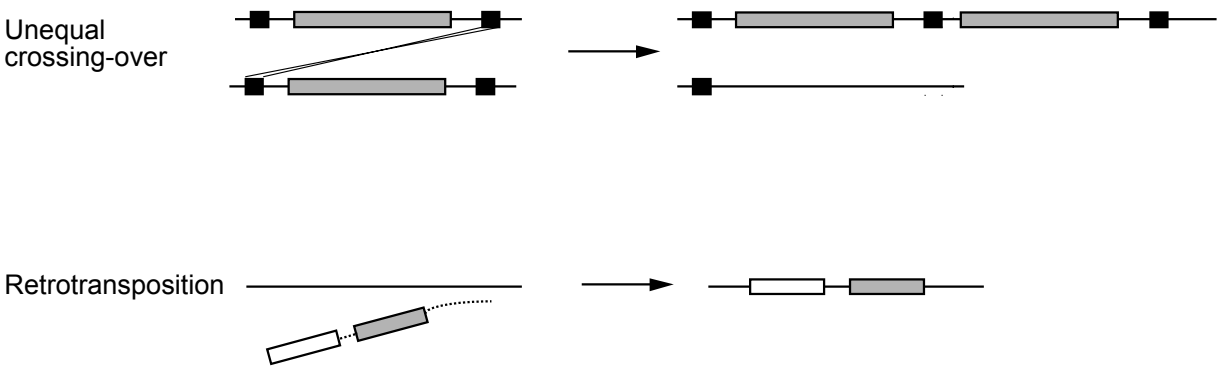


Figure 2: Two mechanisms for the origin of gene duplication. An unequal crossing over event occurs between two regions of sequences similarities (black) at nonhomologous sites. As a result one chromosome harbors a duplication, whereas the other chromosome is lacking that area. Sloppy transcription of a retrotransposon (white) can lead to the additional transcription of a downstream gene (grey). This can lead to insertion in another gene area after reverse transcription (here only depicted as a dashed line). Modified from Lynch 2007.

Duplicated genes retained by selection will experience strong purifying selection after their emergence. The appearance of NSP more than 80 million years ago was an evolutionary novelty enabling a host range shift onto Brassicaceae and a subsequent increase in species diversity of the group (Wheat et al. 2007). The persistence of this adaptation and its now ancient status must be due at least in part to purifying selection, but the current level of intraspecific protein variation suggests limits to the intensity of purifying selection today, as illustrated in chapter II. It appears that NSP is exhibiting an unexpectedly high amount of amino acid polymorphisms in *P. rapae* populations. These polymorphisms are however unequally distributed across the gene, showing regions of high conservation neighboring those with relaxed functional constraint. Surprisingly, this diversity is common in nearly all populations with little to no genetic differentiation among four populations on two continents. The potential implications of this intraspecific protein variation in NSP will be discussed below.

Selection pressure and evolutionary dynamics acting on genes will change over time. Novel genes will experience strong purifying selection enabling their 'survival' in the genome. Adaptive evolution may however - over an evolutionary time frame - cause different selection pressures to act on a given gene. The now ancient status of NSP could allow other factors to influence the evolutionary dynamics acting on it. NSP consists of three repetitions of the same basic domain. Although the precise mode of action of NSP is not known, the repeat structure presents the possibility of functional redundancy and, or, independence among the individual three domains. The increased dN/dS values that we observe in NSP in comparison to the reference genes (chapter II) could be explained in part by functional redundancy reducing purifying selection in certain regions of the domain structure. Slightly deleterious mutations in one domain could be compensated for by the independent functionality of the other two domains. Hence, the possible independent functionality of the NSP domains could obscure the signal of purifying selection in the NSP gene, resulting in the high dN/dS values we observe at the NSP locus compared to random genes.

Adaptation to variable environments can demand variability in the insect responses. In addition to the above proposed compensation effect afforded by the repeated domain structure, the high variability in the NSP alleles might also be driven by environmental factors. *P. rapae* has been reported to feed on at least 16 different plant species, and although specialized on glucosinolate containing plants, it will encounter a large variety of

glucosinolates within its host plant range. Preliminary experiments show that the ability of NSP alleles to convert glucosinolates to nitriles is highly variable. Individuals from the same genetic background appear to have differently performing NSP variants when exposed to glucosinolates and myrosinases in an *in vitro* experiment. These findings suggest that the known high variability of glucosinolates in brassicaceous plants (Windsor et al. 2005) might affect the microevolutionary dynamics of NSP. The variability of the NSP loci might in that respect be a response to the highly variable set of glucosinolates and myrosinases between plant species and even within one plant species enabling *P. rapae* to feed successfully on a wide variety of brassicaceous species. However, the data so far can not rule out the possibility that differences in NSP performances might also be due to different expression intensities of NSP in the caterpillars. Functional antibodies are necessary to test this hypothesis. Nevertheless, the variability of the NSP gene could be a response to the diversity of the host plant system.

The formation of nitriles rather than isothiocyanates has not only evolved in the gut of Pieridae species but also in brassicaceous plants. In *Arabidopsis thaliana* the hydrolysis of glucosinolates will result in epithionitriles and simple nitriles in the presence of the specifier protein ESP (epithio specifier protein) or to simple nitriles in the presence of a nitrile specifier protein (AtNSP1-AtNSP5) (Lambrix et al. 2001; Burow et al. 2006a; Burow et al. 2007b; Burow et al. 2008; Burow et al. 2009). The Arabidopsis nitrile forming proteins all belong to one gene family and show no structural or sequence similarity to the Pieridae NSP. However, insect and plant derived nitrile specifier proteins are all not iron dependent for their activity, whereas ESP activity is strictly iron dependent (Burow et al. 2006a; Burow et al. 2009). The functional role of the plant proteins is not clear yet as nitriles are less toxic than isothiocyanates and generalist herbivore performance is significantly better on nitrile producing plants (Burow et al. 2006b). Nevertheless, recent studies have shown that ovipositing *P. rapae* females prefer wild type over nitrile producing Arabidopsis plants and the parasitoid *Cotesia rubecula*, a specialist on *P. rapae* larvae, is significantly more attracted to *P. rapae* infested, nitrile producing plants than to infested wild type plants (de Vos, Kriksunov, and Jander 2008; Mumm et al. 2008), suggesting that both species might use the highly volatile nitriles as cues for infestation. Simple nitrile formation also is inducible by *P. rapae* herbivore damage in Arabidopsis wild type plants, possibly to attract parasitoids (Burow et al. 2009). Hence, regulation of nitrile versus isothiocyanate formation might

enable brassicaceous plants to adjust their chemical defense in response to specific herbivore attacks and by that adapt better to their environment.

Generalist herbivores need different response strategies to their host plants than specialized herbivores as they will encounter a broad variety of plant defenses. In contrast to the specialized lepidopteran herbivore *P. rapae*, *P. c-album* is feeding on a wide variety of plant species and will therefore encounter a large variety of different secondary compounds as illustrated in chapter III. Either broad detoxifying enzymes could be utilized by *P. c-album* to cope with this complexity or many different detoxifying enzymes that are induced by compounds in the different host plants.

The gene expression profile of *P. c-album* on different hostplants is complex, but the utilization pattern of detoxification genes is not clear. While many differentially expressed genes could be detected in the different treatments, no known detoxifying enzymes could be found to be differently regulated on different host plants in the midgut tissue. This finding can have two divergent explanations. One reason for this finding could be a constant expression of broad acting detoxifying genes in the caterpillar independent of the actual host plant. A constant expression would not have been detected by the experimental approach taken in chapter III. A couple of arguments support this hypothesis. Different to induced genes, that will be slowed down by transcription and translation, constantly expressed genes will be acting immediately, which might be important especially straight after hatching when caterpillars can be expected to be very vulnerable to toxins. This is important since the plant species young *P. c-album* caterpillar will ingest is largely dependent on the availability of host plants as well as on the oviposition choice of the mother and therefore variable and hard to predict. Ovipositing female *P. c-album* butterflies of one population show a surprisingly large variability in host preference, that is not strictly correlated with the performance of their offspring (Janz, Nylin, and Wedell 1994). Hence, as the larvae may hatch on one of many possible hostplants, *P. c-album* might be required to have a detoxifying system that is acting on a very broad substrate range reasonably well. In a second explanation *P. c-album* might have acquired novel inducible detoxifying enzymes that are specific for certain plant compounds. Such enzymes would also not be identified with the approach in chapter III as the sequences found were only compared to known sequences in the databases. This scenario nevertheless appears to be less likely for two reasons. Firstly, induction appears to take too long when host plant availability is very variable as discussed above. Secondly, generalist

lepidopterans have been shown before to use phase I and phase II enzymes to detoxify allelochemicals (Li, Berenbaum, and Schuler 2002; Sasabe et al. 2004; Zeng et al. 2007), making it likely that *P. c-album* is also utilizing them. Therefore, while the data are not clear on detoxification, *P. c-album* caterpillars most likely have a constantly expressed very broad functioning detoxification enzyme system that will act on a broad range of substrates to deal with the variability of plant compounds they encounter.

In generalist herbivores, feeding on chemically-defended plants can be expected to always induce a stronger stress response than in specialist herbivores. In spite of the supposedly successfully metabolized and or excreted allelochemicals *P.c-album* caterpillars will encounter a variety of other factors that will differ between their different host plants. Nutritional contents can be expected to vary greatly as well as slightly toxic compounds that have not been completely metabolized. Many genes involved in digestion, metabolism and translation regulation are differentially expressed in the different treatments. While some of these genes might simply be involved in digestion, it illustrates the complexity of the response, whereby the whole metabolism of the caterpillar needs to adapt to the different plant systems. However, also *P. rapae* will encounter different conditions on different brassicaceous plants and the data in chapter II furthermore suggest that some allelic variants might not be as efficient in converting the encountered glucosinolates, so that differential gene expression to a lesser extent could also be expected in *P. rapae*.

While adaptation to variable environments is important for both butterfly species investigated here, the molecular scale of adaptation appears to differ. *P. rapae* fine-tunes its detoxifying system to different glucosinolate-myrosinase systems, while *P.c-album* is most likely using very broad detoxifying enzymes that will cope with a large variety of allelochemicals. The depth of molecular understanding varies greatly between the two model systems. The origins of adaptation are now better understood for the Pieridae butterflies, while the mode of action of NSP is still unknown. Only after a general understanding of the interaction of NSP with the myrosinase is reached, will it be possible to fully understand the genetic variability detected within the NSP gene. NSP activity could only be detected in glucosinolate feeding Pieridae species, not in basal or derived species (Wheat et al. 2007) and while it could be shown in this thesis that NSP evolved via domain and gene duplication, an interesting future approach would be to understand the mechanisms of loss of NSP in derived non glucosinolate feeding species. The mechanisms of detoxification in *P. c-album* are not understood yet. While the

thesis presented here provides a preliminary glance into the complexity of the caterpillars' response to different hostplants, a better understanding of the host plant chemistry would be useful to identify target genes that serve detoxification in this polyphagous butterfly species.

6. Summary

Although the defensive and counter-defensive molecules underlying many ecological interactions are known, the genetic mechanisms controlling these molecules are often unknown. Knowledge of these mechanisms, as well as the selective forces and adaptations that have shaped them, is necessary if we are to understand the evolution of ecological interactions. In the thesis presented here the molecular mechanisms underlying two plant-insect interaction systems were investigated. Adaptive mutations allowing an insect to utilize a new food plant can have different molecular origins, affecting the regulatory regions as well as the coding sequence of genes. In general it is assumed that phase I and phase II enzymes are important in insects to detoxify plant allelochemicals, but detailed knowledge is still scarce.

The first system involves the Pieridae butterflies and the Brassicaceae plants that have been in a coevolutionary arms race for about 80 million years. To circumvent the activated defense system of the plants, the Pieridae caterpillars posses a unique detoxifying enzyme called Nitrile-specifier protein (NSP), that redirects the hydrolysis of glucosinolates to less toxic nitriles rather than the toxic isothiocyanates in the caterpillar gut. Here the molecular origins of this novel detoxifying mechanism were investigated. It was found that NSP is a member of an insect specific gene family, called the NSP-like gene family. Members of this family consist of variable tandem repeats, are expressed in the gut lumen of the insect and are evolving in an ongoing birth-death process. NSP and its paralog MA evolved through two tandem duplications of the single domain gene SDMA in the Pieridae caterpillars that feed on glucosinolate containing plants. While gene duplication is a common mechanism to adapt to new environments, the molecular evolution of NSP provides a rare example of proven neofunctionalization after duplication. Future research on location, mode of action and genomic location of NSP are necessary to shed light on the mechanism of duplication and neofunctionalization that gave rise to NSP and MA.

Population studies on the little cabbage white, *Pieris rapae,* using four different populations from two continents revealed that NSP is exhibiting unexpectedly high rates of amino acid polymorphism with little to no genetic differentiation among the four surveyed populations. The amino acid substitutions are unequally distributed across the NSP gene and comparisons of synonymous (dS) to nonsynonymous (dN) substitutions between 70 randomly chosen

genes of *P. rapae* and its close relative *Pieris brassicae* (Large Cabbage White) revealed NSP to be evolving much faster than the genomic average. Preliminary experiments indicate performance differences between NSP alleles. Therefore the variability of the NSP loci might be a response to the highly variable set of glucosinolates and myrosinases in and between Brassicaceae plant species, enabling *P. rapae* to feed successfully on a wide variety of plants. Future studies on the mode of action of NSP will also facilitate a better understanding of the variability on the NSP loci.

The second study system in this thesis is the polyphagous Comma Butterfly *Polygonia c-album*. In contrast to *P. rapae*, specialized on glucosinolate containing plants, *P. c-album* feeds on a wide variety of plant species and will therefore encounter a large variety of different secondary compounds. A differential gene expression analysis of *P. c-album* caterpillars on different hostplants revealed a complex picture of gene regulation. Many genes involved in digestion and metabolism and ribosomal proteins were differentially expressed in the caterpillars on the different hostplants, indicating that each plant species requires a very specific set of genes to be regulated. However, no differentially regulated detoxifying enzymes could be identified suggesting that *P. c-album* possesses very broad acting detoxifying genes that are continuously expressed independent of the hostplant. A more detailed knowledge of host plant allelochemicals and genetic resources of *P. c-album* caterpillars are necessary to gain a better understanding of the adaptation mechanisms of this species to its host plants.

In conclusion, adaptations to variable environments are important for both butterfly species investigated, but the molecular scale of adaptation appears to differ. *P. rapae* fine tunes its detoxifying system to different glucosinolate-myrosinase systems, while *P.c-album* is most likely using very broad detoxifying enzymes to cope with a large variety of allelochemicals.

7. Zusammenfassung

Obwohl die Abwehr- und Gegenabwehrmoleküle, die vielen ökologischen Interaktionen unterliegen, bekannt sind, sind die molekularen Mechanismen die diese Moleküle kontrollieren oft unbekannt. Kenntnisse über diese Mechanismen, als auch das Wissen über die selektiven Kräfte und Anpassungen die diese Mechanismen geformt haben, sind erforderlich, wenn wir die Evolution von ökologischen Interaktionen verstehen wollen. In der hier vorliegenden Dissertation wurden die molekularen Mechanismen untersucht, die zwei Pflanzen-Insekten-Interaktionssystemen unterliegen. Adaptive Mutationen, die Insekten erlauben neue Wirtspflanzen zu nutzen, können unterschiedliche molekulare Ursprünge haben. Im Allgemeinen wird angenommen, dass Phase I- und Phase II-Enzyme für die Entgiftung von Pflanzensekundärstoffen in Insekten notwendig sind, aber ein detailliertes Wissen und Verständnis ist noch nicht vorhanden.

Eines der zwei untersuchten Systeme umfaßt die Weißlinge (Pieridae: Lepidoptera) und deren Wirtspflanzen, die Kreuzblütler (Brassicacae), die seit 80 Millionen Jahren in einem coevolutionären Wettlauf stehen. Um das aktivierte Abwehrsystem der Kreuzblütler zu umgehen, besitzen die Raupen der Weißlinge ein einzigartiges Entgiftungsenzym, das sogenannte Nitrile-Specifier-Protein (NSP), welches die Hydrolyse der Glukosinolate zu Gunsten von Nitrtilen beeinflußt, so dass keine giftigen Isothiocyanate entstehen können. In der hier vorgelegten Arbeit werden die molekularen Grundlagen dieses neuen Entgiftungsenzyms untersucht. Es wurde herausgefunden, dass NSP zu einer insektspezifischen Genfamilie gehört, die NSP-like gene family benannt wurde. Mitglieder dieser Genfamilie bestehen aus einer variierenden Anzahl von Tandemwiederholungen, werden im Mitteldarm der Insekten exprimiert und entwickeln sich in einem kontinuierlich dynamischen "birth-death" Prozess. NSP und sein paraloges Gen MA sind durch Tandemduplikation aus dem Einzeldomän-Gen SDMA in den Weißlingen entstanden, die auf glukosinolat-haltigen Pflanzen fressen. Obwohl Genduplikation ein verbreiteter molekularer Mechanismus ist, um sich an neue Umwelten anzupassen, stellt die molekulare Evolution von NSP ein seltenes Beispiel für die nachgewiesene Neufunktionalisierung nach stattgefundener Genduplikation dar. Zukünftige Forschung an der Lokalisation des Proteins, der Wirkungsweise und der genomischen Lokalisation von NSP sind nötig, um die Mechanismen der Duplikation und Neufunktionalisierung aufzuklären, die die Entstehung von NSP und MA ermöglicht haben.

In Populationsstudien am Kleinen Kohlweißling *Pieris rapae* in vier verschiedenen Populationen von zwei Kontinenten wurde festgestellt, dass NSP eine unerwartet hohe Variation an Aminosäurepolymorphismen besitzt, die wenig bis keine Differenzierung zwischen den vier untersuchten Populationen aufweist. Die Aminosäureaustausche sind ungleichmäßig über das NSP Gen verteilt und Vergleiche von synonymen (dS) mit nichtsynonymen (dN) Austauschen zwischen 70 zufällig ausgewählten Genen von *P. rapae* und dem nahe verwandten *Pieris brassicae* (Großer Kohlweißling), zeigte, dass NSP sich schneller entwickelt als der genomische Durchschnitt. Vorläufige Versuche deuten Umsatzunterschiede zwischen unterschiedlichen NSP-Allelen an. Die Variabilität des NSP Gens könnte daher eine Antwort auf das hoch variable Repertoire an Glukosinolaten und Myrosinasen in Kreuzblütlern sein und *P. rapae* befähigen, an einer großen Vielzahl von Kreuzblütlern zu fressen. Zukünftige Forschung an der biochemische Wirkungsweise von NSP wird auch das Verständnis über die Variabilität diese Genes fördern.

Das zweite untersuchte System umfaßt den polyphagen C-Falter *Polygonia c-album*. Im Gegensatz zu der nur auf glukosinolat-haltigen Pflanzen vorkommenden spezialisierten Raupe von *P. rapae*, frißt *P. c-album* an einer großen Vielfalt von Pflanzenarten und wird als Konsequenz auch auf eine Vielzahl von unterschiedlichen sekundären Pflanzenstoffen treffen. Eine differentielle Genexpressionsanalyse der *P. c-album* Raupen nach Fraß auf verschiedenen Wirtspflanzen zeigte ein komplexes Bild der Genregulation. Viele Gene die in der Verdauung und dem Stoffwechsel involviert sind und auch ribosomale Proteine zeigten differentielle Regulation nach Fraß auf verschiedenen Pflanzen und geben dadurch zu erkennen, dass für jede Pflanzenart ein sehr spezifischer Satz an Genen reguliert werden muß. Es wurden jedoch keine differentiell regulierten Entgiftungsenzyme gefunden, was auf Entgiftungsenzyme mit einem breiten Wirkungsspektrum hindeuten könnte, welche kontinuierlich und unabhängig von der Wirtspflanze exprimiert werden. Es ist jedoch ein besseres Verständnis der in den Wirtspflanzen zu findenden Allelochemikalien und der genetischen Ressourcen von *P. c-album* nötig, um zu einer tieferen Einsicht der Anpassungsmechanismen dieser Art zu gelangen.

Als Fazit sind Anpassungen and eine veränderliche Umwelt wichtig für beide untersuchte Schmetterlingsarten, aber der molekulare Maßstab der Anpassungen scheint sich zu unterscheiden. Währen *P. rapae* seine Entgiftungsmaschinerie mit dem Glukosinolat-

Myrosinase-System der Kreuzblütler sehr fein abstimmt, benutzt *P. c-album* weitreichend aktive Entgiftungsenzyme, um die große Vielfalt an pflanzlichen Allelochemikalien zu bewältigen.

8. Acknowledgements

I would like to thank first of all my supervisors Dr. Heiko Vogel, Dr. Christopher W. Wheat and Professor David G. Heckel, who completed one another perfectly. I would like to thank them for giving me the opportunity to work on this project under their advice as well as for their professional support technically and scientifically.
I also would like to thank the following people:
Dr Choon Wei Wee for highly appreciated molecular help and advice,
Henriette Ringys-Beckstein for professional technical advice,
Domenica Schnabelrauch for technical support,
Yannick Pauchet for valuable biochemical advice,
Dalial Freitak for being a companion PhD sharing office, lab and dissertation sorrows with me,
The Entomology Department for creating a fruitful working atmosphere,
Ute Wittstock for giving me the opportunity to work on the topic,
Professor Sören Nylin and Associate Professor Niklas Janz for a fruitful cooperation,
Professor Robert Raguso for Butterfly supplies from the USA,
Professor Jonathan Gershenzon and Professor Daniel Kliebenstein to agree to examine my thesis,
Annett Börner for helping me in graphical emergencies,
and last but not least, my friends, my family and my husband.

9. References

Aminetzach, YT, Macpherson, JM, and Petrov, DA. 2005. Pesticide resistance via transposition-mediated adaptive gene truncation in Drosophila. Science 309:764-767.

Barrett, RDH, and Schluter, D. 2008. Adaptation from standing genetic variation. Trends in Ecology & Evolution 23:38-44.

Barth, C, and Jander, G. 2006. Arabidopsis myrosinases TGG1 and TGG2 have redundant function in glucosinolate breakdown and insect defense. Plant Journal 46:549-562.

Becerra, J. X. 1997. Insects on plants: Macroevolutionary chemical trends in host use. Science, 276: 253-256.

Benderoth, M, Textor, S, Windsor, AJ, Mitchell-Olds, T, Gershenzon, J, and Kroymann, J. 2006. Positive selection driving diversification in plant secondary metabolism. Proceedings of the National Academy of Sciences of the United States of America 103:9118-9123.

Bernays, E. A. 1989. Host range in phytophagous insects: the potential role of generalist predators. Evol. Ecol., 3: 299-311.

Berenbaum, MR. 2002. Postgenomic chemical ecology: From genetic code to ecological interactions. Journal of Chemical Ecology 28:873-896.

Berenbaum, M. R., C. Favret, and M. A. Schuler. 1996. On defining "key innovations" in an adaptive radiation: Cytochrome P450s and papilionidae. Am. Nat., 148: S139-S155.

Berenbaum, M. R., and Feeny, P. 2008. Chemical mediation of host-plant specialization: The papilionid paradigm. pp. 3-19. In K. J. Tilmon (ed.), Specialization, Speciation, and Radiation: the Evolutionary Biology of Herbivorous Insects. University of California Press, Berkeley, California.

Bjorklund, AK, Ekman, D, and Elofsson, A. 2006. Expansion of protein domain repeats. Plos Computational Biology 2:959-970.

Braby, MF, Vila, R, and Pierce, NE. 2006. Molecular phylogeny and systematics of the Pieridae (Lepidoptera: Papilionoidae): higher classification and biogeography. Zoological Journal of the Linnean Society 147:239-275.

Brattsten, LB. 1992. Metabolic defenses against plant allelochemicals, in: Herbivores: their interaction with secondary plant metabolites, Volume 2: Evolutionary and ecological processes. Academic Press Inc.:175-242.

Brenes-Arguedas, T, Horton, MW, Coley, PD, Lokvam, J, Waddell, RA, Meizoso-O'Meara, BE, and Kursar, TA. 2006. Contrasting mechanisms of secondary metabolite accumulation during leaf development in two tropical tree species with different leaf expansion strategies. Oecologia 149:91-100.

Briscoe, AD. 2001. Functional diversification of lepidopteran opsins following gene duplication. Molecular Biology and Evolution 18:2270-2279.

Burow, M, Bergner, A, Gershenzon, J, and Wittstock, U. 2007a. Glucosinolate hydrolysis in Lepidium sativum - identification of the thiocyanate-forming protein. Plant Molecular Biology 63:49-61.

Burow, M, Losansky, A, Müller, R, Plock, A, Kliebenstein, DJ, and Wittstcok, U. 2009. The genetic basis of constitutive and herbivore-induced ESP-independent nitrile formation in Arabidopsis. Plant Physiology 149:561-574.

Burow, M, Markert, J, Gershenzon, J, and Wittstock, U. 2006a. Comparative biochemical characterization of nitrile-forming proteins from plants and insects that alter myrosinase-catalysed hydrolysis of glucosinolates. Febs Journal 273:2432-2446.

Burow, M, Muller, R, Gershenzon, J, and Wittstock, U. 2006b. Altered glucosinolate hydrolysis in genetically engineered Arabidopsis thaliana and its influence on the larval development of Spodoptera littoralis. Journal of Chemical Ecology 32:2333-2349.

Burow, M, Rice, M, Hause, B, Gershenzon, J, and Wittstock, U. 2007b. Cell- and tissue-specific localization and regulation of the epithiospecifier protein in Arabidopsis thaliana. Plant Molecular Biology 64:173-185.

Burow, M, Zhang, ZY, Ober, JA, Lambrix, VM, Wittstock, U, Gershenzon, J, and Kliebenstein, DJ. 2008. ESP and ESM1 mediate indol-3-acetonitrile production from indol-3-ylmethyl glucosinolate in Arabidopsis. Phytochemistry 69:663-671.

Chad Gore, J, and Schal, C. 2005. Expression, production and excretion of Bla g 1, a major human allergen, in relation to food intake in the German cockroach, *Blattella germanica*. Medical and Veterinary Entomology:127-134.

Chew, F. S., and S. P. Courtney. 1991. Plant apparency and evolutionary escape from insect herbivory. Am. Nat., 138: 729-750.

Chew, FS, and Watt, WB. 2006. The green-veined white (Pieris napi L.), its Pierine relatives, and the systematics dilemmas of divergent character sets (Lepidoptera, Pieridae). Biological Journal of the Linnean Society 88:413-435.

Chinen, A, Hamaoka, T, Yamada, Y, and Kawamura, S. 2003. Gene duplication and spectral diversification of cone visual pigments of zebrafish. Genetics 163:663-675.

Daborn, PJ, Yen, JL, Bogwitz, MR et al. 2002. A single P450 allele associated with insecticide resistance in Drosophila. Science 297:2253-2256.

Dethier, V. G. 1941. Chemical factors determining the choice of food plants by *Papilio* larvae. Am. Nat., 75: 61-73.

Dethier, V. G. 1954. Evolution of feeding preferences in phytophagous insects. Evolution, 8: 33-54.

de Vos, M, Kriksunov, KL, and Jander, G. 2008. Indole-3-acetonitrile production from indole glucosinolates deters oviposition by Pieris rapae. Plant Physiology 146:916-926.

Dean, AM, and Thornton, JW. 2007. Mechanistic approaches to the study of evolution: the functional synthesis. Nature Reviews Genetics 8:675-688.

Dowd, P. F., Smith C. M. and Sparks T. C. 1983. Detoxification of plant toxins by insects. Insect Biochemistry, 13 (5): 453-468.

Ehrlich, PR, and Raven, PH. 1964. Butterflies and Plants: A Study in Coevoution. Evolution 18:586-608.

Excoffier, L, Laval, G, and Schneider, S. 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1:47-50.

Excoffier, L, Smouse, PE, and Quattro, JM. 1992. Analysis of Molecular Variance Inferred from Metric Distances among DNA Haplotypes - Application to Human Mitochondrial-DNA Restriction Data. Genetics 131:479-491.

Fay, JC, and Wu, CI. 2000. Hitchhiking under positive Darwinian selection. Genetics 155:1405-1413.

Feeny, P. 1976. Plant apparency and chemical defence. pp. 1-40. In J. W. Wallace, and N. R.L. (eds.), Biological Interactions Between Plants and Insects. Plenum Press.

Felsenstein, J. 2004. Infering Phylogenies. Sunderland: Sinauer Associates.

Ferreira, AH, Cristofoletti, PT, Pimenta, DC, Ribeiro, AF, Terra, WR, and Ferreira, C. 2008. Structure, processing and midgut secretion of putative peritrophic membrane ancillary protein (PMAP) from Tenebrio molitor larvae. Insect Biochemistry and Molecular Biology 38:233-243.

Fischer, HM, Wheat, CW, Heckel, DG, and Vogel, H. 2008. Evolutionary Origins of a Novel Host Plant Detoxification Gene in Butterflies. Molecular Biology and Evolution 25:809-820.

Force, A, Lynch, M, Pickett, FB, Amores, A, Yan, YL, and Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531-1545.

Fraenkel, G. S. 1959. The raison d'être of secondary plant substances. Science, 129: 1466-1470.

Freitak D., C.W. Wheat, D.G. Heckel, and H. Vogel. 2007. Immune system responses and fitness costs associated with consumption of bacteria in larvae of *Trichoplusia ni*. *BMC Biology* 5:56.

Frentiu, FD, Bernard, GD, Sison-Mangus, MP, Brower, AVZ, and Briscoe, AD. 2007. Gene duplication is an evolutionary mechanism for expanding spectral diversity in the long-wavelength photopigments of butterflies. Molecular Biology and Evolution 24:2016-2028.

Fu, YX, and Li, WH. 1993. Statistical Tests of Neutrality of Mutations. Genetics 133:693-709.

Futuyma, D. J. 1976. Food plant specialization and environmental predictability in Lepidoptera. Am. Nat., 110: 285-292.

Futuyma, D. J., M. C. Keese, and S. J. Scheffer. 1993. Genetic constraints and the phylogeny of insect-plant associations - responses of *Ophraella communa* (Coleoptera, Chrysomelidae) to hostplants of its congeners. Evolution, 47: 888-905.

Gould, SJ, and Vrba, ES. 1982. Exaptation - a Missing Term in the Science of Form. Paleobiology 8:4-15.

Gross, E. M., A. Brune, and O. Waleciak. 2008. Gut pH, redox and oxygen levels in an aquatic caterpillar: Potential effects on the fate of ingested tannins. J. Insect. Physiol, 54: 462-471.

Halkier, BA, and Gershenzon, J. 2006. Biology and biochemistry of glucosinolates. Annual Review of Plant Biology 57:303-333.

Hall, TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program fro Windows 95/98/NT. Nucl. Acids. Symp. Ser. 41:95-98.

He, XL, and Zhang, JZ. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169:1157-1164.

Hegenauer, R. 1973. Chemotaxonomie der Pflanzen, Band 6, Birkhäuser Buch

Heidel, AJ, Clauss, MJ, Kroymann, J, Savolainen, O, and Mitchell-Olds, T. 2006. Natural variation in MAM within and between populations of Arabidopsis lyrata determines glucosinolate phenotype. Genetics 173:1629-1636.

Herde, M, Gartner, K, Kollner, TG, Fode, B, Boland, W, Gershenzon, J, Gatz, C, and Tholl, D. 2008. Identification and regulation of TPS04/GES, an Arabidopsis geranyllinalool synthase catalyzing the first step in the formation of the insect-induced volatile C-16-homoterpene TMTT. Plant Cell 20:1152-1168.

Hiryama C., Ono H., Tamura Y., Konno K., Nakamura M. 2008. Regioselective formation of quercetin 5-*O*-glucosides from orally administered quercetin in the silkworm *Bombyx mori.Phytochemistry* 69: 1141-1149

Hoffmann, M, Tripathi, N, Henz, SR, Lindholm, AK, Weigel, D, Breden, F, and Dreyer, C. 2007. Opsin gene duplication and diversification in the guppy, a model for sexual selection. Proceedings of the Royal Society B-Biological Sciences 274:33-42.

Horne I., Haritos V. 2008. Multiple tandem gene duplication in a neutral lipase gene cluster in *Drosophila*. Gene 411: 21-37.

Huang, HS, Hu, NT, Yao, YE, Wu, CY, Chiang, SW, and Sun, CN. 1998. Molecular cloning and heterologous expression of a glutathione S-transferase involved in insecticide resistance from the diamondback moth, Plutella xylostella. Insect Biochemistry and Molecular Biology 28:651-658.

Hudson, RR, Kreitman, M, and Aguade, M. 1987. A Test of Neutral Molecular Evolution Based on Nucleotide Data. Genetics 116:153-159.

Hughes, AL. 2007. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity 99:364-373.

Janz, N, Nylin, S, and Wedell, N. 1994. Host-Plant Utilization in the Comma Butterfly - Sources of Variation and Evolutionary Implications. Oecologia 99:132-140.

Janz, N., and S. Nylin. 1998. Butterflies and plants: a phylogenetic study. Evolution, 52: 486-502.

Janz, N., and Nylin, S. 2008. The oscillation hypothesis of hostplant-range and speciation. pp. 203-215. In K. J. Tilmon (ed.), Specialization, Speciation, and Radiation: the Evolutionary Biology of Herbivorous Insects. University of California Press, Berkeley, California.

Jones, RE, Gilbert, N, Guppy, M, and Nealis, V. 1980. Long-Distance Movement of Pieris-Rapae. Journal of Animal Ecology 49:629-642.

Janz, N., S. Nylin, and K. Nyblom. 2001. Evolutionary dynamics of hostplant specialization: a case study of the tribe Nymphalini. Evolution, 55: 783–796.

Janz, N., S. Nylin, and N. Wahlberg. 2006. Diversity begets diversity: host expansions and the diversification of plant-feeding insects. BMC Evol. Biol., 6: 4.

Johnson, K. S., and G. W. Felton. 1996. Physiological and dietary influences on midgut redox conditions in generalist Lepidopteran larvae. J. Insect Physiol., 42: 191-193.

Johnson K.S. 2005.Plant phenolics behave as radical scavengers in the context of insect (*Manduca sexta*) hemolymph and midgut fluid. *Journal of Agriculture and Food Chemistry*. 53: 10120-10126

Jörnvall, H., B. Persson, M. Krook, S. Atrian, R. Gonzales-Duarte, J. Jeffery, and D. Ghosh. 1995. Short-chain dehydrogenases/reductases (SDR). Biochemistry, 34: 6003-6013.

Jordan, IK, Wolf, YI, and Koonin, EV. 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. Bmc Evolutionary Biology 4.

Kallberg, Y., U. Oppermann, H. Jörnvall, and B. Persson. 2002. Short-chain dehydrogenase/reductase (SDR) relationships: A large family with eight clusters common to human, animal, and plant genomes. Protein Science, 11: 636-641.

Kawahara, R, and Nishida, M. 2007. Extensive lineage-specific gene duplication and evolution of the spiggin multi-gene family in stickleback. Bmc Evolutionary Biology 7.

Kergoat, G. J., A. Delobel, G. Fediere, B. Le Ru, and J. F. Silvain. 2005. Both host-plant phylogeny and chemistry have shaped the African seed-beetle radiation. Mol. Phyl. Evol., 35: 602-611.

Kliebenstein, D, Pedersen, D, Barker, B, and Mitchell-Olds, T. 2002. Comparative analysis of quantitative trait loci controlling glucosinolates, myrosinase and insect resistance in Arabidapsis thaliana. Genetics 161:325-332.

Kliebenstein, DJ, Kroymann, J, and Mitchell-Olds, T. 2005. The glucosinolate-myrosinase system in an ecological and evolutionary context. Current Opinion in Plant Biology 8:264-271.

Koroleva, OA, Davies, A, Deeken, R, Thorpe, MR, Tomos, AD, and Hedrich, R. 2000. Identification of a new glucosinolate-rich cell type in Arabidopsis flower stalk. Plant Physiology 124:599-608.

Kroymann, J, Donnerhacke, S, Schnabelrauch, D, and Mitchell-Olds, T. 2003. Evolutionary dynamics of an Arabidopsis insect resistance quantitative trait locus. Proceedings of the National Academy of Sciences of the United States of America 100:14587-14592.

Kroymann, J, Textor, S, Tokuhisa, JG, Falk, KL, Bartram, S, Gershenzon, J, and Mitchell-Olds, T. 2001. A gene controlling variation in arabidopsis glucosinolate composition is part of the methionine chain elongation pathway. Plant Physiology 127:1077-1088.

Lambrix, V, Reichelt, M, Mitchell-Olds, T, Kliebenstein, DJ, and Gershenzon, J. 2001. The Arabidopsis epithiospecifier protein promotes the hydrolysis of glucosinolates to nitriles and influences Trichoplusia ni herbivory. Plant Cell 13:2793-2807.

Li, WM, Schuler, MA, and Berenbaum, MR. 2003. Diversification of furanocoumarin-metabolizing cytochrome P450 monooxygenases in two papilionids: Specificity and substrate encounter rate. Proceedings of the National Academy of Sciences of the United States of America 100:14593-14598.

Li, X., J. Baudry, M. R. Berenbaum, and M. A. Schuler. 2004. Structural and functional divergence of insect CYP6B proteins: From specialist to generalist cytochrome P450. Proc Natl Acad Sci U S A, 101: 2939-2944.

Li, XC, Berenbaum, MR, and Schuler, MA. 2002. Plant allelochemicals differentially regulate Helicoverpa zea cytochrome P450 genes. Insect Molecular Biology 11:343-351.

Lopez-Vaamonde, C., H. Charles, J. Godfray, and J. M. Cook. 2003. Evolutionary dynamics of host-plant use in a genus of leaf-mining moths. Evolution, 57: 1804-1821.

Loppin, B, Lepetit, D, Dorus, S, Couble, P, and Karr, TL. 2005. Origin and neofunctionalization of a Drosophila paternal effect gene essential for zygote viability. Current Biology 15:87-93.

Lundström, A., G. Liu, D. Kang, K. Berzins, and H. Steiner. 2002 *Trichoplusia ni* gloverin, an inducible immune gene encoding an antibacterial insect protein. Insect Biochemistry and Molecular Biology, 32: 795-801.

Lynch, M. 2007. The origins of genome architecture. Sinauer Association.

Lynch, M. 2002. Genomics - Gene duplication and evolution. Science 297:945-947.

Lynch, M, and Conery, JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151-1155.

Maischak, H, Grigoriev, PA, Vogel, H, Boland, W, and Mithöfer, A. 2007. Oral secretion from herbivores lepidopteran larvae exhibit ion channel forming activities. Federation of European Biochemical Society Letters 581:898-904.

Mao, W., M. A. Berhow, A. R. Zangerl, J. McGovern, and M. R. Berenbaum. 2006. Cytochrome P450-Mediated Metabolism of Xanthotoxin by Papilio multicaudatus. J. Chem. Ecol.

Mao, W., M. A. Schuler, and M. R. Berenbaum. 2007. Cytochrome P450s in Papilio multicaudatus and the transition from oligophagy to polyphagy in the Papilionidae. Insect Mol Biol, 16: 481-490.

Mauricio, R. 1998. Costs of resistance to natural enemies in field populations of the annual plant Arabidopsis thaliana. American Naturalist 151:20-28.

Menken, S. B. J. 1996. Pattern and process in the evolution of insect-plant associations: Yponomeuta as an example. Entomol. Exp. Appl., 80: 297-305.

McBride, CS, and Arguello, JR. 2007. Five drosophila genomes reveal nonneutral evolution and the signature of host specialization in the chemoreceptor superfamily. Genetics 177:1395-1416.

McDonald, JH, and Kreitman, M. 1991. Adaptive Protein Evolution at the Adh Locus in Drosophila. Nature 351:652-654.

Meyling N.V., and J. Eilenberg. 2006. Isolation and characterisation of *Beaveria bassiana* isolates from phylloplanes of hedgegrow vegetation. *Mycological Research* 110: 188-195.

Michalakis, Y, and Excoffier, L. 1996. A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics 142:1061-1064.

Miller, A. M., C. McArthur, and P. J. Smethurst. 2007. Effects of within-patch characteristics on the vulnerability of a plant to herbivory. Oikos, 116: 41-52.

Mumm, R, Burow, M, Bukovinszkine'Kiss, G, Kazantzidou, E, Wittstock, U, Dicke, M, and Gershenzon, J. 2008. Formation of Simple Nitriles upon Glucosinolate Hydrolysis Affects Direct and Indirect Defense Against the Specialist Herbivore, Pieris rapae. Journal of Chemical Ecology 34:1311-1321.

Mumm, R, and Hilker, M. 2006. Direct and indirect chemical defence of pine against folivorous insects. Trends in Plant Science 11:351-358.

Murphy, S. M., and P. Feeny. 2006. Chemical facilitation of a naturally occurring host shift by Papilio machaon butterflies (Papilionidae). Ecological Monographs, 76: 399-414.

Nauen R.. and Denholm I. 2005. Resistance of insect pest to neonicotinoid insecticides: Current status and future prospects. *Archives of Insect Biochemistry and Physiology* 58:200-215.

Nei, M. 1987. Molecular Evolutionary Genetics. Columbia Univ. Press, New York.

Nei, M, and Rooney, AP. 2005. Concerted and birth-and-death evolution of multigene families. Annual Review of Genetics 39:121-152.

Nordborg, M, and Innan, H. 2003. The genealogy of sequences containing multiple sites subject to strong selection in a subdivided population. Genetics 163:1201-1213.

Nury, H, Dahout-Gonzalez, C, Trezeguet, V, Lauquin, GJM, Brandolin, G, and Pebay-Peyroula, E. 2006. Relations between structure and function of the mitochondrial ADP/ATP carrier. Annual Review of Biochemistry 75:713-741.

Nylin, S. 1988. Host Plant Specialization and Seasonality in a Polyphagous Butterfly, Polygonia C-Album (Nymphalidae). Oikos 53:381-386.

Nylin, S., and N. Janz. in press. Butterfly hostplant range: an example of plasticity as a promoter of speciation? Evol. Ecol., DOI: 10.1007/s10682-007-9205-5

Ohno, S. 1970. Evolution by gene duplication. New York: Springer-Verlag.

Ohsaki, N. 1980. Comparative Population Studies of 3 Pieris Butterflies, Pieris-Rapae, Pieris-Melete and Pieris-Napi, Living in the Same Area .2. Utilization of Patchy Habitats by Adults through Migratory and Non-Migratory Movements. Researches on Population Ecology 22:163-183.

Ohsaki, N. 1979. Comparative Population Studies of 3 Pieris Butterflies, Pieris-Rapae, Pieris-Melete and P-Napi, Living in the Same Area .1. Ecological Requirements for Habitat Resources in the Adults. Researches on Population Ecology 20:278-296.

Ohshima, I., and K. Yoshizawa. 2006. Multiple host shifts between distantly related plants, Juglandaceae and Ericaceae, in the leaf-mining moth Acrocercops leucophaea complex (Lepidoptera : Gracillariidae). Mol. Phyl. Evol., 38: 231-240.

Pearson, K, Saito, H, Woods, SC, Lund-Katz, S, Tso, P, Phillips, MC, and Davidson, WS. 2004. Structure of human apolipoprotein A-IV: A distinct domain architecture among exchangeable apolipoproteins with potential functional implications. Biochemistry 43:10719-10729.

Pomes, A, Melen, E, Vailes, LD, Retief, JD, Arruda, LK, and Chapman, MD. 1998. Novel allergen structures with tandem amino acid repeats derived from German and American cockroach. Journal of Biological Chemistry 273:30801-30807.

Ponnuvel K. M., H. Nakazawa, S. Furukawa, A. Asaoka, J. Ishibashi, H. Tanaka, and M. Yamakawa. 2003. A lipase isaolated from the silkworm *Bombyx mori* shows antiviral activity against nucleopolyhedrovirus. Journal of Virology, 77: 10725-10729.

Ponting, CP, Mott, R, Bork, P, and Copley, RR. 2001. Novel protein domains and repeats in Drosophila melanogaster: Insights into structure, function, and evolution. Genome Research 11:1996-2008.

Posada, D, and Crandall, KA. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817-818.

Rask, L, Andreasson, E, Ekbom, B, Eriksson, S, Pontoppidan, B, and Meijer, J. 2000. Myrosinase: gene family evolution and herbivore defense in Brassicaceae. Plant Molecular Biology 42:93-113.

Rastogi, S, and Liberles, DA. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. Bmc Evolutionary Biology 5:28.

Ratzka, A, Vogel, H, Kliebenstein, DJ, Mitchell-Olds, T, and Kroymann, J. 2002. Disarming the mustard oil bomb. Proceedings of the National Academy of Sciences of the United States of America 99:11223-11228.

Raymond, M, and Rousset, F. 1995. An exact test for population differentiation. Evolution 49:1280-1283.

Rice, WR. 1989. Analyzing Tables of Statistical Tests. Evolution 43:223-225.

Ronquist, F, and Huelsenbeck, JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Ronquist, F., and J. Liljeblad. 2001. Evolution of the gall wasp-hostplant association. Evolution, 55: 2503-2522.

Rost, B, Yachdav, G, and Liu, J. 2004. The PredictProtein Server. Nucleic Acids Res 32 (Web Serever Issue).

Roy, B. A. 2001. Patterns of association between crucifers and their flower- mimic pathogens: host jumps are more common than coevolution or cospeciation. Evolution, 55: 41-53.

Rozas, J, Sanchez-DelBarrio, JC, Messeguer, X, and Rozas, R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496-2497.

Sasabe, M, Wen, ZM, Berenbaum, MR, and Schuler, MA. 2004. Molecular analysis of CYP321A1, a novel cytochrome P450 involved in metabolism of plant allelochernicals (furanocoumarins) and insecticides (cypermethrin) in Helicoverpa zea. Gene 338:163-175.

Schlenke, TA, and Begun, DJ. 2004. Strong selective sweep associated with a transposon insertion in Drosophila simulans. Proceedings of the National Academy of Sciences of the United States of America 101:1626-1631.

Shao, L, Devenport, M, Fujioka, H, Ghosh, A, and Jacobs-Lorena, M. 2005. Identification and characterization of a novel peritrophic matrix protein, Ae-Aper50, and the microvillar membrane protein, AEG12, from the mosquito, Aedes aegypti. Insect Biochemistry and Molecular Biology 35:947-959.

Slatkin, M. 1995. A Measure of Population Subdivision Based on Microsatellite Allele Frequencies. Genetics 139:457-462.

Slatkin, M. 1991. Inbreeding Coefficients and Coalescence Times. Genetical Research 58:167-175.

Spaethe, J, and Briscoe, AD. 2004. Early duplication and functional diversification of the opsin gene family in insects. Molecular Biology and Evolution 21:1583-1594.

Steppuhn, A, and Baldwin, IT. 2007. Resistance management in a native plant: nicotine prevents herbivores from compensating for plant protease inhibitors. Ecology Letters 10:499-511.

Stipanovic, RD, Puckhaber, LS, and Bell, AA. 2006. Ratios of (+)- and (-)-gossypol in leaves, stems, and roots of selected accessions of Gossypium hirsutum var. marie galante (Watt) Hutchinson. Journal of Agricultural and Food Chemistry 54:1633-1637.

Strode, C, Steen, K, Ortelli, F, and Ranson, H. 2006. Differential expression of the detoxification genes in the different life stages of the malaria vector Anopheles gambiae. Insect Molecular Biology 15:523-530.

Swofford, DL. 2003. PAUP: Phylogenetic Analysis Unsing Parsimony. Sunderland: Sinauer Assocoiates.

Tajima, F. 1989. Statistical-Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics 123:585-595.

Takami, Y, Koshio, C, Ishii, M, Fujii, H, Hidaka, T, and Shimizu, I. 2004. Genetic diversity and structure of urban populations of Pieris butterflies assessed using amplified fragment length polymorphism. Molecular Ecology 13:245-258.

Terra W. R., and C. Ferreira. 2005. in *Comprehensive Molecular Insect Science* 4, eds. Gilbert LI, Iatrou K, Gill SS (Elsevier, Oxford), pp 171-224.

Thompson, JD, Gibson, TJ, Plewniak, F, Jeanmougin, F, and Higgens, DG. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acid Research 25:4876-4882.

Thoorez, L., Van Deun, K., Tranchevent, L.-C., Van Lommel, L. Engelen, K., Marchal, K.,

Thorsteinson, A. J. 1960. Host selection in phytophagous insects. Annu. Rev. Entomol., 5: 193-218.

Moreau, Y., Van Mechelen, I. and Schuit, F. 2008. Using ribosomal protein genes as reference: A tale of caution. Plos One 3(3) e1854.

Volkoff A-N, J. Rocher, E. d'Alençon, M. Bouton, I. Landais, E. Quesada-Moraga, A. Vey, P. Fournier, K . Mita, and G. Devauchelle. 2003. Characterization and transcriptional profiles of three *Spodoptera frugiperda* genes encoding cysteine-rich peptides. A new class of defensin-like genes from lepidopteran insects? *Gene* 319: 43-53.

Vontas, JG, Small, GJ, Nikou, DC, Ranson, H, and Hemingway, J. 2002. Purification, molecular cloning and heterologous expression of a glutathione S-transferase involved in insecticide resistance from the rice brown planthopper, Nilaparvata lugens. Biochemical Journal 362:329-337.

Wadleigh, RW, and Yu, SJ. 1988. Detoxification of Isothiocyanate Allelochemicals by Glutathione Transferase in 3 Lepidopterous Species. Journal of Chemical Ecology 14:1279-1288.

Wahlberg, N. 2001. The phylogenetics and biochemistry of hostplant specialization in melitaeine butterflies (Lepidoptera: Nymphalidae). Evolution, 55: 522-537.

Wang, NM, Lee, MF, and Wu, CH. 1999. Immunologic characterization of a recombinant American cockroach (Periplaneta americana) Per a 1 (Cr-PII) allergen. Allergy 54:119-127.

Wasserman, S. S. 1979. Allelochemic diversity and plant apparency: Evidence from the detoxification systems of caterpillars. American Midland Naturalist, 102: 401-403.

Watterson, GA. 1975. Number of Segregating Sites in Genetic Models without Recombination. Theoretical Population Biology 7:256-276.

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

Wen, ZM, Rupasinghe, S, Niu, GD, Berenbaum, MR, and Schuler, MA. 2006. CYP6B1 and CYP6B3 of the black swallowtail (Papilio polyxenes): Adaptive evolution through subfunctionalization. Molecular Biology and Evolution 23:2434-2443.

West-Eberhard, M. J. 2003. Developmental plasticity and evolution. Oxford University Press, New York.

Wheat, CW, Vogel, H, Wittstock, U, Brayby, M, Underwood, D, and Mitchell-Olds, T. 2007. The genetic basis of a coevolutionary key innovation and its effect on diversification rate. PNAS in press.

Wheat, CW, Watt, WB, Pollock, DD, and Schulte, PM. 2006. From DNA to fitness differences: Sequences and structures of adaptive variants of Colias phosphoglucose isomerase (PGI). Molecular Biology and Evolution 23:499-512.

Wikstrom, N, Savolainen, V, and Chase, MW. 2001. Evolution of the angiosperms: calibrating the family tree. Proceedings of the Royal Society of London Series B-Biological Sciences 268:2211-2220.

Windsor, AJ, Reichelt, M, Figuth, A, Svatos, A, Kroymann, J, Kliebenstein, DJ, Gershenzon, J, and Mitchell-Olds, T. 2005. Geographic and evolutionary diversification of glucosinolates among near relatives of Arabidopsis thaliana (Brassicaceae). Phytochemistry 66:1321-1333.

Wittstcok, U, Kliebenstein, DJ, Lambrix, V, Reichelt, M, and Gershenzon, J. 2003. Glucosinolate hydrolysis and its impacts on generalist insect herbivores. In: Recent Advances in Phytochemitry - Intergrative Phytochemitry. From Ethnobotany to Molecular Ecology Elsivier, Amsterdam:101-126.

Wittstock, U, Agerbirk, N, Stauber, EJ, Olsen, CE, Hippler, M, Mitchell-Olds, T, Gershenson, J, and Vogel, H. 2004. Successful herbivore attack due to metabolic diversion of a plant chemical defense. Proceedings of the National Academy of Sciences of the United States of America 101:4859-4864.

Wittstock, U, and Halkier, BA. 2002. Glucosinolate research in the Arabidopsis era. Trends in Plant Science 7:263-270.

Wright, SI, and Charlesworth, B. 2004. The HKA test revisited: A maximum-likelihood-ratio test of the standard neutral model. Genetics 168:1071-1076.

Yang, ZH. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer applications in the biosciences: CABIOS 13:555-556.

Yang, ZH, and Bielawski, JP. 2000. Statistical methods for detecting molecular adaptation. Trends in Ecology & Evolution 15:496-503.

Yu, SJ. 1982. Host Plant Induction of Glutathione S-Transferase in the Fall Armyworm. Pesticide Biochemistry and Physiology 18:101-106.

Zangerl, A. R., and M. R. Berenbaum. 2003. Phenotype matching in wild parsnip and parsnip webworms: causes and consequences. Evolution, 57: 806-815.

Zeng, RS, Niu, GD, Wen, ZM, Schuler, MA, and Berenbaum, MR. 2007. Toxicity of aflatoxin B1 to Helicoverpa zea and bioactivation by cytochrome p450 monooxygenases (vol 32, pg 1459, 2006). Journal of Chemical Ecology 33:1-1.

Zavala J. A., A.G. Patankar, K. Gase, D. Hui, and I. T. Baldwin. 2004. Manipulation of endogenous trypsin proteinase inhibitor production in *Nicotiana attenuata* demonstartes their function as antiherbivore defenses. Plant Physiology, 134: 1181-1190.

Zhang, Z, and Kishino, H. 2004. Genomic background predicts the fate of duplicated genes: Evidence from the yeast genome. Genetics 166:1995-1999.

10. Selbständigkeitserklärung

Die zur Zeit gültige Promotionsordnung der Biologisch-Pharmazeutischen Fakultät der Friedrich-Schiller-Universität ist mir bekannt. Die vorliegende Arbeit wurde von mir selbst und nur unter Verwendung der angegebenen Hilfsmittel erstellt und alle benutzten Quellen angegeben. Alle Personen, die an der experimentellen Durchführung, Auswertung des Datenmaterials oder bei der Verfassung der Manuskripte beteiligt waren, sind benannt.

Es wurden weder bezahlte noch unbezahlte Hilfe eines Promotionsberaters in Anspruch genommen.

Die vorliegende Arbeit wurde bisher weder als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung noch bei einer anderen Hochschule als Dissertation eingereicht.

Jena, den 2009 ……

Hanna M. Heidel-Fischer

11. Curriculum Vitae

Personal

Name:	Hanna Marieke Heidel-Fischer, née Fischer
Date of birth:	14. 07.1977
Place of birth:	Berlin Charlottenburg, Germany
Address:	Herderstr. 16, 07734 Jena; Germany
Martial status:	married
Email:	hanna.fischer@gmx.net, hfischer@ice.mpg.de

Scientific career

Feb 2005 – current	PhD student at the Max-Planck-Institute for Chemical Ecology in Jena and the Friedrich-Schiller-University, Jena.
Dec 2003 – Aug '04	Diploma Thesis at the FU-Berlin and the Max-Planck-Institute for Chemical Ecology: 'Molecular Analysis of *Pinus sylvestris* after Insect Oviposition' supervised by Prof. Monika Hilker (FU Berlin, Department of Applied Zoology) and Prof. Jonathan Gershenzon (MPICE, Biochemistry Department)
Jan - Dec 2001	'Postgraduate Diploma in Environmental Science' at the University of Canterbury, New Zealand
1998-2004	University studies in biology at the Freie-Universität Berlin Major: zoology, ecology, microbiology, bio–psychology

Education

Aug 1990 - July 1997	Paul-Natorp-Oberschule, Berlin
July1984 – June1990	Paul-Klee-Grundschule and Löcknitz Grundschule, Berlin

Practical training and stays abroad

1997:	Volunteer at the 'Cape Tribulation Tropical Research Station' in Northern Queensland , Australia
1998:	Volunteer at the 'Wilderness Trust', Hamilton, New Zealand
2000:	work experience student at the Chester Zoo, Chester, UK
2002 - 2003:	student worker at the 'Freie Universität Berlin' in the Institute for Chemical Ecology / Applied Zoology:
Sept-Oct 2002:	work experience student at the Robert Koch Institute

12. Publications

FISCHER HM, WHEAT CW, HECKEL DG, VOGEL H (2008): Evolutionary origins of a novel host plant detoxification gene in butterflies. *Molecular Biology & Evolution* 25, 809-820

KÖPKE D, SCHRÖDER R, **FISCHER HM**, GERSHENZON J, HILKER M, SCHMIDT A (2008): Does egg deposition by herbivorous pine sawflies affect transcription of sesquiterpene synthases in pine? *Planta* 228, 427-438

PAUCHET Y, FREITAK D, **HEIDEL-FISCHER HM**, HECKEL DG, VOGEL H (2008): Immunity or digestion: Glucanase activity in a glucan-binding protein family from lepidoptera. *Journal of Biological Chemistry* 284

Oral Presentations

FISCHER HM, WHEAT CW, HECKEL DG, VOGEL H (2008) Evolutionary origins and local adaptation of a novel host plant detoxification gene in butterflies. *XXIII International Congress of Entomology*, Durban, South Africa.

FISCHER HM, WHEAT CW, HECKEL DG, VOGEL H (2007) Molecular evolution of an ecologically important gene in the butterfly family Pieridae. *13th International Symposium on Insect-Plant Relationships*, Uppsala, Sweden.

FISCHER HM, WHEAT CW, HECKEL DG, VOGEL H (2006) Host plant adaptation and characterization in the Pieridae family. *The seventh International Workshop on Molecular Biology and Genetics of the Lepidoptera*, Crete, Greece.

Poster Presentations

FISCHER HM, WHEAT CW, HECKEL DG, VOGEL H (2007) Evolutionary origins of a novel host plant detoxification gene in butterflies, *11th Congress of the European Society for Evolutionary Biology*, Uppsala, Sweden.

FISCHER HM, WHEAT CW, HECKEL DG, VOGEL H (2006) A first glance in to the evolutionary events leading towards the adaptation to Brassicacae plants in the Pieridae family, 1st *International Conference on Glucosinolates, Photochemical Society of Europe*, Jena, Germany.

Jena, den 2009 ……

Hanna M. Heidel-Fischer